

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
SCIENTIFIQUE
Université des Sciences et de la Technologie HOUARI BOUMEDIENE

Faculté d'Électronique et d'Informatique



THÈSE

Présentée pour l'obtention du **grade** de **DOCTORAT**

En : INFORMATIQUE

Spécialité : Intelligence Artificielle

Par : MOUDJARI Leila

Sujet

**Analyse des opinions dans les réseaux sociaux
et linked data**

Soutenue publiquement, le 18/07/2021, devant le jury composé
de :

Mme.	KHELLAF-HANED Faiza	Professeur	à l'USTHB	Présidente
Mme.	AKLI-ASTOUATI Karima	Professeur	à l'USTHB	Directrice de thèse
M.	AZZOUNE Hamid	Professeur	à l'USTHB	Examineur
M.	KAZAR Okba	Professeur	à l'Univ. Biskra	Examineur
M.	BOUKHALFA Kamel	Professeur	à l'USTHB	Examineur
Mme.	FAREH Messaouda	MCA	à l'Univ. BLIDA	Examinatrice
Mme.	BENAMARA-ZITOUNE Farah	HDR	at Univ. Toulouse	Membre Invitée

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
University of Science and Technology HOUARI BOUMEDIENE

Faculty of Electronics and Computer Science



THESIS

Presented to obtain the **DOCTORATE** grade

In : COMPUTER SCIENCE

SPECIALITY : Artificial Intelligence

By : MOUDJARI Leila

Subject

Sentiment analysis in social media and linked data

Publicly defended, on 18/07/2021, before a jury composed of :

Mme.	KHELLAF-HANED Faiza	Professor	at USTHB	President
Mme.	AKLI-ASTOUATI Karima	Professor	at USTHB	Thesis supervisor
Mr.	AZZOUNE Hamid	Professor	at USTHB	Examiner
Mr.	KAZAR Okba	Professor	at Univ. Biskra	Examiner
Mr.	BOUKHALFA Kamel	Professor	at USTHB	Examiner
Mme.	FAREH Messaouda	MCA	at Univ. BLIDA	Examiner
Mme.	BENAMARA-ZITOUNE Farah	HDR	at Univ. Toulouse	Invited Member

Dedication

I would like to dedicate this thesis to my mum, the most important person in my life. Thank you for being always there supporting, helping, and encouraging me no matter what. Thank you so much for taking care of us and also for always having a smile even in rough times.

Acknowledgements

I would like to thank Mme. AKLI-ASTOUATI Karima, Professor at the University of Science and Technology Houari Boumedien, who supervised me throughout this thesis and who shared with me her intuitions. I would also like to thank her for her kindness, her availability and for her encouragements. Special thanks to Mme. BENAMARA-ZITOUNE Farah, an HDR at the Toulouse University and a researcher at the IRIT laboratory. This thesis is the fruit of a collaboration of more than two years. She welcomed me to her team and guided me the whole time we spent together, whether at the IRIT lab or at distance.

I would like to thank Mme. KHELLAF-HANED Faiza, Professor at the University of Science and Technology Houari Boumedien, for the honour she has given me by accepting to be the president of my jury. She not only introduced me to the semantics theory when I was one of her students, but she also gave me a lot of advice and pushed me to be more serious in class.

I would like to express my gratitude to Mr. AZZOUNE Hamid and Mr. BOUKHALFA Kamel, Professors at the University of Science and Technology Houari Boumedien, who kindly accepted to be examiners of this thesis. Both whom I appreciate deeply. Mr. AZZOUNE Hamid, not only introduced me to the problem resolution and robotics theory when I was one of his students, but he was also my supervisor once. He also gave me a lot of advice on how to start the postgraduate course. I thank him very much for this and for been present in my academic work till this day.

I am also grateful to Mr. KAZAR Okba, Professors at the University of Biskra, and Mme. FAREH Messaouda, MCA at the University of Blida, who also kindly agreed to be examiners of my thesis.

I would also like to thank Mr. Mohand Bughanem for all his contributions and help. Thank you.

Finally, I would like to express my sincere gratitude to all of those who have contributed in any way to the accomplishment and success of this work.

Leila Moudjari
RIIMA Laboratory, USTHB, Algiers, Algeria.

Abstract

Embeddings are very popular representations that allow computing semantic and syntactic similarities between linguistic units from text co-occurrence matrix. These units can vary from character n-grams to words, including more coarse-grained units such as sentences and documents. Recently, multi-level embeddings combining representations from different units have been proposed as an alternative to single-level embeddings to account for the internal structure of words (i.e., morphology) and help systems to generalise well over out of vocabulary words. These representations, either pre-trained or learned, have shown to be quite effective outperforming word-level baselines in several NLP tasks such as machine translation, part of speech tagging and named entity recognition. Our aim here is to contribute to this line of research proposing for the first time in Arabic NLP an in-depth study of the impact of several subwords configurations ranging from character to character n-grams (including word) for social media text classification. Indeed, in this thesis, we aim to investigate the effectiveness of different vectorial representations for the research problem of Opinion and Emotion Analysis (OEA), which attempts to capture both semantic and sentiment information encoded in user-generated content such as tweets. Therefore, we propose several neural architectures to learn character, subword and word embeddings, as well as a combination of these three levels, exploring different composition functions to obtain the final representation of a given text. To evaluate the effectiveness of these representations, extrinsic evaluations were performed on different OEA text classification tasks, namely sentiment analysis, emotion detection and irony detection, while accounting for different Arabic varieties (Modern Standard Arabic, dialects (Levantine and Maghrebi)). This research investigates the effectiveness of these vector representations through extensive experiments over different text data (tweets, blog posts, Facebook comments). For each task, well-known datasets, dialect-agnostic and dialect-specific datasets, were experimented with, including those which have been recently used in shared tasks to better compare the results with those reported in previous studies on the same datasets. The results show that the proposed multi-level embeddings outperform current static and contextualised embeddings as well as best performing state of the art models in sentiment and emotion detection. In addition, these models achieve competitive results in irony detection. The proposed models are also the most productive across dialects observing that different dialects require different composition configurations. Finally, it is shown that these performances tend to increase when coupling the multi-level representations with task-specific features.

Résumé

Les embeddings sont des représentations populaires permettant le calculer des similarités sémantiques et syntaxiques entre les unités linguistiques à partir de la matrice de cooccurrence du texte. Les unités peuvent varier de n-grammes de caractères aux mots, y compris les unités à grain plus large telles que les phrases et les documents. Récemment, les embeddings à plusieurs niveaux combinant des représentations de différentes unités ont été proposés comme alternative aux embeddings à un seul niveau afin de tenir compte de la structure interne des mots (la morphologie) et d'aider les systèmes à généraliser bien au-delà des mots hors vocabulaire. Ces représentations, qu'elles soient pré-entraînées ou apprises, se sont révélées très efficaces pour surpasser les niveaux de référence au niveau des mots dans plusieurs tâches du traitement automatique des langues, telles que la traduction automatique, l'étiquetage morpho-syntaxique et la reconnaissance des entités nommées. Notre objectif ici est de contribuer à cet axe de recherche en proposant pour la première fois en TAL arabe une étude approfondie de l'impact de plusieurs configurations de sous-mots allant

du caractère aux caractère n-grammes (y compris le mot) pour la classification des textes des médias sociaux. En effet, dans cette thèse, nous visons à étudier l'efficacité de différentes représentations vectorielles pour la problématique de recherche de l'Analyse des Opinions et des Emotions (AOE), qui tente de saisir à la fois les informations sémantiques et les sentiments encodés dans les contenus générés par les utilisateurs tels que les tweets. Dans cette optique, nous proposons plusieurs architectures neuronales pour apprendre les embeddings des caractères, des sous-mots et des mots, ainsi qu'une combinaison de ces trois niveaux, en explorant différentes fonctions de composition pour obtenir la représentation finale d'un texte donné. Pour évaluer l'efficacité de ces représentations, nous effectuons des évaluations extrinsèques sur différentes tâches de classification des textes du AOE, à savoir l'analyse des sentiments, la détection des émotions et la détection de l'ironie, tout en variant les dialectes, notamment le levantin et le maghrébin, en plus de l'arabe standard moderne. Nous étudions l'efficacité de ces représentations vectorielles par des expériences approfondies sur différentes données textuelles (tweets, billets de blog, commentaires facebook). Pour chaque tâche, nous expérimentons avec des ensembles de données bien connus, y compris ceux qui ont été récemment utilisés dans des tâches partagées afin de mieux comparer nos résultats avec ceux rapportés dans des études précédentes sur les mêmes jeux de données. Les résultats montrent que nos représentations à plusieurs niveaux sont plus performantes que la représentation au niveau des mots ainsi que les modèles de l'état de l'art en ce qui concerne la détection des sentiments et des émotions et qu'elles obtiennent des résultats compétitifs en ce qui concerne la détection de l'ironie. Les modèles proposés sont également les plus productifs dans l'ensemble des dialectes, en observant que les différents dialectes nécessitent des configurations de composition différentes. Nous montrons enfin que ces performances ont tendance à augmenter lorsque l'on associe les représentations à plusieurs niveaux à des caractéristiques spécifiques à la tâche.

Keywords

Opinion Mining, Social Media, microblog, machine learning, features, feature selection, deep learning, Embedding, dense vectors, multi-level embeddings, Convolutional Neural Networks, Arabic dialects, sentiment analysis, emotion detection, irony detection, Crowdsourcing annotation platform, Algerian dialect, NLP, subjectivity lexicon, corpus creation, linked data.

Contents

Contents	1
List of Tables	4
List of Figures	6
List of Abbreviations	7
1 Introduction	9
1.1 Web, Social Media, a Data Pool	9
1.2 Arabic and its Varieties	12
1.3 Algerian Dialects Characteristics	14
1.4 Data representations	16
1.5 Contributions	17
1.6 Thesis Outline	19
I State of the Art	21
2 Opinion and emotion analysis: A Fascinating Problem	23
2.1 Introduction	23
2.2 Opinion and Emotion Analysis	25
2.2.1 Sentiment Analysis Applications	25
2.2.2 Different Levels of Analysis	26
2.2.3 Different Tasks	27
2.2.4 Cross-Language Sentiment Classification	28
2.2.5 OEA Approches' Classification	29
2.2.6 Feature Selection	30
2.3 Main Existing Resources for Arabic OEA	32
2.3.1 Datasets	32
2.3.2 Lexicons	34
2.4 Proposed Approaches for Sentiment Analysis	34
2.5 Proposed Approaches for Emotion Analysis	38
2.6 OEA Limitations	40
2.7 Conclusion	41
3 Linked Data	43
3.1 Introduction	43
3.2 Linked Data	43
3.3 Linked Data and the Analysis of Opinions and Emotions	45
3.4 Tools	47
3.5 Relevant Projects	48

3.6	CONCLUSION	51
4	Vectorial Representation	55
4.1	Introduction	55
4.2	Word-Level Models	56
4.3	Morpheme-Based Models	59
4.4	Character N-Grams Models	60
4.4.1	Pre-trained Models	60
4.4.2	Learning Subword Embeddings	61
4.5	Arabic Opinion and Emotion Analysis	65
4.6	Conclusion	67
II	Contribution	71
5	An Algerian Corpus and an Annotation Platform for Opinion and Emotion Analysis	73
5.1	Introduction	73
5.2	Data Collection	75
5.3	The Annotation Platform: TWIFIL	75
5.4	The Generated Annotated Corpus: Twifil	78
5.5	The Generated Annotated Lexicon: Twilex	79
5.6	Conclusion	79
6	Multi-Level Embeddings for Processing Arabic Social Media Contents	81
6.1	Introduction	81
6.2	Leveraging Pre-trained Embeddings at Different Unit Levels	84
6.3	Dialect Specific Language Models	85
6.4	CNN Architectures for Learning Multi-Level Embeddings	87
6.4.1	Methodology	87
6.4.2	General Architecture of the CNN	88
6.4.3	EMBCNN _v : Learning Single Unit Embeddings	92
6.4.4	Constructing Multi-Level Embeddings	93
6.4.5	Enriching the Embeddings with Additional Features	94
6.4.6	Summary of the proposed embeddings	96
6.5	Conclusion	96
7	Results and Discussion	99
7.1	Introduction	99
7.2	Data Pre-treatment and Methodology	100
7.3	Multi-Level Embeddings for Sentiment Analysis	101
7.3.1	Datasets	102
7.3.2	Results for the Algerian Dialect	103
7.3.3	Results for the Other Dialects	107
7.3.4	Comparison with State of the Art Results	109
7.4	Multi-Level Embeddings for Emotion Detection	110
7.4.1	Datasets	111
7.4.2	Results	111
7.5	Multi-Level Embeddings for Irony Detection	112
7.5.1	Dataset	113
7.5.2	Results	113

7.6	Discussions and Error Analysis	114
7.6.1	Main Conclusions	114
7.6.2	Error Analysis	115
7.7	Conclusion	116
CONCLUSION GÉNÉRALE ET PERSPECTIVES		121
Bibliography		125

List of Tables

1.1	Examples of Algerian dialect terms.	14
1.2	Which Number Replaces Which Letter? (ar) refer to Arabic and (fr) French.	15
2.1	Exploiting machine techniques learning for OEA.	30
2.2	Main Existing Corpora for Arabic OEA.	33
2.3	Main Existing Lexicons for Arabic OEA.	34
3.1	EmotionML example.	48
3.2	Limitations and suggestions of opinion and emotion analysis' projets.	52
4.1	Summary of main existing subword models. <i>D</i> means that a given reference treats Arabic dialects without targeting a specific dialect while <i>A</i> stands for MSA.	63
5.1	An excerpt of the generated corpus via TWIFIL.	78
5.2	Sentiment characteristics of the corpus.	78
5.3	Emotion characteristics of the corpus.	78
5.4	An excerpt of the generated lexicon via TWIFIL.	79
6.1	AlgSet : The Algerian dialect dataset for building word embeddings.	85
6.2	Different unit permutations when learning multi-level embeddings.	88
6.3	EMBCNN _U experimental settings.	93
6.4	MGCNN _{LEVEL} experimental settings.	94
6.5	Summary of the different embeddings proposed in this study.	96
6.6	Summary of state of the art language models we compare with.	97
7.1	Sentiment analysis datasets used in our experiments. Min(Max) length refers to the minimum(maximum) number of words per instance (tweet, post, review).	103
7.2	Baseline for the SA task	104
7.3	Lexicon-based results of the SA task	104
7.4	Sentiment analysis results on the Algerian dialect datasets when evaluated on a binary classification task (positive vs. negative). ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model per dataset.	106
7.5	Sentiment analysis results on other Arabic dialect datasets when evaluated on a binary classification task (positive vs. negative). ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model per dataset.	108
7.6	Percentage of shared vocabulary between the train and test sentiment datasets and out of vocabulary rates as given by ALG2VEC _w (OOV _w) and EMBCNN _{sw} (OOV _{sw}).	109
7.7	Comparison with state of the art results on the same datasets.	110
7.8	Emotion datasets used in the experiments carried out.	111

7.10	Percentage of shared vocabulary between the train and test emotion datasets and out of vocabulary rates as given by ALG2VEC_w (OOV_w) and EMBCNN_{sw} (OOV_{sw}).	112
7.9	Results of the emotion detection experiments when evaluated on a multi-class classification task (8 classes). ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model per dataset. .	117
7.11	Tweets distribution in the IDAT dataset.	117
7.12	Results of the irony detection experiments when evaluated on the IDAT dataset. ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model for the whole dataset.	118
7.13	Percentage of shared vocabulary between the train and test irony datasets and out of vocabulary rates as given by ALG2VEC_w (OOV_w) and EMBCNN_{sw} (OOV_{sw}).	118
7.14	Examples of the predicted classes in each dataset as given by the best performing models and their associated gold label.	119

List of Figures

1.1	An overview of the web expansion from the web 1.0 to the web 5.0.	10
1.2	The Arab world's map.	12
1.3	A map of spoken Arabic dialects.	12
4.1	CBOW and Skip-gram models architecture.	57
5.1	A general architecture of our work for OEA of the DALG.	74
5.2	TWIFIL platform.	76
5.3	Login to TWIFIL.	76
5.4	Choose a category.	76
5.5	Annotate: choose a label.	76
5.6	Example of polarity annotation interface of the proposed platform.	77
5.7	Example of emotion annotation interface of the proposed platform.	77
6.1	Double-CNN architecture for embedding generation and text classification .	89
6.2	Multi-group CNN for combining embeddings from different units.	94
6.3	Enriching embeddings with additional features at the CNN dense layer. . . .	95
7.1	The evolution of the classifier's performance after each pre-processing step .	101

List of Abbreviations

ALGD	Algerian dialects
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BOW	Bag Of Words
CBOW	Continuous Bag-Of-Words
CNN	Convolutional Neural Network
DA	Arabic Dialects
DA	Dialectical Arabic
DL	Deep Learning
DT	Decision Tree
EA	Emotion Analysis
ED	Emotion detection
GloVe	global representation of word vectors
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long Short-Term Memory
ME	Maximum Entropy
ML	Machine Learning
MLP	multi-layer perceptron
MNB	Multinomial Naïve Bayes
MSA	Modern Standard Arabic
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NMT	Neural Machine Translation

NN Neural Network

OEA Opinion and Emotion Analysis

OOV Out-Of-Vocabulary

POS part-of-speech

RF Random Forest

RNN Recurrent Neural Network

SA Sentiment Analysis

SG Skip-Gram

SSA Subjectivity and Sentiment Analysis

SVM Support Vector Machine

TF*IDF Term Frequency-Inverse Document Frequency

Word2Vec Word2Vec

Chapter 1

Introduction

Since the beginning of time, people have communicated using verbal languages as a means of facilitating interactions. Thereafter, people began to communicate through texts (postal letters, telegraphs). However, communication is a form of sharing ideas, opinions and thoughts. People are known for their curiosity towards other peoples' sayings and thoughts. When listening to or reading another's statement the mind receives more information than that which is seen or heard. Indeed, through texts, opinions are shared which convey thoughts and feelings about specific subjects.

A human being is built to live in a society, of which each person can be and is a subject influenced by the opinions of others. This is due to what a person hears or reads not only influences behaviours but also affects the decision-making and how things are perceived.

Human reactions and interactions with his surroundings differ from person to person, due to how the human brain perceives and translates the knowledge it receives. As Lazarus argued in [Lazarus, 1999], there are three concepts of mind, namely cognition, emotion and motivation. Commonly a person's opinion is blurred with his/her emotional state. It can be said that impressions are what arouse an emotion within and vice versa, as emotions control most of our impressions.

Opinion and emotion analysis is the science of extracting and analysing sentiments in a text, image or other data formats. This is quite a recent and active field of research and thus there is still so much more to discover. Due to social media proliferation, researchers started showing a growing interest in the early 2000s. Hence, it has become one of the Natural Language Processing's most active research areas. It was quickly sought after in the data mining field, web mining and information retrieval [Liu, 2012a]. The Natural Language Processing (NLP) community claims that emotions are briefer than sentiments and that opinions are personal interpretations of information and are not necessarily as emotionally charged as sentiments [Sánchez-Rada and al, 2016]. However, the most robust systems infer new knowledge from such a rich source of data.

1.1 Web, Social Media, a Data Pool

People have an insatiable need to share and communicate their experiences and life events, which led to the invention of what is called "social media" since communications by the telephone were extremely expensive for long conversations not to mention long-distance communications.

The roots of Social media can be traced as far back as November 1994 with the appearance of one of the Internet's earliest social networking websites "GeoCities", followed by "Classmates" in December 1995 and SixDegrees in May 1997. The latter is widely considered to be the very first social networking site, as it included profiles, friends lists and school

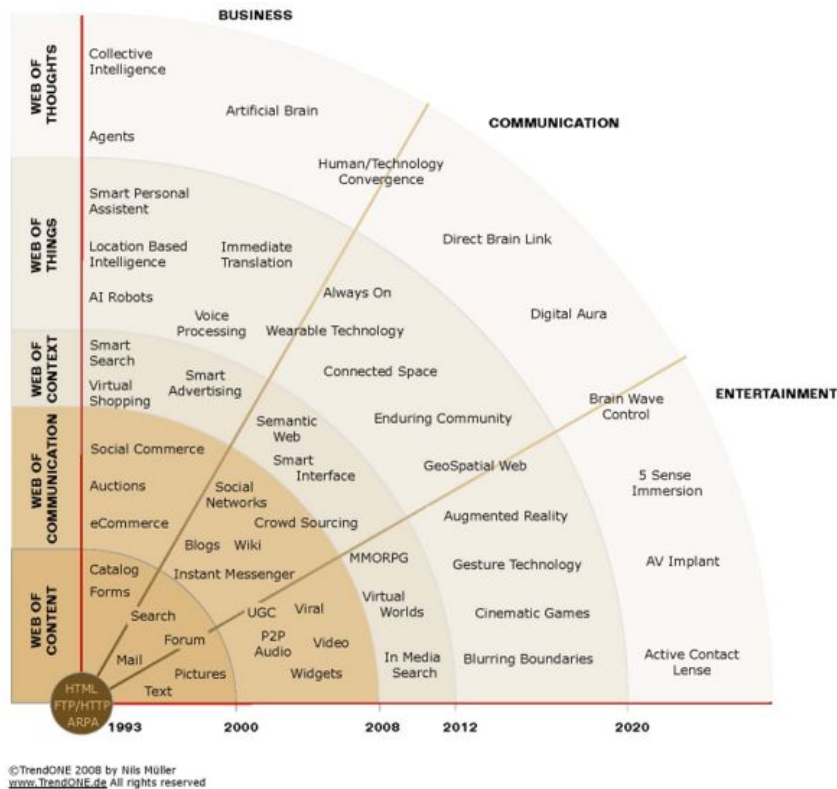


Figure 1.1: An overview of the web expansion from the web 1.0 to the web 5.0.

affiliations which could be used by registered users ¹. These social networking platforms are also known as the "social web" or the "web 2.0". The users of the platforms could share different types of data: text, images, videos and vocal messages.

The web is the largest transformable-information construction, firstly introduced by Tim Burners-Lee in 1989 [Berners-Lee et al., 1992]. Considerable progress has been made during the past two decades in developping the web and related technologies in (Figure 1.1²). Web 1.0 referred to as a web of cognition, information, or percipience, web 2.0 as a web of communication or verbalisation, web 3.0 as a web of co-operation or affiliation, web 4.0 as a web of integration and Web 5.0 as the web of a decentralised smart communicator or the sensory and emotive Web, which is designed to develop computers that interact with human beings. These are introduced such as five generations of the web since the advent of the web [Aghaei et al., 2012, Patel, 2013, Benito-Osorio et al., 2013].

The web 3.0 is also a web of interlinked datasets and new inferred knowledge, which is also called "Linked Data" and "semantic web". Web 3.0 aims to decrease human tasks and decisions and leave them to machines by providing machine-readable content on the web. In general, web 3.0 is included in two main platforms, semantic technologies and social computing environments. The semantic technologies represent open standards that can be applied on the top of the web. The social computing environment allows human-machine co-operations and the organisation of several social web communities[Aghaei et al., 2012].

Semantic markup refers to the communication gap between human web users and comput-

¹<https://www.cbsnews.com/pictures/then-and-now-a-history-of-social-networking-sites/3/>

²source: <https://flatworldbusiness.wordpress.com/flat-education/previously/web-1-0-vs-web-2-0-vs-web-3-0-a-bird-eye-on-the-definition/> visited 18/03/2020

erised applications. *"One of the largest organisational challenges of presenting information on the web was that web applications were not able to provide context to data, and, therefore, did not really understand what was relevant and what was not"*. While this is still evolving, this notion of formatting data to be understood by software agents leads to the "execute" portion of the process, and provides a way to discuss web service ³.

Unlike the semantic web, the social web is considered the number one source for subjective data, where the data is not well written nor structured. The web 2.0 interlinks people from different backgrounds and allows them to share thoughts and express opinions about their interests and disinterests. Social networks have played an important role in communicating and sharing human information since day one. Their content often transmits information about the emotional state of the author, his or her judgments and assessments of a certain person or subject, or premeditated emotional communication (the emotional effect the sender wants to have on the receiver). Message receivers can sense senders' emotions through verbal cues such as emotional words and language markers (e. g., lexical or syntactic coding of emotions), as well as non-verbal paralinguistic cues (e. g., emoticons, feedback, reactions or sharing their messages), which falls under the category of visual cues. The advent of the web 2.0 gave people more freedom and more platforms to express themselves [Arti Buche, 2013].

The data present in these platforms created a very large pool of opinionated and exploitable data. This led to the emergence of techniques that have proven their effectiveness in customised opinion and emotion modelling related to applications such as NLP and automatic learning.

The Linked Data technology tries to offer a more sustainable solution. Its goal is a better interoperability and a single data encoding. This is translated by the use of different ontologies, one of the latest that tackled the Opinion and Emotion Analysis (OEA) problem is Onyx [Sánchez-Rada and Iglesias, 2016] used in two large European projects: EuroSentiment [Sánchez-Rada et al., 2014] and MixedEmotions⁴ (cf. Chapter 2.7).

The advent of the social web produced a crescent interest for opinion and emotion analysis, along with the appearance of new challenges. Nowadays, more affable technologies are talked about which can understand the complexity of human emotions. It is a fact that the Web's evolutionary process has changed business orientation and online education as mentioned in [Benito-Osorio et al., 2013]. It suggests that emotive technologies will be in higher demand (indeed they are). Therefore, machines must be able to process human speech spoken and written in a more advanced way, whether it is formal or informal. As well as take into consideration the unstructured and non-standardised nature of social media texts. This is due to online communications via short messages, often using non-standard language variations. Analysing this type of texts is a challenging task. Especially when dealing with short posts or tweets.

Indeed, currently, there are more than 4 billion Internet users worldwide. More than 50% of the North African population has access to the Internet. The region has also seen a growth of more than 17% in the number of social media users compared to 2017 ⁵. In Algeria, more than 50% of the population are registered users on different social platforms and around 46% use mobile devices for such activity ⁶. These numbers represent a growth of 17% in social media use and more than a 19% growth in the use of mobile devices for such platforms. Twitter users in Algeria reached 18.23% in February 2021 compared to February

³<https://flatworldbusiness.wordpress.com/flat-education/previously/web-1-0-vs-web-2-0-vs-web-3-0-a-bird-eye-on-the-definition/> visited 18/03/2020

⁴<https://mixedemotions-project.eu/linked-emotions-data/>

⁵<https://wearesocial.com/blog/2018/01/global-digital-report-2018> (visited on 23rd, November 2019).

⁶<https://www.slideshare.net/wearesocial/digital-in-2018-in-northern-africa-86865355>

2020 (3.3% and 8.76% in December 2020). This number has almost sextupled over a year, making Twitter the third most used platform by active social media users ⁷.

On social media platforms, 76% of users express their sentiments by clicking corresponding buttons when available, such as "Like", "Dislike" or using "emoticons", "emojis" or "smileys". Across the Arab region, more than 30% of the users use Arabic script and 26% uses Latin script (mostly English and French) and about 15% combine both [Salem, 2017]. Compared to other Arabic dialects, the North African dialects have other peculiarities, as several languages are used in everyday conversations. For example, the expression "*Nro7o ensemble?*" is a combination of the French word "*ensemble*" meaning "*together*", and the Arabic word "*nro7o* (نرؤحو)", meaning "*we go*".

Over the past two decades, opinion and emotion analysis has achieved significant progress, where the highest proportion of available resources and research publications are dedicated to English. The Arabic language is considered an under-resourced language, especially the dialectal Arabic. Therefore, this research target Modern Standard Arabic (MSA) as well as different Arabic dialects (Levantine and Maghrebi dialects) with a focus on the Algerian dialect. In the next section, section, different Arabic varieties are presented.

1.2 Arabic and its Varieties

Arabic is the fifth most spoken language in the world with more than 315 million native speakers (of different different Arabic varieties) ⁸ across the Arab world (Figure 1.2).



Figure 1.2: The Arab world's map.

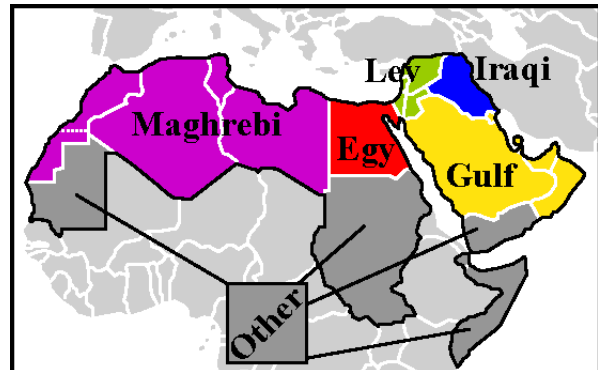


Figure 1.3: A map of spoken Arabic dialects.

Arabic is considered as one of the most morphologically and phonologically rich languages [Meftouh et al., 2012]. Indeed, there are three varieties of Arabic: the classical written Arabic (used in the Holy Quran (Islam's Holy Book) and old literature), MSA or written Modern Standard Arabic (used nowadays in schools, official documents and formal communications), and lastly, the most used in daily conversations and on social media is the "Dialectal Arabic (DA)" [Boudad et al., 2018a]. Figure 1.3, shows a categorisation of 6 different Arabic dialects: Maghrebi (Moroccan, Algerian and Tunisian), Egyptian (Egyptian, Sudanese), Levantine (Palestinian, Lebanese, Syrian and Jordanian), Iraqi, and Gulf (Kuwait, Bahrain, Qatar, the United Arab Emirates, Saudi Arabia and Oman)⁹. In addition to emoticons and emojis, the use of the three varieties of Arabic complicates texts analysis even further [Zaidan and Callison-Burch, 2014]. Indeed, Arabic texts are often characterised by non-diacritised¹⁰ texts

⁷<https://gs.statcounter.com/social-media-stats/all/algeria> (visited on 30th, March 2021).

⁸<https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world> visited 19/03/2020

⁹<https://discoverdiscomfort.com/arabic-dialects-maghrebi-egyptian-levantine-gulf-hejazi-msa/>

¹⁰Diacritics are dedicated letters to represent short vowels.

and large variations of unstandardised dialectal Arabic (among Egypt, Gulf, Levantine and Maghrebi) and linguistic code switching between modern standard Arabic (MSA) and the dialects. Arabic social media are also characterised by a large variation of unstandardised texts, where speakers use Arabic letters (see examples (1) and (2)) but also arabizi [Yaghan, 2008], the informal Arabic chat alphabet in which words are written in their transliterated form using Latin characters and numbers that replace some letters/syllables (see examples (3) and (4) respectively).

1. The word طَاوَلَة (*table*) has two plurals in Algerian dialect (طَابَلَات (*tablat*) and طَوَابِل (*twabel*)) while it has one in MSA (طَاوِلَات (*tawilat*)).
2. The word باي باي written in Arabic refers to the English expression "bye-bye".
3. "Win raho ktabi" (*Where is my book*). This is the Latin script version of "وِين رَاهُو كِتَابِي".
4. "3andi" written in Arabic as "عَنْدِي" (*I have*) where the number 3 replaces the letter "ع".

In addition, a word in MSA can correspond to different words depending on the dialect, involving either intra-word switching [Sankoff and Poplack, 1981] where a switch could occur in one or more places in the same word (cf. (1)) or significant lexical changes that go beyond pronunciation, as in (2). In addition, within each dialect, significant variations due to regions may occur (cf. (3)). Finally, dialectal Arabic is also characterised by linguistic code-switching between MSA and the dialects and between Arabic and other languages like English and French [Bentahila and Davies, 1983, Abu-Melhim, 1991], as in the expression "A3tini la serviette" which is a combination of the French word "*la serviette*" (*towel*), and the transliterated Arabic word "*a3tini*" (أَعْطِنِي), meaning *give me*.

1. The Arabic word فَجَّان (*cup*) becomes فَجَّال in Maghrebi dialect after changing the last letter from "ن" (*n*) to "ل" (*l*).
2. The word سَيَّارَة (*sayara*) (*car*), can be translated into عَرَبَة (*araba*) in Egyptian, سَيَّارَة (*siyara*) in Levantine, كَهْرَبَا (*kahreba*) in Tunisian, طُونَوِيل (*tonobil*) or لَوَطُو (*loto*) in Algerian, and طُومُوِيل (*tomobil*) in Moroccan dialect.
3. The word اَنْزَلَ (*I go down*) corresponds to three different words in Algerian dialect depending on the region: نَحْدِر (*nheder*) in the East, نَهْوَد (*nhawed*) in the West and نَحْبِط (*nahbet*) in the North.

Although MSA has been the target language in most opinion analysis research for Arabic, recent efforts extend the focus to other Arabic dialects (Palestinian [Abdul-Mageed and Diab, 2012], Egyptian [Shoukry and Rafea, 2012a], Tunisian [Medhaffar et al., 2017]). However, works on the Algerian dialect are still scarce.

North African dialects contain words formed from different languages and are still considered under-resourced languages. Therefore the linguistic situation becomes more complicated as there is no dialect derived solely from Arabic (monolingual dialects), but rather multilingual dialects.

Table 1.1: Examples of Algerian dialect terms.

Term	Origin	Pronunciation	Explication
Cozina	Spanish	Cocina	Kitchen
فَنجَال	Arabic	فَنجَان	Cup
وَلَادِ الْجَاجِ, عِظَامِ, بِيض	Arabic	بِيض	Egg
تَاس, طَاس, كَاس	French	Tasse	Glass
جَزَايَا, مَطْرَح	French	Matelas	Mattress

The major problems faced when dealing with Arabic dialects is the lack of resources, the complexity of the language morphology and specific characteristics of each dialect [Baly et al., 2017b, Üstün et al., 2018]. In the next section we highlight specific Algerian Arabic characteristics and common ones with other North African and Arabic dialects.

1.3 Algerian Dialects Characteristics

Algerians have always known how to make the most out of the cultural mix and the heritage bequeathed by all civilisations passing through Algeria [Moudjari et al., 2021]. Indeed, the Algerian dialect is one of the richest languages in the world, and it is constantly being enriched. Nowadays, the publications found on social networks contain not only Arabic, Algerian Arabic, but also: Turkish, Italian, Spanish, French, Korean, Japanese and increasingly English. Moreover, there are not only Western languages, but also Arabic dialects such as Egyptian and Levantine.

Algerian vocabulary shares common words and expressions with other Arabic dialects. This is particularly salient for loan words which are words that have been altered due to code-switching or used as they are when employing other dialects on social media communications, such as the Egyptian expression *سِتِين دَاهِيَة* (calamity) or the Levantine expression *اهلين* (hello or welcome) that are often employed in Algeria.

The Algerian Arabic or Algerian dialect (Darja) is, therefore, less normalised and standardised compared to MSA and other Arabic dialects [Harrat et al., 2014, Mataoui et al., 2016, Harrat et al., 2017]. It has a vocabulary inspired by Arabic, although the original words have been altered phonologically and morphologically [Meftouh et al., 2012]. Algerians also use Tamazight, the original language of the first inhabitants of the region and the second official language of Algeria, which is also divided according to regions, for example, Kabyl, Chaoui, Mzabi and Tergui. Take the expression "I am coming down", in Kabyl it is "Adsoubagh" and in Mzabi "adwatigh". More than 99% of Algerians have Tamazight and ALGD as their native languages, where 73% of the country's population speaks ALGD, and 27% speaks Tamazight ¹¹. The Algerian language is characterised by non-standardised spellings and spelling variations based on the phonetic transcription of many local variants. Table 1.3 highlights some examples of words from different origins.

It is practically the same case for Tunisian and Moroccan. However, the following properties are not specific to the Algerian dialect.

- *Code-switching*: North Africans alternate between two or more languages, or language

¹¹<https://www.worldatlas.com/articles/what-languages-are-spoken-in-algeria.html> visited on the 14th march, 2019

varieties, in the context of a single conversation. This is illustrated, in the following example: "A3tini la serviette (give me the towel)" in which the user used the Arabic word "اعطني" and the French word "la serviette". However, the Algerian dialect is also formed by transformed words from the languages which have inspired Algerians through the ages. Take the word "فنجال" which was inspired from the Arabic word "فنجان (a cup)", where the last letter was changed from "ن" to "ل", or طاس in Algerian and كأس in MSA (a glass) This phenomenon is known as "Intra-word switching" in linguistics [Sankoff and Poplack, 1981], where a switch could occur in one or more places in the same word.

- *Encoding language in another language*: either Arabic expressions encoded in Roman letters known as arabizi or foreign expressions encoded in Arabic letters. As an example of arabizi: "khdma kbira", written in Arabic as "خدمة كبيرة (huge task)", or "باي باي (bye-bye)" written in Arabic.
- *The combination of the two: code-switching and encoding a language in another one*. "wkass jus", an expression of a mix of an Arabic expression "وكاس (and a glass)" and a French one "jus (juice)".
- *The use of numbers instead of letters or words*: this phenomenon has been observed with the proliferation of mobile phones and the social web, where users started to use more and more abbreviations. Since numbers resemble some letters and some syllables, they were used to replace those letters and syllables. Table 1.2 gives examples of the meaning of each number with its use.

Table 1.2: Which Number Replaces Which Letter? (ar) refer to Arabic and (fr) French.

Number	It replaces	Eg : full word == meaning
1	ien (fr)	b1 : bien (fr) == good
2	أ	2men: أمين (ar) == Secure
3	ع	3neb : عنب (ar) == raisins
4	fort (fr)	strong
5	خ	5ali: خال (ar) == uncle
6	ط	6abla == table
7	ح	7ot: حوت (ar) == fish
8	hui	s8: suis (fr) == I am
9	ق	fou9 : فوق (ar) == over

- *Derivatives of the Algerian dialect*. North Africa is well known for having each region speaking a different variety of the dialect. In Algeria, each region is characterised by its own spoken dialect variety. Those from the East and the West speak with totally different accents. Let's take the word "coming", in the East it is pronounced "نحدر" : nheder" in the west "نهود (nhawed)" and in the North "نحبط (nahbet)". Another example is the MSA word "سكر (sugar)", which in Algeria has more than one meaning: sugar/close.

- *Social media language.* Social web users, especially the younger generation, use a lot of emoticons and emojis. Besides abbreviations (already mentioned), social media has its own language. Since emoticons help express emotion in a single character, their use has widely spread. "Hashtags", are used to find, follow, and contribute to a conversation. "Sharing/retweeting" a post is a way of showing support, participating, or even trivialising the post.

Another characteristic of social media chatting is the use of capital letters, the internet code for Yelling and Shouting. In most cases, this is considered rude. In other instances, typing everything in capitals conveys the importance of the text. There are other methods to emphasis a word or a text such as the use of *asterisks* and s p a c i n g words out or even the repetition of letters to emphasise non-verbal signals (joy, anger, etc.). This repetition is also used to overstate comments. For example: yaaaaay, stooooop¹².

- *Common expressions and idioms.* Finally, Arabic as most languages has idiomatic expressions that cannot be understood literally; therefore, they are a source of a semantic difficulty. These common expressions are used in both MSA and different Arabic colloquial dialects. They are mostly used for sarcasm or for suggesting something indirectly or covertly. For instance, "عُمَر أو قهوة" is a way of asking for a bribe, where the exact meaning of the expression "قهوة" is "coffee" and "عُمَر" is a common name.

Above all, there is the possible existence of more than one language in the same sentence, with many possible writing styles, possible writing errors and new words frequently appearing, making the Algerian dialect very difficult to understand and very complex to process automatically. These linguistic diversities call for special attention, which is why the spoken and written dialects are very rich and varied languages¹³.

1.4 Data representations

*"One of the largest organisational challenges of presenting information on the web was that web applications were not able to provide context to data, and, therefore, did not really understand what was relevant and what was not"*³.

Developing meaningful representations of data has been one of the primary goals of computer science since its inception. Computer scientists still struggle to find the best way of representing the data to be treated, to maintain the most relevant features, related either to the task or the context of the data.

In the opinion and emotion analysis field, the literature reviewed showed that standard features (binary, word count, word frequency, TF*IDF, n-grams) are still used and provide acceptable results [Nabil et al., 2015, Al-Khatib and El-Beltagy, 2017, Baly et al., 2019]. While other works exploited more complex features such as dense vectors in the Euclidean space, also known as embeddings [Mikolov et al., 2013c, Soliman et al., 2017].

Embeddings have become very popular representations allowing computing semantic and syntactic similarities between linguistic units from text co-occurrence matrix [Baroni and Lenci, 2010]. Units can vary from character n-grams to words, including more coarse-grained units such as sentences and documents. As highlighted by several authors [Bojanowski et al., 2017, Li et al., 2018], embeddings at word level **suffer from two main issues**. Firstly,

¹²<https://newrepublic.com/article/150506/universal-basic-income-future-of-pointless-work>

¹³All words quoted from the Algerian dialect were given by the authors, who are regular users of the dialect and social media

they ignore the morphology of words treating different forms of the same word as being unrelated vectors. Secondly, **the models are limited to learn vectors for the words present in the vocabulary only**, making NLP systems fail to generalise when the Out-Of-Vocabulary (OOV) set of unseen words is large during the training phase. These limitations are particularly problematic for morphologically rich languages such as Arabic. Indeed, Arabic social media is often characterised by a large variation of unstandardised dialectal Arabic [Zaidan and Callison-Burch, 2014]. Recently, multi-level embeddings combining representations from different units have been proposed as an alternative to single-level embeddings to account for the internal structure of words (i.e., morphology) and help systems to generalise well over out of vocabulary words. These representations, either pre-trained or learned, have shown to be quite effective, outperforming word-level baselines in several NLP tasks such as machine translation, part of speech tagging and named entity recognition.

More recently, a new form of vectorial representations has gained popularity, known as "Transformers" or "contextual embeddings". These are models in which several representations are provided for each word according to the context in which it appears. Most known models are ELMo (Embeddings from Language Models) [Peters et al., 2018a] and Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019]. The main feature of the Transformer is that it uses attention, the concept which helps with alignment in seq2seq architectures for translation, to capture correlations between the words of a sentence, the process is quite similar to the convolutional feature maps extraction.

For the Arabic language, FastText [Bojanowski et al., 2017] where each word is represented as a bag-of-character n-grams, has been widely used: [Erdmann et al., 2018] proposed multi-dialectal word embeddings (Levantine, Gulf and Egyptian) and used it for dictionary induction, [Kim et al., 2018] experimented with word similarity while [Shapiro and Duh, 2018] neural machine translation from Arabic to English. Other studies developed their own neural architectures learning a combination of character and word embeddings (see Chapter 3.6 for an overview). These models are generally evaluated on a single task. Most studies target Middle-East dialects, although few investigate character embeddings for North African dialects, such as Algerian [Adouane et al., 2018].

Representations can either be monolingual or multilingual where units in different languages are mapped to the same area in the latent space [Ammar et al., 2016]. In this thesis, we focus on vectorial representations at different word and character n-grams levels, to capture the semantic aspect of the data for the Arabic language, targeting Modern Standard Arabic and its different dialects emphasising Algerian Arabic.

1.5 Contributions

As far as we know, there has been no analysis conducted on the impact of subword information ranging from character to character n-grams (including word) when varying both the dialects and the classification tasks. We are particularly interested in answering the following research questions: (1) *What are the best vectorial representations for processing Arabic social media textual contents?* (2) *Are their performances stable across different tasks?* (3) *Are their performances stable across different tasks?* (4) *To what extent are these representations dialect dependent?* This last question opens up a fundamental issue regarding transferring knowledge from embeddings trained on a given dialect to tasks involving other dialects, in particular: *Are language models trained on dialect-aware datasets more effective than the ones trained on dialect-agnostic datasets? Do the improvements concern both seen and unseen dialects during training time?*

To this end, we newly propose in Arabic NLP an in-depth study of the impact of several subwords configurations ranging from character to character n-grams (including word) and

explore different composition functions to obtain the final representation of a given text. To evaluate the effectiveness of this approach, extrinsic evaluations were performed on different text classification tasks while varying the dialects. In this thesis, the contribution covers all three levels of a classification task. Namely, resources, data representation and classifiers, which are as follows:

- **A crowd-sourcing annotation platform.** Tools and resources are essential if progress to be made in this field of research. To ensure the credibility of resources, using crowd-sourcing was considered. Hence, an open platform was created for manual annotation which we called "TWIFIL". Where multi-grained annotations were done at both word and tweet level, as well as multi-level annotations including sentiment, emotion and extra-linguistic information (age, gender and topic). Resulting in the largest Algerian dataset for multiple opinion and emotion analysis tasks with multi-labels and multi-regional Algerian dialects.
- **A multi-level representation (MLR) of Arabic texts.** Manual feature extraction is a challenging and time-consuming task, especially in a morphological and phonologically rich language such as Arabic and Algerian dialect. Therefore, various neural architectures to learn character, subword and word embeddings were proposed, as well as a combination of these three levels. In addition, enriched models with surface features (text-length, emoticons, count of positive and negative words and intensifiers) were proposed. All these pre-trained embeddings will be made available for the research community¹⁴.
- **A fair comparison with state of the art pre-trained models across tasks and dialects.** We first compare with pre-trained subword-agnostic embeddings such as Word2vec [Mikolov et al., 2013c], then with well-known pre-trained character and subword embeddings, namely FastText and BPEmb [Heinzerling and Strube, 2018]. We finally compare with contextual embeddings as given by Flair [Akbik et al., 2018] and BERT [Devlin et al., 2019]. We show in this work that MLR are effective for capturing the meaning of Arabic dialects beating current state of the art vectorial representations (both static and contextualised).
- **A systematic evaluation of MLR across *different Arabic varieties*.** We target MSA but also Levantine and Maghrebi dialects using both dialect-agnostic and dialect-aware embeddings in order to investigate whether embeddings learned from a given dialect can be successfully used to classify texts written in other dialects. To evaluate knowledge transferability across dialects, we select Algerian dialect (ALGD) as a source language for two main reasons: its vocabulary contains many words/expressions taken from other dialects and Algerians used to employ other dialects in their daily conversations¹⁵. When compared to competitive baselines, the results show that MLR models presented in this work are the most productive across dialects observing that different dialects require different composition configurations. These results suggest that transferring knowledge across dialects is feasible which opens the door to leveraging dialect-aware embeddings even for dialects that lack pre-trained models.
- **A systematic evaluation of these representations in *three tasks*,** namely sentiment analysis, emotion detection and irony detection. For each task, well-known

¹⁴https://github.com/kinmokusu/oea_algd

¹⁵See for example [Harrat et al., 2016] who studied the dialect spoken in Algiers, the Algerian's capital, showing that it shares different grammatical characteristics with all Arabic dialects including word declension.

datasets were experimented with including those recently used in shared tasks to better compare the results with those reported in previous studies on the same datasets. These results show that the multi-level representations outperform state of the art results in sentiment analysis and emotion detection and achieve competitive results in irony detection, and more importantly, the performances are stable across the three tasks. Finally, it is shown that these performances tend to increase when coupling the multi-level representations with task-specific features.

1.6 Thesis Outline

This thesis is divided into two main parts. The first part is an overview of prior research, which includes opinion and emotion analysis as the first chapter. Then, the second chapter lays out the efforts made in the Linked Data field which are related to opinion mining. The last chapter of this part presents the different syntactic and semantic features used for the opinion and emotion analysis tasks, with a special focus on the different deep vectorial representations proposed.

Part two dedicates its pages to shine the light on the contributions presented in this thesis. Furthermore, part two is also divided into three chapters. The first describes the open platform for crowdsourcing annotation addressing the lack of available resources. The second describes the feature specifications, the data exploited and the methods and algorithms implemented as part of the experiments undertaken in this research. The third, however, details the experiments which were undertaken and gives a detailed error analysis for each task and each corpus. Lastly, this work ends with the final chapter, depicting the conclusions drawn from the evaluation and outlining possible future areas of work.

Part I

State of the Art

Chapter 2

Opinion and emotion analysis: A Fascinating Problem

Social media proliferation allowed people with different backgrounds to share and engage in discussions on various topics. This phenomenon led to what the twenty-first century described as a data explosion. Researchers and marketers sought this opportunity to exploit such data to gain meaningful insights over different domains (e-commerce, politics, cinema).

Over the last two decades, researchers were engaged to assess and analyse the different opinions on the web. Several approaches were proposed, ranging from simple hand-crafted to complex and automated methods, using different types of features. However, the Arabic language and its dialects still suffer from a lack of resources.

This chapter will define the axis of research in the field of OEA, as well as synthesise existing work and list available resources. It will also underline areas of possible contributions and improvements.

Keywords Opinion Mining, Sentiment Analysis, Emotion Analysis, Social Media, microblog, machine learning, features, feature selection, document-level, sentence-level, aspect-level.

2.1 Introduction

Emotions control almost all human activities and are key influencers of their behaviours. Our beliefs and perceptions of reality, and the choices we make, are, to an important level, conditioned upon how we see and interact with our entourage and how much we are influenced by the way others see and evaluate the world. We are therefore accustomed to seeking the advice of others when deciding upon matters. This also applies to organisations and large companies.

Nowadays, such analytics control the marketing strategies, "*what, how, which products and features should I offer my clients?*", is an essential question and the answer can be found in the different social media posts and comments related to the brand and its previous products.

According to dictionaries sentiment, emotion and opinion are synonyms. The Cambridge online dictionary ¹ defines an *emotion* as a strong feeling or sentiment, a *sentiment* as a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something, and an *opinion* as a thought or a belief about something or someone or judgment that we make.

Sentiment analysis is a semantic analysis that aims to extract and analyse subjectivity

¹<https://dictionary.cambridge.org/dictionary/english/>

out of user-generated content on the social web, generally textual sources [Arti Buche, 2013]. It generally assigns a polarity (positive, negative, mix or neutral) to a resource. It is an approach that relies mainly on text mining and semantic analysis to determine the "position" of the individuals studied, concerning a brand or an event. It is based on other elements as well, such as the use of the "emoticons and emojis", likes, interactions, voice analysis or even facial expressions [Ekman, 1999b].

When based on textual searching, sentiment analysis can include verbatim statements from social networks, blogs, microblogs, review sites and forums. However, it is considered vaguer and general since it is limited to labelling the document with one of the three classes (positive, neutral, negative) with respect to someone or something.

Emotions analysis is similar to sentiment analysis. Nevertheless, it goes beyond a binary/ternary distinction. Since it provides the emotional class (neutral, anger, disgust, fear, joy, sadness, surprise). Some sources claim that these are the same [Cambria et al., 2013], others claim that sentiment analysis is a part of the emotion analysis [Liu, 2012a].

However, opinion analysis takes the process even further. It depends not only on the situation in which the opinion is expressed yet also on the person expressing the opinion. Some studies consider all as the same and are called opinion mining, sentiment analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis and review mining. Other studies consider sentiments as the emotional part of opinions. Opinion mining provides automatic analysis of feedbacks and a tool that answers questions such as "What do people think of the product?" [Arti Buche, 2013]. It is also a way to automate and simplify the analysis of feedbacks. It is mainly based on the different resources on the web, since the advent of web 2.0 gave people more freedom and more platforms to express themselves [Arti Buche, 2013].

An opinion expressed could be subjective or objective (neutral). Subjectivity analysis helps determine whether data is subjective or objective. If it is subjective, it means that the opinion carries an emotion. If there is a need to know its polarity, sentiment analysis is performed and if there is a need to know beyond this scope then this enters the realm of emotion analysis. Therefore, the expression "*Opinion and Emotion Analysis*" is used to cover all meanings of the terms used, allowing to detect the subjectivity and emotional state of a subject studied and understand the opinion expressed.

Sentiment analysis and emotion detection in Arabic have been widely studied [Baly et al., 2017b, Al-Smadi et al., 2018, Abo et al., 2018]. Most related works focus on Modern Standard Arabic (MSA), although a few investigated Arabic dialects, such as Jordanian [Atoum and Nouman, 2019, Duwairi, 2015], Egyptian [Shoukry and Rafea, 2012a], Iraqi [Alnawas and Arici, 2019], Levantine [Baly et al., 2019, Qwaider et al., 2019] and Tunisian [Medhaffar et al., 2017]. North African dialects, including Algerian dialects (ALGD) are less normalised compared to MSA. They have been enriched by many languages over the years, which resulted in a complex linguistic situation. Also, a significant lack of resources was found for most of these dialects such as lexicons, dictionaries, and annotated corpora.

The rest of the chapter is organised as follows, the second section is an introduction to the analysis of opinions and emotions. The third section focuses on the public resources available for Arabic and its dialects. The fourth section presents the different approaches of sentiment analysis. Section 5 presents the existing literature on emotion analysis. Before concluding this chapter, section 6 highlights some of the main issues affecting the methods described in previous sections.

2.2 Opinion and Emotion Analysis

Since the early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole [Liu, 2012a].

The first language which received the largest share of research, tools and resources is English [Maas et al., 2011, Poria et al., 2016]). However, other European languages quickly joined such as Spanish [Brooke et al., 2009] and Greek [Kermanidis and Maragoudakis, 2013]. Arabic researchers were not too far behind [Rushdi-Saleh et al., 2011, Shoukry and Rafea, 2012b, Abdulla et al., 2013]. Unlike other languages, Arabic is characterised by its rich morphology and different varieties.

Since its emergence as a powerful data analysis branch, different domains quickly adopted it to help analyse customer's and client's feedbacks and to provide more context to their predictive systems. During the past two decades, several approaches were proposed in the literature, and therefore adopted in different applications that facilitate several daily tasks.

2.2.1 Sentiment Analysis Applications

Since the advent of the internet, data has become crucial. Nowadays, leading web influencers sell the data collected via their platforms. Opinion and emotion analysis has raised the bar even higher by simplifying multiple tasks and offering complementary information such as "why do customers like/dislike my product?". In this section, several are mentioned;

- Huge volumes of opinion-rich data are published on social media at an unprecedented rate, opinion and emotion analysis facilitates the analysis of humanly impossible data sources to analyse [Cheng et al., 2017].
- Decision making, on a personal or business level.
- Rapid analysis of customers' feedbacks for branding or predicting future needs as well as trend analysis with forecasts for future trends [Sánchez-Rada and al, 2016].
- Often used in monitoring social networks, tracking customer attitudes towards brands and politicians [Sánchez-Rada and al, 2016].
- Prediction of poll ratings [Hu et al., 2013].
- Forecasting box office revenues of movies [Asur and Huberman, 2010].
- Spam detection [Peng and Zhong, 2014].
- Predicting election results and analysing public opinion towards different policies [Sobkowicz et al., 2012] and politicians [Hu et al., 2013].
- Detection of violent social movement groups [Zimbra and Chen, 2012] and cyberbullying [Dadvar et al., 2012].
- Analyse information and infer knowledge in different domains, such as business, research and industry. In business, it helps to obtain information about competitors, developing new strategies to conquer the market and what features must be included in the minimum viable products. In research, it helps to understand popular societies and cultures, which can help better understand people's traces on social networks (messages, comments, and publications).

- E-Learning Environments. The analysis of opinions and emotions can be used by adaptive e-learning systems to support personalised learning by taking into consideration the learner’s emotional state and recommending the most appropriate tasks or courses. In addition, students’ feedbacks about a course are useful for teachers.
- Affective user interfaces, for example in the automotive industry. Where human-machine interaction is enhanced by intelligent user interfaces, which are adapted to recognize users’ emotional states and respond to these emotions by adapting to the current situation through an affective modelg. This is intended to improve human-machine interaction and provide more appropriate recommendations.
- Advertising is a popular means of online monetisation. Opinion and emotion analysis is exploited in a variety of ways: 1) placing advertisements in the most appropriate places on social networks to attract more attention, or 2) placing the advertisements in the sight of people criticising a competitor.

The analysis of opinions and emotions could be applied to a greater number of use cases. This has inspired several methods and approaches over the years. From natural language processing to machine learning. These approaches treated different units of a given text (aspect, sentence, document).

2.2.2 Different Levels of Analysis

Online opinions expressed through comments or blogs, can directly affect future sales in a wide range of fields such as hotels, e-commerce, and restaurants [Bravo-Marquez et al., 2014].

Consequently, interest in OEA techniques as a way of automatically extracting and analyzing user-generated opinions has risen.

OEA centers around the classification of sentiments, opinions or attitudes expressed in human-generated texts. To this end, texts can be labelled into several categories, being positive and negative the most common. Since the study of opinions in a text comprises of a broad range of possibilities, where Sentiment Analysis (SA) can be classified into three categories, attending to the granularity level [Zhao et al., 2016]. The literature considers the following categorisation, in increasing order of specificity:

- A. **Feature/Aspect Level**, in which case, a sentiment is associated with a word or a small group of words (e.g., good food). The aspect-level performs finer-grained analysis, which is called feature level, aspect level or entity level. It helps determine the sentiment of entities and/or their aspects. For example, "The Xperia camera takes great pictures, but its battery does not operate efficiently". The example evaluates two *aspects*, camera resolution and battery quality of the *entity* Xperia. The sentiment on the first aspect is positive, but it is negative on the second aspect. The camera resolution and battery quality are the opinion *targets*.

Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses [Liu, 2012a].

- B. **Sentence Level**, where sentiment is bounded to a singular sentence of text. The task at this level goes beyond a feature level and considers several aspects presented in one sentence [Shoukry and Rafea, 2012b]. This level of analysis is closely related to subjectivity classification. However, it should be noted that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions [Liu, 2012a], e.g., "I bought a hairdryer last week, but the cold air button was not working". Other researchers

have also studied sentiments at a clause level [Wilson et al., 2004]. According to [Liu, 2012a] the analysis at this level is not enough since the opinion of the sentence could be divided between its clauses, e.g., "I liked the movie better when it ended".

- C. **Document Level**, as in a movie review that can span one or more paragraphs; the analysis help classify whether a whole opinionated document expresses a positive or negative sentiment [Pang et al., 2002, Farra et al., 2010, Rushdi-Saleh et al., 2011].

The different levels of classifications are highly challenging and consist of several sub-problems and are exploited in different tasks.

2.2.3 Different Tasks

Opinion and Emotion Analysis is a large field. It includes subjectivity, sentiment, emotion, opinion. However, there are other NLP tasks that relied on the OEA information as extra-linguistic features to their predictive systems such as irony and sarcasm detection. However, the resulted prediction could also enhance OEA systems.

In this section, two tasks are presented, which were of particular interest as there was a need to explore their effect on this study.

Age and Gender Detection

The literature of age and gender detection is vast. However, this research is interested in its cross path with OEA. The reader can refer to [Rangel et al., 2014] for a comprehensive overview of the fields.

Nowadays all people are using social media which advanced opinion and emotion analysis. However, the proposed approaches still have some limitations [Guimaraes et al., 2017]. Hence the interest in exploring other parameters that could affect the person's opinion. Social media sites are the major drivers for most of the web's traffic. Consequently, the ability to harness the power of social metadata is a vital skill for Opinion and Emotion Analysis. Such metadata can affect the results and precision rate hugely.

Knowledge of the user can be crucial when it comes to developing complex systems that are capable of being more efficient and eventually adaptable. This knowledge is often based, mostly on the construction of a user profile, defined as a set of features characterising the user, his context and usage.

People do not provide their personal data, such as age, gender, and other demographics. Although OEA can benefit from such data to overcome some of its limitations. Often, applications do not consider the most diverse parameters that can influence the sentiments in a sentence, if the users' profile information and their writing characteristics are considered we can obtain more reliable results.

Gender is reflected in the topical trends of the documents, as well as the stylistic choices and discourse (the use of profane language, Emoticons, and communication verbs). An interesting work [Škrjanec and Pollak, 2016] compares the language of women and men in Slovenian content user-generated, based on stylometric² approaches, which aims to generalise the writing style into measurable templates that can be compared between individual authors or groups of authors. Indeed, NLP has allowed researchers to model the variation linguistics in large text corpora using automated approaches. This work endeavours to detect the user's gender through his/her posts and comments on Twitter, showing that there are typical behaviours among users of the same age group and the same gender. Especially, when these users write about the same topics. The work consists of finding differences between

²Stylometry tries to identify the style of a text, inherent to its author, as well as its era, author's gender.

female and male authors with regard to the subject matter, linguistic style and referential grammatical gender, and whether these variations can be used to automatically identify the gender of the author of Slovene blog entries and tweets based on topic ontologies of male and female entries.

Irony and Sarcasm Detection

The literature on the topic is vast, and it is out of the scope of this thesis to reproduce it. The recent work by [Wallace, 2015] contains a comprehensive state of the art on the topic, as well as an introduction to novel models.

Since people use social media to express their opinions, they often apply linguistic devices such as irony. From the sentiment analysis perspective, such utterances represent a challenge being a polarity reverser, e.g., "The soup that they served was awesome it could be the last thing you would eat". Irony represents an interesting way for opinion communication towards a particular target in social media. Therefore, it affects opinion analysis and is considered one of the hardest tasks in NLP. It is treated as a classification problem, generally binary where a message is classified into ironic vs. non-ironic. It has gained relevance these years particularly due to its importance in improving sentiment analysis [Zhang et al., 2019, Hernández Farías et al., 2016, Rosso et al., 2018].

In recent years, several approaches have been proposed to deal with irony detection in social media ([Jia et al., 2019], [Ghosh and Veale, 2018], [Karoui et al., 2017a]), where classifiers such as decision trees and Support Vector Machine (SVM) give the best results [Hernández-Farías et al., 2015]. Many shared tasks on irony have been proposed in several languages, such as SemEval 2018 Task 3 for English [Hee et al., 2018], DEFT@TALN2017 for French [Benamara et al., 2017a], IronITA@Evalita 2018 for Italian [Cignarella et al., 2018], IroSvA@IberLEF-2019 for Spanish variants [Ortega-Bueno et al., 2019] (from Spain, Cuba and Mexico), and IDAT@FIRE2019 for Arabic [Ghanem et al., 2019]. This last competition was built on the initial work by [Karoui et al., 2017b] who proposed the first irony detection system for Arabic. The attempts in which irony has been addressed in Arabic literature are fewer. According to [Rosso et al., 2018], there are no automatic approaches to detect irony, which is valid for older works, such as [Sigar and Taha, 2012], who manually analysed the similarities and differences between ironic expressions in English and Arabic from books, articles and the internet. However, recent works rely on different features to infer the irony existence such as surface, sentiment, false assertion and exaggeration, [Karoui et al., 2017b] used these features to infer the context needed to detect irony in Arabic social media texts. [Agrawal et al., 2020], on the other hand, split each text into some consecutive non-overlapping chunks, where each chunk was labelled with one of the six Ekman's emotions. Afterwards, an LSTM model classifies each text as being sarcastic or not while taking into consideration the emotional transitions between the text's chunks.

2.2.4 Cross-Language Sentiment Classification

It is the Opinion and Emotion Analysis of multi-lingual opinionated documents. There are generally two motives for such analysis. The first is the lack of resources of a given language. Therefore, researchers performed the translation of datasets or lexicon usually from English to their language [Mohammad et al., 2016b]. The second motive is that there are several multinational companies which want to analyse the feedbacks of their clients in different countries [Liu, 2012a].

The motive in this research is a bit different. Due to the striking differences between the Arabic Dialects, each dialect can be considered as a different language. Therefore, in this

work, the aim is to build robust models capable of analysing documents of different Arabic Dialects.

In the literature, two methods are found which have been used. The first combines different resources from different languages and trains one model using the large corpus created [Ruder et al., 2019, Nguyen and Le Nguyen, 2018]. The second either trains the model on one dataset, of one language at a time and iterates the learning of the model on different datasets or train a model using an annotated dataset of a language and then annotates another dataset of a low resourced language with the model [Dashtipour et al., 2016].

2.2.5 OEA Approches' Classification

Opinion and emotion analysis is viewed as an established field of research in a growing industry [Pang et al., 2008] which has at its center the natural language processing domain and exploits different machine learning techniques. Sentiment analysis systems can be classified into corpus or lexicon-based approaches, machine learning-based approaches and hybrid approaches [Ravi and Ravi, 2015, Dashtipour et al., 2016].

Feature and Lexicon-Based Approaches

In lexicon-based approaches, either dictionary or corpus-based approaches can be used, where sentiment dictionaries are created with or without using ontologies, and corpus approaches rely on probabilities, based on the occurrence of the overall sentiment in a document. Corpus-based methods usually use labelled data [Ye et al., 2009], lexicon-based methods rely on lexicons and optionally on unlabelled data [Wan, 2008]. A sentiment lexicon is a collection of known sentiment terms [Medhat et al., 2014]. The subjective nature of OEA depends on key indicators of sentiment polarity which are sentiment terms. Good, great and brilliant can be considered positive words while bad, worse and disastrous can express negative sentiments towards a subject. Consequently, sentiment lexicons, which gather sentiment words, are used extensively by the research community [Ravi and Ravi, 2015]. They can be organized into three categories [Wu et al., 2013, Liu, 2012a]: (i) those who contain only sentiment words (a list of words), (ii) the ones that are formed by both sentiment words and polarity orientations (a list of words with only positive and negative annotations), and (iii) the lexicons that offer sentiment words with orientation and intensity [Taboada et al., 2011] (a list of words with scalar numerical values).

Deep and Machine-Learning Based Approaches

Machine learning methods can be divided into supervised (SL) and unsupervised methods (UL) or deep learning (DL), whether it is a Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN).

Supervised methods train a robust sentiment classifier, generally from manually labelled training data. The commonly used classifiers for supervised learning are; Decision Tree (DT), SVM, Neural Network (NN), Naïve Bayes (NB), and Maximum Entropy (ME). Unsupervised learning approaches are usually mixed with a supervised learning process, such as a Learning Word Vectors (LWV) for Sentiment Analysis, a mixture of unsupervised and supervised techniques to learn word vectors capturing both semantic and sentiment similarities between words in order to perform a document level polarity classification [Bilal et al., 2016, Maas et al., 2011]. Or, the K-nearest neighbours (KNN) [Bilal et al., 2016]. Others used different sentiment and emotion cues such as emoticons and product ratings [Cheng et al., 2017], by providing unified modelling for two main categories of emotional signals: emotion indication

Table 2.1: Exploiting machine techniques learning for OEA.

SL	DT	[Dixit et al., 2017][Bilal et al., 2016][Nasser et al., 2016]
	NB	[Pak and Paroubek, 2010][Bilal et al., 2016][Shoukry and Rafea, 2012b] [Rahab et al., 2017][Nabil et al., 2015][ElSahar and El-Beltagy, 2015]
	SVM	[Pak and Paroubek, 2010][Nasser et al., 2016][Al-Smadi et al., 2018] [Rahab et al., 2017][Nabil et al., 2015][ElSahar and El-Beltagy, 2015]
	NN	[Cambria et al., 2013]
	ME	[Xie et al., 2017]
UL	KNN	[Bilal et al., 2016][Nabil et al., 2015][ElSahar and El-Beltagy, 2015]
	LWV	[Bilal et al., 2016][Maas et al., 2011]
	Others	[Hu et al., 2013][Al-Radaideh and Al-Qudah, 2017] [Cheng et al., 2017][Al-Smadi et al., 2018]
DL		[Glorot et al., 2011][Kim, 2014][Al Sallab et al., 2015][You et al., 2016] [Poria et al., 2016][Baly et al., 2017b][Al-Smadi et al., 2018]

and emotion correlation [Hu et al., 2013]. Table 2.1, give examples of the methods mostly used, given by supervised learning methods, where SVM proves effective in most cases and gives the best results in some cases.

Hybrid Approaches

Hybrid methods are used based on both labelled data and lexicon resources, optionally with unlabelled data [Sindhwani and Melville, 2008] for learning more useful sentiment analysis resources. Examples include techniques that combine visual and textual Sentiment Analysis of Social Multimedia. [You et al., 2016] proposed a cross-modality consistent regression (CCR) model, utilising both visual and textual techniques, a fine-tuned CNN for image sentiment analysis and a trained paragraph vector model for textual sentiment analysis.

2.2.6 Feature Selection

The goal of feature-selection techniques is to find the most relevant features for a specific task by removing irrelevant, redundant and noisy data. The approaches defined in the previous section use different features ranging from syntactic to linguistic and stylistic writing characteristics including meta-information which can offer some user profiling information. Information gain (IG) and chi-square distribution (chi2) are two well-known algorithms for features extraction. [Zheng et al., 2018] performed a binary SA using 1-4 character n-grams and 1-4 POS n-grams features. They relied on Chi2 and document frequency to extract relevant features. While [Sayadi et al., 2016] relied on IG. [Elouardighi et al., 2017], on the other hand, used the Between Group to within Group Sum of Squares (BSS/WSS) to filter different n-grams (1-2 grams) TF*IDF weights. These features can be classified into:

Statistical Features

In information retrieval, feature vectors have been used to represent a text. The vector's entries correspond to the individual terms in the text. In NLP this vector is called Bag Of Words (BOW), representing a given text. Words are therefore replaced by one of the following configurations:

1. Term presence: a binary-valued feature vector wherein the entries simply show whether a term appears in a text taking the value of 1, or not, taking the value of 0. This

approach was more effective in reviewing the polarity classification than the real-valued feature vectors [Liu, 2012b].

2. Word count: a simple count of the words in the text.
3. Term frequency: the frequency of each word as a ratio of words within each text (formula 2.1).

$$freq = \frac{term\ count\ in\ tweet}{max(freq\ all\ terms\ in\ tweet)} \quad (2.1)$$

4. Term Frequency-Inverse Document Frequency (TF*IDF), a statistic that reflects the importance of a word in a document, in case of this work the corpus. Where each word was replaced by its weight calculated by the formula 2.2.

$$TF * IDF = freq(term_i) * log_e \frac{Number\ of\ tweets}{Number\ of\ tweets\ containing\ term_i} \quad (2.2)$$

Grams are words which are frequently repeated in the corpus and represented with one of these features, where generally unigrams perform better than bi-grams and other n-grams.

Linguistic, Stylistic and Surface features

Several surface features have already been successfully employed in the sentiment analysis literature for the Arabic language [Mohammad et al., 2016a, Al-Twairesh et al., 2018, Al-Twairesh and Al-Negheimish, 2019]. After performing feature selection using the chi-squared test, it resulted in the following list:

- Surface and stylistic features: number of punctuation marks, presence of URLs and user mentions, number of positive/negative emoticons, number of positive/negative emojis and the sum of positive/negative emoticons and emojis.
- Sentiment features. Several available Arabic sentiment and emotion lexicons were used, including³: NRC emotion lexicon [Mohammad and Turney, 2013b], Arabic NRC hashtags, the dialectal Arabic hashtags, the Arabic translation of Bing Liu’s lexicon, the Arabic MPQA subjectivity [Mohammad et al., 2016a], and EmoLex [Alsmearat et al., 2015]. As the Algerian dialect is specifically targeted, finally a manually built lexicon was relied upon, composed of 6,350 positives, 5,160 negatives and 10,090 neutrals (i.e., both positive and negative) Algerian words/expressions [Moudjari et al., 2020]. Given a text, the number of positive/negative words that belong to one of these lexicons is counted.
- Extra-linguistic features: for tweets, the number of followers and the number of retweets is also accounted for.

Semantic Orientation Features

These features are given by embeddings. These representations, either pre-trained or learned, have shown to be quite effective outperforming word-level baselines in several NLP tasks such as machine translation, part of speech tagging and named entity recognition (cf. Chapter 3.6).

In the next section, the different corpora and lexicons proposed for the Arabic OEA tasks will be presented.

³These lexicons can be found here <https://saifmohammad.com/WebPages/ArabicSA.html>

2.3 Main Existing Resources for Arabic OEA

2.3.1 Datasets

There have been many contributions aiming to develop annotated Arabic language corpora, with the main objective of facilitating Arabic NLP applications [Jarrar et al., 2014], several Arabic datasets have been built for sentiment and emotion analysis. Table 2.2 summarises the main proposed datasets, along with the dialects used and the type of annotation. These corpora were generally created manually relying on crowdsourcing⁴ annotation [Duwairi and Qarqaz, 2014, Duwairi, 2015, Mohammad and Kiritchenko, 2018a].

The table is organised into three parts. The first for the Sentiment Analysis corpora, the second for the Subjectivity and Sentiment Analysis (SSA) corpora and the last for the Emotion Analysis (EA) corpora. Each part is sorted by the number of classes in an ascending manner.

For SA corpora, each corpus has a different annotation scale. The SA-n is used to indicate the number (n) of classes. We distinguish from the classical ternary classification (positive(pos), negative (neg) and neutral(neu)) and other ternary classifications by numbering the classes explicitly in the table. The same process is followed for SSA and EA.

Most corpora are not available and those available are marked with (*). Also, manually annotated corpora are labelled with (M) and those automatically annotated by (A). Those with (MA) label refer to corpora created using seeds manually annotated but the corpus was automatically harvested using these seeds. Some corpora were validated automatically by human annotators [Abdul-Mageed et al., 2016].

It can be observe that most corpora for SA deal with different Levantine (Palestinian, Syrian, Jordanian, Lebanese) and Egyptian dialects. Among Maghrebi dialects, we cite [Oussous et al., 2019] for Moroccan, [Medhaffar et al., 2017] for Tunisian and [Moudjari et al., 2020] for Algerian.

Most of these corpora provide an SA-2 annotation but the SA-3 annotations are as popular, wherein some corpora, the third class is "mix", representing texts carrying both positive and negative sentiments such as [Qwaider et al., 2019] and [Ibrahim et al., 2015b], which also provided the polarity intensity (-10 to 10). Although, some corpora present different annotation scales. [Siddiqui et al., 2015] proposed a multilingual corpus (Arabic and English) for hierarchical sentiment analysis of 650 hotel reviews in Saudi Arabic. The corpus contains more than 191k annotated reviews including nominal sentiment labels, sentiment intensity, language, and dialect. The reviews are divided into eleven different categories ranging from Exceptional to Very poor, within each category, reviews are divided into positive and negative categories.

The same behaviour can be observed in the 4-class annotation such as the corpus proposed in [Adouane and Johansson, 2016], where texts are classified into the SA-3, in addition to the "mix" class. [Nabil et al., 2015] on the other hand uses a different scale where texts are classified as objective, subjective positive, subjective negative, and subjective mixed.

The 5-class corpora provide different scales. [Baly et al., 2017b, Baly et al., 2019] used the {very negative, negative, neutral, positive, very positive} sentiment set, and [Rosenthal et al., 2017] used the classical SA-3 set for the first corpus and the {highly positive, positive, neutral, negative, highly negative} for the ordinal corpus.

The SSA corpora use different annotation scales as well. The AWATIF dataset [Abdul-Mageed and Diab, 2012] used two different scales, one for subjectivity (objective(obj), pos, neg, neu) and one for the sentiment (pos, neg, neu and mix). SAMAR corpora [Abdul-

⁴It is a technique that relies on a group of individuals to annotate a given text using a given taxonomy [Zaidan, 2012].

Table 2.2: Main Existing Corpora for Arabic OEA.

Corpus (name/source)	Dialect	Annotation level	Size	Source
OCA[Rushdi-Saleh et al., 2011] ^{MA*}	MSA	SA-2	500	Movie reviews
SIAAC[Rahab et al., 2017] ^M	ALGD(MSA)	SA-2	92	FB
[Mataoui et al., 2016] ^{M*}	MSA+ALGD	SA-2	5,039	FB
[Abdelli et al., 2019] ^{M*}	MSA+ALGD	SA-2	49,864	FB
TSAC[Medhaffar et al., 2017] ^{M*}	Tunisian	SA-2	17k	FB
MEC[Elouardighi et al., 2017] ^M	MSA+Morrocan	SA-2	10,253	FB
ASA[Oussous et al., 2019] ^M	MSA+Morrocan	SA-2	2,000	Tweets
BBN[Mohammad et al., 2016b] ^{M*}	Levantine	SA-2	1,200	SM
[Mohammad et al., 2016b] ^{M*}	Syrian	SA-2	2,000	Tweets
ArTwitter [Abdulla et al., 2013] ^{M*}	Jordanian+MSA	SA-2	2,000	Tweets
[Shoukry and Rafea, 2012a] ^M	Egypt+MSA	SA-2	1,000	Tweets
[Ibrahim et al., 2015a] ^M	MSA+Egypt	SA-2	2,000	Tweets+reviews
[DoniaGamal et al., 2018] ^A	Egypt+MSA	SA-2	151,500	Tweets
[Siddiqui et al., 2015] ^A	Gulf+English	SA 11/2	191,011	Hotels reviews
LABR [Aly and Atiya, 2013] ^{A*}	MSA	SA-3	63,000	Book reviews
OoredooRayek[Klouch et al., 2020] ^M	ALGD+MSA	SA-3	18,141	FB
TEC[Sayadi et al., 2016] ^M	MSA+Tunisian	SA-3	3,043	Tweets
MIKA[Ibrahim et al., 2015b] ^M	Egypt+MSA	SA intensity + SA-3	4,000	Tweets+reviews
[Al-Kabi et al., 2016] ^M	MSA, Egypt, Levant, Iraq, Maghreb	SA-3	1,442	Maktoob Yahoo
[Duwairi, 2015] ^M	Jordanian	SA-3	22,550	Tweets
[Atoum and Nouman, 2019] ^M	Jordanian	SA-3	3,550	Tweets
[Duwairi and Qarqaz, 2014] ^M	Dialectal	SA-3	2,591	Tweets + FB
Elsahar[ElSahar and El-Beltagy, 2015] ^{A*}	MSA/Dialectal	SA-2 + mix	33,000	Reviews (restaurants, products,hotels)
[Qwaider et al., 2019] ^M	Levantine	SA-2 + mix	2,000	Tweets
[Adouane and Johansson, 2016] ^M	Gulf	SA-4 (3+mix)	4,072	Restaurant reviews
ASTD[Nabil et al., 2015] ^{M*}	Dialectal	SA-4	10,006	Tweets
Arasenti-tweet[Al-Twaires et al., 2017] ^{M*}	MSA+Saoudi	SA-4 (3+mix)	17,573	Tweets
ArSAS[Elmadany et al., 2018] ^M	Dialectal	SA-4 (3+mix)	21k	Tweets
[Baly et al., 2017b] ^M	MSA	SA intensity + SA-5	1,177	News comments
ARSENTD[Baly et al., 2019] ^{M*}	Levantine	SA-5	4,000	Tweets
SemEval2017[Rosenthal et al., 2017] ^M	Dialectal	SA 3/5	9,455/4,413	Tweets
AWATIF[Abdul-Mageed and Diab, 2012] ^M	MSA	SSA-4	5,382	Wikipedia Talk Pages/Forums
SAMAR[Abdul-Mageed et al., 2014] ^M	MSA	SSA-3 + objective	3,008	Wikipedia Talk Pages
SAMAR[Abdul-Mageed et al., 2014] ^M	MSA+Dialectal	SSA-3 + objective	3,015	Tweets
SAMAR[Abdul-Mageed et al., 2014] ^M	MSA+Dialectal	SSA-3 + objective	3,097	Forums
[Refaee and Rieser, 2014] ^M	Dialectal	SSA-5	8,868	Tweets
[Moudjari et al., 2020] ^{M*}	ALGD	SA-3/Emotion	9,040/5,044	Tweets
[Al-Aziz et al., 2015] ^{MA}	Egypt	EA-5	1,552	Tweets
[Al-Khatib and El-Beltagy, 2017] ^M	Egypt+Dialectal	EA-8	10,065	Tweets
[Almahdawi and Teahan, 2019] ^M	Iraqi	Emotion(6)	1,365	FB
DINA [Abdul-Mageed et al., 2016] ^{MA}	Dialectal	EA-6	2,984	Tweets
LAMA [Alhuzali et al., 2018] ^{MA}	Dialectal	EA-8	7,268	Tweets
SemEval2018[Mohammad et al., 2018] ^{M*}	Dialectal	SA-4/7 + EA-11	4,381	Tweets

Table 2.3: Main Existing Lexicons for Arabic OEA.

	Lexicon	Total	Positive	Negative	Neutral
1	Twilex [Moudjari and Akli-Astouati, 2019a]	21,600	6,350	5,160	10,090
2	Arabic Bing Liu’s [Mohammad et al., 2016a]	6,789	2,006	4,783	–
3	Arabic MPQA Subjectivity [Mohammad et al., 2016a]	8,199	2,718	4,911	570
4	Arabic NRC Emotion (EmoLex) [Alsmearat et al., 2015]	14,182	2,312	3,324	8,546
5	Arabic NRC Emoticon Lexicon [Mohammad and Turney, 2013b]	62,468	38,312	24,156	–
6	Arabic translation of NRC Hashtag Sentiment Lexicon	54,129	32,048	22,081	–
7	Arabic Hashtag Lexicon	21,964	13,118	8,846	–

Mageed et al., 2014] used the same subjectivity scale. The same set was also used by [Refaee and Rieser, 2014], to which they added the "polar" class to represent subjectivity (positive or negative emotion, evaluation, attitude). The label "mix" covers the cases where tweets are composed of positive and negative sentiments simultaneously.

Arabic emotion detection research on the other hand is obstructed by the lack of publicly annotated data for written Arabic for both MSA and dialectal Arabic as can be see in the table 2.2. The EA proposed corpora also use different classification sets (Ekman, Plutchik and others). [Al-Aziz et al., 2015] used five emotional labels (happiness, sadness, fear, anger, disgust or none), [Almahdawi and Teahan, 2019, Abdul-Mageed et al., 2016] both used the Ekman set, while [Alhuzali et al., 2018] used the Plutchik set. Other corpora use larger sets including {anger, fear, joy, love, sadness, surprise, sympathy, and none} [Al-Khatib and El-Beltagy, 2017], or eleven classes set (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust) [Mohammad et al., 2018].

It is worth mentioning that the highest proportion of available resources and research publications in Arabic OEA are devoted to MSA. Regarding Arabic dialects, the Middle Eastern and Egyptian dialects received the largest share of all research efforts and funding. On the other hand, very little work has been conducted for the Maghrebian dialects [Medhaffar et al., 2017]. In addition, research into ALGD is rare, which in turn causes a lack of resources.

2.3.2 Lexicons

In this study, the use of lexicon-based methods was also investigated. The main existing Arabic lexicons were manually annotated, and are presented in Table 2.3 with details, such as the name and/or origin of each lexicon, the source and its size. Others were generated automatically. [ElSahar and El-Beltagy, 2015] generated a lexicon automatically using 1-2grams with SVM. These lexicons are publicly available ⁵.

2.4 Proposed Approaches for Sentiment Analysis

Over the years OEA has been widely used in a variety of applications such as marketing and politics. These have inspired several methods ranging from lexicon-based approaches [Al-Ayyoub et al., 2015, Al-Moslmi et al., 2018] to corpus-based [Abdul-Mageed and Diab, 2012] to recently deep learning [Al-Smadi et al., 2018]. The literature in sentiment analysis being relatively vast, the focus here was on main existing datasets and methods for MSA and its dialects as well as the common challenges of OEA. The reader can refer to [Liu, 2012b, Liu, 2015, Mohammad, 2016, Benamara et al., 2017b, Zhang et al., 2018] for a comprehensive overview in the field.

⁵<https://saifmohammad.com/WebPages/ArabicSA.html> (The reader can refer to the same link for a comprehensive overview of each lexicon)

As mentioned in the introduction, the analysis of Arabic-speaking opinions and emotions have received much attention in recent years [Abdul-Mageed and Diab, 2012, Nabil et al., 2015, Baly et al., 2017b, Badaro et al., 2018b, Al-Smadi et al., 2018, Abo et al., 2018]. Most studies focus on MSA [Rushdi-Saleh et al., 2011, Aly and Atiya, 2013] and in general they gave satisfactory results [Diab et al., 2010, Maamouri and Cieri, 2002, Manning et al., 2014, Habash et al., 2009, Habash, 2010], although few have investigated sentiment analysis in Arabic dialects. The Arabic dialects are a variety of MSA which includes languages with less normalisation and standardisation [Saadane and Habash, 2015], and differ from MSA at all levels of linguistic representation, from phonology and morphology to lexicon and syntax. Among them we cite Gulf [Elnagar et al., 2018], Jordanian [Duwairi, 2015, Atoum and Nouman, 2019], Iraqi [Abdulla et al., 2013, Alnawas and Arici, 2019], Levantine [Baly et al., 2019, Qwaider et al., 2019], Egyptian [Nabil et al., 2015, Shoukry and Rafea, 2012a], Sudanese [Ismail et al., 2018], Tunisian [Medhaffar et al., 2017, Sayadi et al., 2016], Moroccan mixed with MSA [Oussous et al., 2019], and Algerian [Moudjari and Akli-Astouati, 2019a]. Most of these studies rely on mixed data written in both MSA and the dialect. Some dialect-agnostic datasets have also been proposed such as the SemEval2017 Task 4 sentiment analysis [Rosenthal et al., 2017] and [DoniaGamal et al., 2018].

In [Guellil and Boukhalfa, 2015] three main tasks characterising the opinion mining were identified. The first is how the opinion is formalised "modelling of opinion". The second is the "extraction of opinion" that can affect either a general subject or several subjects or an expression including opinion or the holder of the opinion (the individual who expressed the opinion). The third task is the "analysis of the subjectivity" of a text. The latter is considered objective if it contains facts and subjective if it represents an opinion.

To meet the challenge of the Arabic OEA, many researchers have developed models based on computational linguistics and machine learning. Methods vary from standard feature-based techniques [Duwairi and Qarqaz, 2014, Rahab et al., 2017, Nabil et al., 2015, ElSahar and El-Beltagy, 2015, Elouardighi et al., 2017, Baly et al., 2019] to lexicon-based approaches [Boudad et al., 2018b, Mostafa, 2017, Al-Ayyoub et al., 2015]. Deep learning has also been used relying either on CNN [You et al., 2016, Dahou et al., 2019] or RNN [Al-Smadi et al., 2018, Oussous et al., 2019, Qwaider et al., 2019], where Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) were the most popular models.

The most popular feature-based techniques were n-grams [Duwairi and Qarqaz, 2014, Rahab et al., 2017] and calculated weights such as TF*IDF [Nabil et al., 2015, ElSahar and El-Beltagy, 2015, Ibrahim et al., 2015a] and term-frequency [DoniaGamal et al., 2018]. The combination of the two, namely n-grams and calculated weights, were also exploited in some works, where [Baly et al., 2019, Elouardighi et al., 2017] used TF*IDF weights for the calculated uni-grams and bi-grams.

In lexicon-based methods, most of the lexicons have been created manually. [ElSahar and El-Beltagy, 2015, Al-Ayyoub et al., 2015]. Well known sentiment and emotion lexicon, include the Arabic Twitter Sentiment Lexicon [Kiritchenko et al., 2016] and others [Mohammad and Turney, 2013a] (Arabic Emoticon Lexicon, Arabic Emoticon Lexicon (MSA and dialectal), the Arabic translation of MPQA Subjectivity Lexicon, Arabic translation of NRC Emoticon Lexicon). The most common classical machine-learning methods used are NB [Rushdi-Saleh et al., 2011, Shoukry and Rafea, 2012a, Atoum and Nouman, 2019], Logistic Regression (LR) [Nabil et al., 2015, ElSahar and El-Beltagy, 2015, Baly et al., 2019] and SVM [Shoukry and Rafea, 2012a, Abdulla et al., 2013, Ibrahim et al., 2015a, Atoum and Nouman, 2019, Al-Twairesh et al., 2017].

These approaches represent documents as BOWs using different word features such as n-grams, stems and part-of-speech (POS) tags. Indeed, these works rely on manually engineered features in their classification. The most prominent features are the existence of

specific words from a sentiment lexicon. Lexical-based features have been heavily exploited alongside the n-grams, POS and statistical features (word existence and word frequency).

[Rushdi-Saleh et al., 2011] experimented with their OCA corpus relying on 1-3-grams and cross-validation with different machine learning algorithms such as SVM and NB where the results were encouraging. [Shoukry and Rafea, 2012a] followed the same process while experimenting with their proposed corpus of MSA and Egyptian (1,000 manually annotated tweets by two raters). They experimented with 1-3grams and different combinations (eg: uni+bi grams) using SVM. The best results were obtained using uni-grams with raw tweets. In addition, the combination of the three 1-3 grams and the removal of stop words only outperformed other experiments, while in [Shoukry and Rafea, 2012b], they used NB and SVM with probabilistic BOW for binary Arabic sentiment classification at the sentence-level.

[Aly and Atiya, 2013] on the other hand experimented with different settings using the same features (uni-grams, bi-grams and tri-grams with/without TF*IDF weighting) for SA and rating classification. The SVM performed better in the unbalanced settings (unbalanced corpus in terms of documents per class), while the Multinomial Naive Bayes (MNB) was slightly better in the balanced settings.

[Duwairi and Qarqaz, 2014] used K-Nearest Neighbors (KNN), NB, and SVM classifiers, where SVM gave the best performances in a 10-folds cross-validation. They further enhanced the performance of their classifiers using transliteration [Duwairi, 2015] to transform Jordanian tweets to MSA. Following the same line of research, [Abdul-Mageed et al., 2014] proposed an SVM-based system, called SAMAR for Subjectivity and Sentiment Analysis (SSA), wherein the subjectivity of a text is determined prior to its polarity. They tested different features including word n-grams, POS tagging, word stems and lexicon-based features.

[Adouane and Johansson, 2016] exploited positive and negative restaurant reviews using four different setups with an NB classifier; each setup used a specific lexicon or a combination of lexicons that were created manually. The highest accuracy reported was 90.54%, achieved using the Gulf Lexicon alone. Notably, negation was dealt with by reversing the sentiment of the lexeme of opinion whenever it was directly preceded by an inverter, while [Al-Khatib and El-Beltagy, 2017] experimented on their corpus using an NB, a Complement Naive Bayes and a Sequential Minimal Optimisation classifier. The best results using cross-validation were obtained using Complement Naive Bayes Classifier with stems n-grams (1-3) of words that appeared at least 3 times in the data set.

[Al-Radaideh and Al-Qudah, 2017] followed another direction, using the Rough Set (a mathematical theory used to analyse uncertainty, incomplete information and data reduction) for feature selection in Arabic tweets. Results indicate that the use of the Rough Set theoretical framework can improve overall accuracy and reduce the number of used terms, thus leading to a faster classification process, especially for larger datasets.

Other works have used translated data from English to treat Arabic OEA. [Mohammad et al., 2016b] studied the loss in sentiment predictability when translating Arabic contents into English, manually and automatically. Experiments showed that the translation gave good results with a slight loss in accuracy compared to untranslated texts, and thus concluding that translation does not affect valence. On the contrary, [Salameh et al., 2015] and [Mohammad et al., 2015] show that even with manual translations of texts, cultural differences can lead to significantly different valence associations between speakers of the two languages. They also showed that automatic Arabic translations of English valence lexicons improve the accuracy of an Arabic sentiment analysis system. The experiments with automatic translations showed that English classifiers can be used for languages with scarce resources as benchmarks.

Recently, deep learning applications have been used extensively. [Al Sallab et al., 2015]

presented four architectures, three were based on Deep Belief Networks and Deep Auto Encoders (DAE), and the fourth was based on the Recursive Auto Encoder (RAE) to tackle the lack of context handling in the others. DAE models gave a better representation of the input sparse vector. However, RAE gave the best results, although it required no sentiment lexicon.

A key requirement for Recursive Neural Tensor Networks (RNTN) is the availability of a sentiment Treebank; a collection of syntactic parse trees annotated for sentiment classification. [Baly et al., 2017b] proposed the first Arabic Sentiment Treebank (ArSenTB) which is morphologically and orthographically enriched. The model achieved significant improvements at the sentence-level outperforming well-known classifiers including SVM. When comparing SVM to other Deep Learning (DL) models, the reported results were not conclusive, where the comparison of [Al-Smadi et al., 2018] between RNN and SVM in an aspect-based sentiment analysis of Arabic Hotels’ reviews, showed that SVM outperformed RNN in terms of identifying, extracting and calculating sentiment polarity. However, RNN gave a better execution time.

Shared tasks such as SemEval 2017 task 4 [Rosenthal et al., 2017] help to attract more researchers into the field. The task consists of five sub-tasks. Subtask A involves classifying a tweet into one of the polarity classes, and the other tasks focus on the sentiment related to a topic for two languages: Arabic and English. DL methods were the most popular, such as CNN and LSTM. The top-ranked team scored 0.581 in terms of accuracy. It used an NB classifier with a combination of lexical and sentiment features and increased the training dataset to about 13K examples using external tweets. The second team achieved an accuracy of 0.563, their system used a feature-rich SVM with lexical features and embedding representations. The third team used a convolutional recurrent neural network (CRNN) achieving an accuracy of 0.508.

The use of MSA resources for the DA is seen as a very challenging task given the differences between them [Qwaider et al., 2018]. Consequently, some researchers believe that it would be preferable to treat each dialect as a stand-alone language. Heading towards North Africa, where the linguistic situation is more complicated as there is no dialect derived solely from Arabic, but rather from multilingual dialects. These are known for diversity in spoken dialects, which are less normalised compared to MSA and other Arabic dialects since their vocabulary contains terms from other languages. Additionally, most of these dialects have a significant lack of resources. Nevertheless, recently the community attempting to overcome such shortcomings, where [Qwaider et al., 2019] showed a decrease in performances when training a neural model on an MSA corpus and applying it directly on a Levantine corpus. They obtained no more than 60% accuracy. However, when they trained different Machine Learning (ML) models on the Levantine corpus there was a gain of more than 15% in terms of accuracy. The same approach was adopted to tackle the ALGD. At first, it seemed promising, although yielded significantly low performances [Saadane and Habash, 2015]. So, it was deemed necessary to develop solutions and build resources for the Maghrebi dialects.

[Medhaffar et al., 2017] proposed a Tunisian corpus which contains 17k Facebook users’ comments, manually annotated to positive and negative polarities. They tested their corpus using SVM, NB, and Multi-Layer Perceptron (MLP). For the three machine learning algorithms pre-trained vector representations were used. The vectors were created using the Doc2Vec gensim tool ⁶. SVM and MLP gave similar results and both outperformed NB. While [Oussous et al., 2019] proposed a framework for Moroccan OEA at sentence-level. The framework implements various techniques and models for both Arabic text pre-processing and OEA. The authors built a Moroccan Arabic dataset which consisted of 2,000 tweets/comments annotated for 2-class SA (positive, negative). Tests revealed that DL mod-

⁶<https://radimrehurek.com/gensim/>

els (CNN and LSTM) using pre-trained word embedding were more efficient and achieved higher accuracy than classic ML classifiers (SVM, NB, and ME).

[Harrat et al., 2014] constructed a parallel dataset for Algerian dialects, intending to build Machine Translation solutions for MSA and ALGD, in both directions. Their low performances were due to the significant language deviance, where they presented a Maghrebi multi-dialect study [Harrat et al., 2017], including dialects from Algeria, Tunisia and Morocco which were compared to MSA. [Saadane and Habash, 2015] on the other hand, proposed a list of phonetic rules to be followed, to facilitate the automatic translations of Algerian Arabic and MSA, in both directions. Such tools could be used in several NLP applications, such as OEA. The authors rely on the CODA [Habash et al., 2018] spelling model (Conventional Orthography for Dialectal Arabic) proposed by [Habash et al., 2012], for the Egyptian dialect and Palestinian [Jarrar et al., 2014]. Furthermore, [Zribi et al., 2014] extended the CODA guidelines to consider the Tunisian dialect.

[Mataoui et al., 2016] proposed a manually annotated dataset composed of posts and comments from Algerian Facebook pages and three lexicons (negation words, intensification-words and a list of emoticons with their assigned polarities and a lexicon of common phrases of the ALGD) which were used for SA-2 polarity computation. [Guellil et al., 2018] also relied on a lexicon-based approach for polarity calculation and corpus automatic annotation. Where the lexicon was automatically created using the English lexicon "SOCAL" [Taboada et al., 2011]. Words were translated into Algerian Arabic, although their polarity remained the same. The generated lexicon was used to annotate an Algerian corpus.

Sentiment Polarity Identification on Algerian Arabic Newspaper Comments (SIAAC), a proposed corpus by [Rahab et al., 2017] of Arabic comments from Algerian Newspapers. Tests carried out using SVM and NB gave competitive results in terms of precision, also the use of bigrams increased the results in the two models. Furthermore, when compared to OCA (Opinion Corpus for Arabic [Rushdi-Saleh et al., 2011]) SIAAC showed competitive results. [Klouche et al., 2020] used the same approach, where they collected customers' feedbacks of the Telephone operator Ooredoo written in MSA, Arabizi or Algerian dialect (3,626 positives, 9,063 negatives and 5,784 neutrals). Using a hybrid classification at the document level, employing a lexicon-based approach with SVM and NB, where NB outperforms the SVM using TF*IDF weighting. The "Ooredoo Rayek" system is designed to help Ooredoo's decision-makers to support and evaluate strategies to improve the quality of the operator services based on customers' sentimental information, thus providing a formal, grounded and concise decision. On the other hand, [Abdelli et al., 2019] also proposed a dataset collected from Algerian Facebook pages (24,932 positives and 24,932 negatives) and enriched it with the LABR [Yamada et al., 2018], Elshahr and [Mataoui et al., 2016] corpora. However, the results show that SVM with TF*IDF outperformed LSTM, which uses a CBOW constructed model.

2.5 Proposed Approaches for Emotion Analysis

Internet's transition from the static web to the dynamic web, allowing users to share opinions, ideas, and more, created a large and exploitable data pool. The exploitation of this data has drawn the attention of the NLP community to the importance of the Opinion and Emotion Analysis, which is increasingly gaining attraction. Whereas Sentiment Analysis (SA) is used to determine the polarity of a given text. Emotion Analysis is focused on the emotional state of the author as expressed through the text. Various models for emotions exist, ranging from the most simplistic and classical models, originating from Chinese philosophers to the most modern theories, which refine and extend earlier models [Ekman, 1999a, Prinz, 2004]. The Hourglass of emotions, inspired by the studies of Plutchik [Plutchik, 1984], which has

been widely used in the field of emotion analysis and affective computing [Al-Khatib and El-Beltagy, 2017, Alhuzali et al., 2018], connecting all the different emotions in what is called the wheel of emotions. Emotion detection is generally cast as a multi-class problem which aims at detecting the emotions conveyed in a text from a pre-defined taxonomy of emotions [Mohammad, 2016, Al-Ayyoub et al., 2019, Acheampong et al., 2020]. Well known taxonomies include the Ekman’s classification [Ekman et al., 1980] composed of six basic emotions (*Anger, Disgust, Fear, Joy, Sadness* and *Surprise*), and Plutchik’s eight emotions [Plutchik, 1984], which in addition to Ekman’s classification, count for *Anticipation* and *Trust* emotions.

Although there has been a surge in research on sentiment analysis, little work has been done on ED. Especially for the Arabic language [Abdul-Mageed et al., 2016]. Emotion analysis of DA however, has received even less attention.

Recently, the community is attracting more and more attention to the field with competitions such as the Wassa workshops for English [Klinger et al., 2018]⁷, and the 2018 Semantic Evaluation competition first task for English, Arabic, and Spanish⁸ that helped advance researches in the field.

Very limited work has been carried out for MSA and other Arabic dialects. However, there has been some interest over the last decade. Thus, some attempts can be found for Arabic content. As far as it is known, literature addressing the analysis of ALGD emotions is non-existent. Among them, we cite [Rabie and Sturm, 2014, Abdul-Mageed et al., 2016, Almahdawi and Teahan, 2019] who attempted to predict Ekman’s basic emotions or [Alhuzali et al., 2018] who follow the Plutchik’s framework. Other studies such as [Hussien et al., 2016] propose their own personalised sets of emotions.

Emotion datasets for the Arabic language resources are scarce. Some of the proposed datasets include DINA [Abdul-Mageed et al., 2016], LAMA [Alhuzali et al., 2018], the SemEval2018 emotion detection corpus [Mohammad et al., 2018], and Twifil [Moudjari et al., 2020]. The last two are the only publicly available corpora.

Concerning automatic detection, different machine learning algorithms were used, such as NB [Rabie and Sturm, 2014, Almahdawi and Teahan, 2019] and SVM [Hussien et al., 2016] relying on the standard bag of word features (term frequency, n-grams and TF*IDF). Other experimented with deep learning models including RNN [Alhuzali et al., 2018], CNN [Baali and Ghneim, 2019], and CNN-LSTM [Abdullah et al., 2018]. Most of these models relied on word embeddings created by using task training sets (which are similar to those of this work RAND_w models (see Chapter 5.6)) [Alhuzali et al., 2018, Abdullah et al., 2018, Baali and Ghneim, 2019] or CBOW [Paetzold, 2018]. As far as it is known, no one has investigated multi-level embeddings on emotion detection. A notable exception is the generation of character-level embeddings as part of an RNN [Paetzold, 2018]. The research presented here aims to go further by exploring different unit combinations, targeting for the first time the Algerian dialect together with other Arabic dialects.

[Rabie and Sturm, 2014] experimented with SMO classifier (a simplification of the SVM classifier) and NB using their corpus which contained tweets from MSA and the Egyptian dialect, as well as a simple term search and frequency using their lexicon annotated for emotion analysis. They also proposed a customisation of the Khoja stemmer⁹ to support the Egyptian dialect contained in their corpus. Tests using the term search and frequency outperformed both SMO and NB. [Hussien et al., 2016] experimented also on their corpus created using automatic annotation based on emojis of four classes (*anger, disgust, joy* and *sadness*). Tests using SVM and Multinomial Naive Bayes (MNB) with term frequency

⁷<http://implicitemotions.wassa2018.com/>

⁸<https://competitions.codalab.org/competitions/17751>

⁹<http://zeus.cs.pacificu.edu/shereen/research.htm>. Accessed 26/08/2020

showed that the automatic labelling approaches outperformed manual labelling approaches. [Alhuzali et al., 2018] on the other hand, followed the same approach as [Abdul-Mageed et al., 2016] and constructed the LAMA corpus which reached 7,268 collected tweets using a set of emotional seeds. However, in this work, the authors used a list of seeds for each of the Plutchik 8 primary emotions. The Deep Gated Recurrent Neural Networks (GRNN) with word embeddings learned directly from the training set outperformed the baseline classifiers (SVM, NB, Passive Aggressive Classifier, Perceptron, and linear SVMs trained with Stochastic Gradient Descent (SVM-SGD)) which used different n-grams (1-3) with different combinations.

An Iraqi corpus was constructed using 1,365 posts from Facebook, collected using geolocation tags and keywords related to each class of the Ekman emotion set [Almahdawi and Teahan, 2019], where the annotations were further checked by four annotators. Using different classifiers with 1-3 grams selected using Weka’s StringToWordVector filter with NGramTokenizer, the results showed that NB gave good results. However, a compression-based classifier called PPM outperformed all the other classifiers reaching 86.9% accuracy.

The 2018 Semantic Evaluation competition included five sub-tasks on inferring the affectual state of a person from their tweet: 1. emotion intensity regression, 2. emotion intensity ordinal classification, 3. valence (sentiment) regression, 4. valence ordinal classification, and 5. emotion classification. For each sub-task, labelled data were provided for English, Arabic, and Spanish [Mohammad et al., 2018]. Affect lexicon features, word n-grams and word embeddings, were some of the most widely used features in the submitted runs for the SemEval2018 task 1. Indeed, this task provided a substantial corpus that combines data from the 2017 tasks, which were annotated manually for the 2018 tasks into 12 emotion classes. The dataset was used in different works [Mohammad and Kiritchenko, 2018b, Alhuzali et al., 2018, Abdullah et al., 2018, Baali and Ghneim, 2019]. SVM/SVR, LSTMs and Bi-LSTMs were some of the most used machine learning algorithms. Where word embeddings, affect lexicon features, and word n-grams were some of the most used features.

The OEA of Arabic content is quite a new field, thus, the challenges and limitations existing works face are numerous. Therefore, this work aims at helping advance the research in this field as well as help overcome some of these limitations.

2.6 OEA Limitations

One of the biggest issues in the opinion and emotion analysis’ field is that the data is domain-based. Although domain-specific applications are beneficial, stronger support for domain and context is needed to provide a better understanding of data and avoid ambiguities. In the example "Honey is not bad!", where due to the lack of context the word "honey" is ambiguous with three possible explanations: the food substance produced by bees, HONEY MOTORS the used car dealer in Dubai and Honey Grove, the city in Texas. [Yerva et al., 2010] tried to distinguish tweets about companies such as Apple (apple has two meanings: the company and the fruit). Others relied on transferred learning to capture the different specifications for different domains, [Glorot et al., 2011] proposed a protocol relying on Stacked Denoising Autoencoder (SDA) with rectifier units to extract features in a supervised manner, and an SVM classifier with a squared hinge loss trained on a labelled dataset for one domain and tested on other domains. The system proved its effectiveness compared to others. The OEA field can benefit from an accuracy enhancement, where some rare works considered parameters such as gender to improve sentiment classification [Volkova et al., 2013] and cyber-bullying detection [Dadvar et al., 2012]. [Abdulla et al., 2014] on the other hand, highlighted that the accuracy level in lexicon-based approaches can be improved using different techniques, among which: expanding the lexicon with a dialect-

specific and domain-specific content, performing more thorough experimentations on bigger and more diverse corpora and better handling of the very difficult cases such as irony and sarcasm. It is also believed that adoption of informal speech -used in everyday conversations on social platforms- alongside the formal can enhance systems' accuracy even further since the results of mixed corpora are very encouraging [Shoukry and Rafea, 2012a, Mataoui et al., 2016, Abdul-Mageed et al., 2014].

Most of the emerging works in Arabic are not yet releasing their resources. Some of these work on a domain basis for news, product reviews or MSAs in Arabic only and exclude dialects and their complexity. Also, the lack of resources in Algerian dialects affects progress in this area. Indeed, Opinion and Emotion Analysis is a recent field, where several aspects remain challenging.

2.7 Conclusion

Due to the considerable differences between MSA and dialectal Arabic, some researchers believe it would be better to process each dialect as a standalone language. Although the community has made remarkable progress in this field, specifically for MSA, a lot remains to be performed for dialectal Arabic [Badaro et al., 2019], where each dialect has different characteristics making sentiment analysis harder. In addition, several limitations can be observed in the proposed approaches, and thus the field is still an open challenge.

Proposed approaches ranging from standard feature-based techniques to lexicon-based approaches to recent deep learning-based approaches. Also, several resources were developed, although few were publicly available. Nevertheless, these resources suffer from different limitations including vocabulary coverage and domain adaptation [Araque et al., 2019]. However, [Abdulla et al., 2014] believe that task-specific and domain-specific features among others can enhance OEA systems since sentiment lexemes can display variations in their polarity values depending on the contextual domain [Turney, 2002], language [Perez-Rosas et al., 2012] or even context [Ding et al., 2008], causing a decrease in lexicon-based approaches performances. These limitations can be tackled using semantic-based approaches, either by using Linked Data or embeddings. Moreover, tweets include a lot of neologisms¹⁰, abbreviations, emojis, elongations, negations, intensifiers, polarity shifters, and sarcasm. Thus, more sophisticated natural language processing tools are needed to detect these critical features for sentiment, as well as morphological features [Badaro et al., 2019]. [Baly et al., 2017a] also suggests that each dialect requires a customised opinion-mining model. However, having an accurate opinion-mining model for each dialect is only possible if there are large-scale annotated sentiment corpora for each of the dialects [Badaro et al., 2019].

Recent techniques exploit the concept of distributed semantics, normally through word embeddings. Where a semantic similarity metric is computed between text words over a large corpus. Distributed representations, or word embeddings, are considered a key characteristic of state-of-the-art SA systems, where embedding models can be used to overcome lexicon limitations [Araque et al., 2019] since they provide semantic aspects that can help with ambiguities. However, a new direction in this field surged, that opted for Linked Data technologies to overcome the semantic lacking in precedent approaches. The next chapter presents such technologies and the efforts made in the OEA's field.

¹⁰A neologism is a relatively recent or isolated term, word, or phrase that may be in the process of entering common use.

Chapter 3

Linked Data

The immense contribution of the social web has greatly motivated researchers. This has led to the emergence of techniques that have proven their effectiveness in customised opinion and emotion modelling related to applications such as NLP and automatic learning. However, when it comes to interoperability and a single agreed-on encoding of opinions and emotions, there remain some gaps, since the modelling is generally custom made for each application. This prompted a new direction of research which combines opinion analysis works with those of "Linked Data".

In this chapter, different solutions and projects are presented, such as Onyx, which was highlighted due to the domain's standardisation and the reasons why we also believe that data semantics is crucial to the accuracy and efficiency of the models. In addition, limitations and the possibility of advancement in the OEA/LD domain are shown.

Keywords Linked Data, opinion mining, sentiment analysis, social media, microblog, machine learning, deep learning, ontology, sentiment, emotion.

3.1 Introduction

Opinion and emotion analysis drew researcher's attention from different fields since the early 2000s. It was quickly sought after in the data mining field, web mining and information retrieval [Liu, 2012a]. Recently, the Linked Data community joined. The social web, the source number one for subjective data, which interlinks people from various backgrounds and allows different content to be shared. Web 3.0 is a place where a collection of datasets is interlinked and new knowledge is inferred, also called "Linked Data". So, what is Linked Data and why were researchers interested in applying it in this field?

The remainder of the chapter is organised as follows, section 2 provides an overview of Linked Data and its use. In section 3 Linked Data technologies and their use for opinion and emotion analysis are introduced. Section 4 presents some tools such as ontologies, while section 5 introduces projects related to the field and discusses their strengths and weaknesses. Section 6, concludes this chapter and presents directions for future research.

3.2 Linked Data

The web of data¹ is a collection of interrelated datasets on the web, also referred to as Semantic Web or Web 3.0, which gained its name in 2006 after Tim Berners-Lee, director of the World Wide Web Consortium (W3C), coined the term. To assist the use of such resources,

¹This section mainly used the references: [Bizer et al., 2008, Bizer et al., 2009, Heath and Bizer, 2011, Bizer, 2009] and the w3c online resources ²

special means are provided, the "Semantic Web technologies", which helps query databases, discover new relations and "inference" new knowledge due to automatic procedures using the data itself or vocabularies. The latter is called ontology in a more formal language -in the Linked Data context. However, it is a set of terms (concepts, relationships) that describe and represent an area of concern, so vocabulary is a classification of terms that can be used in a particular application, characterise possible relationships, and define possible constraints on using those terms. This means that automatic procedures can generate new relationships based on the data and based on some additional information in the form of vocabulary.

Linked Data also designates a list of rules to respect, so that the published data is standardised and structured to be interlinked and better exploited using Semantic Web technologies (RDF, OWL, SPARQL) to infer new knowledge.

URI Uniform Resource Identifier, which identifies particular resources.

RDF Resource Description Framework is a directed, labelled graph data format for representing information in the Web. It represents the data using a triplet:

< subject, predicate, object >

equivalent to:

< identifier, attributename, attributevalue >

the first two are represented using a URI that helps identify the data as unique.

Example:

< employee3, title, Professor >

URI makes sure that the job title of a professor is the one referred to and not a book title. The object as well can be a URI, is, therefore a subject of another triplet, and so relations between data are discovered and data are interlinked creating graphs.

OWL Ontology Web Language, a Semantic Web language designed to represent knowledge. OWL documents, known as ontologies, can be published on the World Wide Web and may refer to or be referred from other OWL ontologies. OWL is part of the W3C's Semantic Web technology stack.

SPARQL a Query Language for RDF, it can query graph patterns along with their conjunctions and disjunctions. The result of SPARQL queries can be sets of results or RDF graphs. It helps access a growing collection of public and private data, add/delete or update the stored data. The web of data is a huge pool of datasets interlinked, ontologies such as DBpedia³, Freebase⁴ and Geonames⁵ have contributed greatly to the process of linking published data and inferring new relations between data, which can be accessed using Linked Data browsers. Instead of following links between HTML pages using HTML browsers, Linked Data browsers enable users to navigate between different data sources by following RDF links.

So basically, Linked Data is a way to link existing sources of data, which helps infer new knowledge and obtain more relevant information to a query.

³<http://dbpedia.org/page/>

⁴Freebase is an open database of the world's information. It is built by the community and free for the community, for anyone to query, contribute to, built applications on top of, or integrate into other websites. Source: <https://cutt.ly/freebase> Accessed: 28/04/2020

⁵<https://cutt.ly/Geonames>

Is this not the same as the current web and does what google is doing? No, it might sound the same, yet the infrastructure technology is not the same. Besides, the web3.0 is a graph-based architecture and the web2.0, in its advanced version is unstructured, non-linked HTML web pages. That is why the advanced version of search engines is migrating slowly but surely to such technology. However, the real question is: what does it have to offer to OEA?

3.3 Linked Data and the Analysis of Opinions and Emotions

One of the factors which makes it difficult to exploit opinions is that emotions and subjectivity are highly sensitive to the context and highly dependent on the field. For example, the ice cream melted (negative) and the cheese melted (positive) in the food sector[Liu, 2012a]. Due to the dependence of the analysis on the context and domain of the opinions and emotions expressed, a semantic network linking entities and features to their exact meaning in a phrase can help tremendously overcome this issue. It will also provide a uniform semantic interface for users and entities will have a uniformed URI nomenclature according to the related data conventions.

In the following, some reasons are cited that explain our interest in Linked Data for the analysis of emotions and opinions.

- The different formats for representing data in the opinion and emotion analysis are heterogeneous, generally modelled according to the need of the application despite the existence of several standards [Sánchez-Rada and al, 2016].
- In social media, it is time and labour consuming to obtain sentiment labels, social media data is often unlabelled [Cheng et al., 2017], Linked Data can unveil and link already labelled published data to help advent the research.
- Although this area has fostered innovation in a range of SMEs (Small and medium-sized enterprises) with strong semantic and linguistic skills, the language resources developed remain limited to their clients. The main reason for this is the fear of losing competitiveness or missed returns on research/business investment ⁶.
- Easily accessible social platforms such as Twitter provide data from people of different origins, different languages and various topics discussed. To capture the meaning of tweets by considering all these factors is beyond the capabilities of existing formats for sentiment analysis, which hinders the emergence of applications that make deep sense of data. The use of Linked Data can help better exploit such information [Sánchez-Rada and al, 2016].
- Expressions such as: "it's cooooool" and "good n8 :)", are not standard, although, are more acceptable and frequently used in social media [Hu et al., 2013], Linked Data can unveil the standard format of these expressions and their meaning, therefore, simplifying the opinion detection.
- The richness of data available on social networks beyond pure text can be exploited to offer a deeper understanding of user-generated content, and ultimately train a system to look for more targeted characteristics, process and categorise the content. Apart

⁶<http://eurosentiment.eu/section/project/>

from the actual content, it is also the context in which it was created that can serve as a rich source of information and be used to generate more powerful data analysis and make smarter business decisions [Sánchez-Rada and al, 2016].

- There are commercial solutions, such as Lexalytics⁷, Sentimetrix⁸, Hootsuite⁹ and Tweetreach¹⁰. Though quite generic, they fail to deal with the multilingual aspect and do not adapt to the enormous volumes of data to be processed. Neither do they achieve the integration of emotional analysis observations between data sources and/or modalities at a significant level and are not trained for domain-specific terminology, idioms and characteristics [Sánchez-Rada and al, 2016].
- Domain-specific resources for multiple languages are potentially valuable, yet not shared. Sometimes this is due to intellectual property and license considerations, although often for technical reasons. Including interoperability, there are some initiatives such as Text Encoding. To date, there is still an absence of a widely accepted global solution for integrating and combining heterogeneous linguistic resources from different sources [Iglesias and al, 2017].

Several studies on the Arabic sentiment analysis have been carried out in recent years. These studies focus mainly on MSA among which few have investigated the case of Arabic dialects. Algerian dialect is considered as one of the most "hard to understand" Arabic dialect varieties. It is far less normalised and standardised compared to MSA. It has been enriched by many languages (Arabic, Turkish, Italian, Spanish, French...) which resulted in a complex linguistic situation [Mataoui et al., 2016]. Popular works in the area are generally limited to sentiment polarity classification using supervised classifiers such as Naïve Bayes and SVM classifiers mostly. The techniques used were mainly based on machine learning where models such as SVM proved their effectiveness. However, the fundamental issue these techniques suffer from is the lack of support for interoperability and a single agreed-on encoding, since the modelling is custom made for each application makes it difficult to find models already trained or data already labelled for different dialects [Moudjari and Akli-Astouati, 2018]. The lexicon-based techniques also lack a semantic understanding of the terms used and face an ambiguity problem more so than the other techniques, which could be avoided if the correct meaning is located through an ontology, providing links to external Linked Open Data (LOD) repositories such as DBpedia or through a semantic query if the application provides such technology. This holds true when it comes to dealing with data published by users from different origins, different languages and various topics. Even though sentiment analysis makes use of such data to develop useful applications for daily use, it still suffers from several weaknesses. Either the limited number of words contained in affect lexicons or the fact that they do not consider the most diverse parameters which can influence the opinion or emotion expressed in a sentence. Using Linked Data is an interesting turning point for opinion and emotion analysis. Not only can it help standardise the modelling of emotions and opinions, but also simplify the extraction of subjective data to ultimately reach a better understanding of opinions, while taking into consideration context, cultural background and other factors.

Some techniques have been implemented using a set of tools as presented in the following.

⁷<https://www.lexalytics.com/>

⁸<http://www.sentimetrix.com/>

⁹<https://hootsuite.com/>

¹⁰<https://tweetreach.com/>

3.4 Tools

Presented here is a list of tools recommended by W3C, since it promotes web standards.

- **WordNet Affect**, an affective lexicon providing affective polarity of words. It is based on WordNet, adding a new set of tags to a selection of synsets to annotate them with affective information [Strapparava et al., 2004].
- **SenticNet** a knowledge base that provides a set of semantics, sentics(emotion categorization values expressed in terms of four affective dimensions (Pleasantness, Attention, Sensitivity, and Aptitude)), and polarity (floating number between -1 and +1) associated with 100,000 natural language concepts¹¹.
- **EmoSenticNet** is a lexical resource that assigns six WordNet Affect emotion labels to SenticNet concepts. It is considered as an expansion of WordNet Affect emotion labels to a larger vocabulary¹².

Chinese Emotion Ontology was developed to simplify the understanding, recognition and classification of emotions in the Chinese language. Based on HowNet, the Chinese equivalent of WordNet [Sánchez-Rada and al, 2016].

- **Emotive Ontology** the work of Sykora et al. [Sykora et al., 2013] in 2013, proposed a mechanism to extract emotions from informal messages.
- **NLP Interchange Format (NIF)** an RDF/OWL-based format, it aims to achieve interoperability between NLP tools, language resources and annotations. NIF consists of specifications, ontologies and software. It defines a semantic format and an API for improving interoperability among NLP services. NIF can be extended via vocabularies modules [Hellmann et al., 2013]. It uses Marl for sentiment annotations [Sánchez-Rada and Iglesias, 2016].
- **Provenance Ontology** provides information on the entities, activities and people involved in producing the data [Sánchez-Rada and Iglesias, 2016].
- **Lemon** (lexicon model for ontologies) built on previous work on standards for lexical resources' representation, the Lexical Markup Framework (LMF2) yet extends the underlying formal model and provides native integration of lexica with domain ontologies. It was designed to model lexicon and machine-readable dictionaries and link them to the Linked Data cloud. Lemon-based ontology lexicalisation is the use of URIs for uniquely identifying all objects defined by the lemon model (lexicons, lexical entries). Its main purpose is representing language resources for opinion and emotion analysis in a Linked Data conform way (RDF-native form), leveraging existing Semantic Web technologies (SPARQL, OWL, RIF) [Buitelaar et al., 2013].
- **Marl** a vocabulary for annotating and describing subjective opinions, it models feature level opinions. Marl follows the Linked Data principles as it is aligned with the Provenance Ontology, it represents lexical resources as linked data, and has been integrated with lemon [Buitelaar et al., 2013].
- **Onyx** a semantic vocabulary of emotions with the focus on lexical resources and emotional analysis services. It follows a linguistic Linked Data approach.

¹¹<http://sentic.net/>

¹²<https://www.gelbukh.com/emosenticnet>

These tools help represent opinions and emotions and provide their sentiment polarity. Onyx on the other hand tried to provide a tool that combines most of these tools. It has been aligned with vocabularies such as NIF and the Provenance Ontology. Also, integrated with lemon to represent lexical entries. Aligned with EmotionML and WordNet-Affect to work with different theories of emotion. The combination of subjective information from opinion and emotion analysis with facts already published as Linked Data would make it possible to offer a wider range of new services. This would require a widely accepted representation of emotionally related data. Onyx aims to fill this gap and allow interoperability of tools and resources [Sánchez-Rada and Iglesias, 2016].

3.5 Relevant Projects

In order to provide standardised metrics for opinion and emotion analysis, the World Wide Web Consortium (W3C), created the forum¹³ "Linked Data Models for Emotion and Sentiment Analysis W3C Community Group" to promote sentiment analysis research, where the main topics discussed are Linked Data based vocabulary and models for emotion and sentiment analysis.

This section introduces projects and approaches which adopted the Linked Data philosophy. Also mentioned are the advantages and disadvantages of each one (see Table 3.2):

- **Emotion Markup Language (EmotionML)**[Schröder et al., 2011], a markup language, recommended by the W3C community and widely used since it provides a common data representation for multiple opinion and emotion analysis' applications.

Table 3.1: EmotionML example.

```
<sentence id="sent1">
  Do I have to go to the dentist?
</sentence>
<emotionml version="1.0" xmlns="http://www.w3.org/2009/10/emotionml">
  <emotion category-set =
    "http://www.w3.org/TR/emotion-voc/xml#everyday-categories">
    <category name="afraid" value="0.4"/>
    <reference role="expressedBy" uri="#sent1"/>
  </emotion>
</emotionml> [https://www.w3.org/TR/emotionml/]
```

Limitations

- The representation of suppressed emotions, simulated and masked by another emotion.
- The use of ontologies to define the terms in an emotion vocabulary, to relate the terms to one another, and to define mappings between emotion vocabularies.
- **ArsEmotica AEO** an ontology-driven approach proposed in 2015, encoded in OWL and incorporates, in a unifying model, multiple ontologies describing different aspects

¹³<https://www.w3.org/community/sentiment/>

of the connections between media objects (e.g. works of art), people and emotions. It helps identify emotions in art magazines, or excerpts from them. In addition, due to the need to model the link between words and the emotions they convey, AEO integrates Lemon to provide the lexical model and WordNet-Affect synset to provide the affective information. With linkage to external LOD repositories (e.g. DBpedia). The ontology contains 87 emotional concepts to categorise emotion denoting words. Each class was populated by instances from English and Spanish words. Over 450 Italian emotional words were added using MultiWordNet and WordNet-Affect [Patti et al., 2015].

Limitations

- Detect emotions in multi-word expressions, that do not explicitly convey emotions but are related to concepts which do.
- To operate on a real multilingual level, additional language resources need to be incorporated.
- **EuroSentiment** provides a pool of shared language resources for multilingual opinion and emotion analysis. The lexicon format is based on a combination of lemon (lexical concepts), Marl (opinion/sentiment) and Onyx (emotions). The corpus format uses NIF, Onyx and Marl for subjectivity. Each entry in the lexicon is described with part of speech tags, morphosyntactic information, links to DBpedia and WordNet and information on the identified opinion and/or emotion. The project provides semantically rich lexical resources, a set of lexicons and corpus, conversion tools from inherited non-semantic formats, an extension of the NIF format and an API for Web services in different programming languages to simplify the development of semantic services for the analysis. It contains resources for the electronics and hotel domains in six languages (Catalan, English, Spanish, French, Italian and Portuguese)[Sánchez-Rada et al., 2014].

Limitations

- It is restrained to six European languages and does not include other popular, cross border languages such as Arabic. Although millions of Arab speaking individuals living in Europe, six million in France alone ¹⁴.
- It has a limited set of resources and domains.
- It only processes text data with no support for other data formats (images and sounds).
- **MixedEmotions**¹⁵ continues the work of EuroSentiment, including other media (image and sound) in different languages. To develop new multi-dimensional multimodal data analysis applications to analyse a more complete emotional profile of user behaviour using data from mixed input channels: multilingual text data sources, A/V signal input, social media and structured data. It implements an integrated Big Linked Data platform for emotion analysis based on heterogeneous data sources, languages and modalities, using existing tools, services and approaches to track the emotional

¹⁴https://apaelo.com/know/view_html.php?q=Arabs%20in%20Europe&sq=Arabs-in-Europe&language=en

¹⁵<https://mixedemotions-project.eu/linked-emotions-data/>

aspects of user interaction. NIF, to assign URIs to language resources, to be annotated following Linked Data principles. Lemon to represent lexicons and linking lexical and semantic entities to lexical forms. Marl and Onyx for annotating sentiments and emotions.

Limitations It is the same as EuroSentiment, it needs more multilingual and multi-domain enrichment.

When talking about the multilingual aspect, with the assumption of referring to the formal speech, contrary to the informal speech used in everyday conversations on social platforms. Integrating support for such vocabulary would enrich the understanding of emotion and opinions on such platforms.

- **SEWA** (European Sentiment Analysis in the Wild¹⁶) develops visualisation, speech processing and automatic learning tools for the automatic understanding of human interactive behaviour in naturalistic contexts for Spatio-temporal analysis of sentiments.

Limitations Another example of an interesting opinion and emotion analysis project. Not only does it deal with text inputs, but facial, vocal and verbal behaviour (interest, calmness, positivity, likeability, ..) as well. It requires more enrichment on the multilingual level, and according to its website, the project has some limitations regarding social signals and image processing.

- **OpeNER** Open Polarity Enhanced Named Entity Recognition provides a set of NLP tools. OpeNER uses the annotation format KAF, with ad-hoc elements to represent the characteristics of opinion and emotion. The results of the project included an annotated corpus of reviews and a Linked Data node that displays this information [Bosma et al., 2009]. It offers a usable language analysis tool for six languages for applications such as Reputation Management and Information Access. The project focus was on a generic application domain than was adapted to the Tourism domain [Agerri et al., 2013].

Limitations OPENER, can be enhanced with the adoption of more domains and context support to avoid ambiguities. More languages are always an advantage and an opportunity for larger use, academic or industrial, not just formal languages, but informal ones as well.

- [Charfuelan and Steiner, 2013] presented a framework for the synthesis of expressive speech based on MARY TTS (text-to-speech) using audiobook data and EmotionML. It creates expressive unit selection and HMM-based voices using audiobook data labelled according to voice styles. Data were split according to voice styles by principal component analysis (PCA) of acoustic features extracted from segmented sentences. EmotionML was used to represent and control expressivity in terms of discrete emotions/opinions.

Limitations In our opinion, what could boost the expressivity of the generated audio is a domain-based extension. The framework could also enlarge its user base by supporting more languages.

¹⁶<https://www.sewaproject.eu/>

- **Sentilo** an unsupervised, domain-independent system that performs sentiment analysis by hybridizing NLP and Semantic Web technologies. It makes use of affective knowledge resources such as SenticNet and SentiWordNet. It recognizes the holder, detects the topics and subtopics of a sentence expressing an opinion, links them to relevant situations and events and evaluates the sentiment expressed, outputting an RDF graph. It uses FRED (semantic Web aware machine reader) which resolves the identity of entities involved in an opinion present on resources like DBpedia and WordNet [Recupero et al., 2015].

Limitations For the lexical aspect, Sentilo uses WordNet a large lexical database of English tokens. Since WordNet provides multilingual support, Sentilo could use it not only to acquire a larger user base but also as a reference and a recommended tool for topic/subtopic detection.

Table 3.2 represents some limitations of the mentioned projects and some suggestions to overcome these limitations.

Adding support for popular, cross border languages such as Arabic since millions of Arab speaking individuals live in Europe ¹⁷. This would enlarge the user base on an academic and industrial level. Not only formal languages but informal ones as well. Integrating support for informal speech would enrich the understanding of emotion and opinions. The adoption of more domains and context support can help avoid ambiguities.

As mentioned, other parameters can enhance the analysis such as "gender", Linked Data can facilitate the process by exploiting users' data already published and link profiles, which takes us to another point "revealing fake profiles on social media". Each person has at least three to four accounts on different social platforms, detecting fake profile can help diminish cybercrimes and violent movements, yet most importantly avoid suspected profiles linked to terrorism.

3.6 CONCLUSION

The aim of this chapter was to illustrate the main linked data techniques and approaches proposed in the literature for the analysis of opinions and emotions and to extract their vulnerabilities. However, on the other hand explore the linked data and their potential to enrich the OEA, to determine fundamental research questions and to assess areas for improvement.

Usually, sentiment lexicons are used for polarity estimation through the matching of words contained in a text and their associated lexicon sentiment polarities. Nevertheless, such resources have limitations in vocabulary coverage and domain adaptation. The use of Linked Data could help overcome some of these issues, where technologies such as RDF, OWL and SPARQL provide a standard format for publishing structured data that could be linked to other data that provide further information when queried. The fact that projects such as ArsEmotica and EuroSentiment have already introduced such technologies to deal with similar problems supports this work's view. Nevertheless, to guarantee the success and adoption of the new approach, common vocabularies and best practices for their use are required. Works such as Onyx and Lemon are the first step in this direction, especially with the presence of initiatives such as the W3C Community Group "Linked Data Models for Emotion and Sentiment Analysis". Data-enriched with linked meta-data and surface

¹⁷https://apaelo.com/know/view_html.php?q=Arabs%20in%20Europe&sq=Arabs-in-Europe&language=en

Table 3.2: Limitations and suggestions of opinion and emotion analysis' projets.

Project	Limitations	Suggestions
EmotionML	- Representation of suppressed emotions, simulated, masked by another emotion.	+ Add an attribute "master" to "emotion" tag to represent the explicit emotion. And use another emotion tag to represent the masked emotion(<i>master = false</i>). + Use ontologies to define and link terms and map emotion vocabularies.
AEO	- Detect emotions in multi-word expressions, that do not explicitly convey emotions, but are related to concepts that do.	+ Incorporate more language resources to operate on a multilingual level. + Integrate support for informal speech.
Euro-Sentiment	- Restrained to six languages. - Limited set of resources and domains. - Handles only textual data.	+ Add other popular, cross border languages such as Arabic.
Mixed Emotions	- Restrained to six languages. - Lacks support for informal speech.	+ Needs more multilingual and multi-domain enrichment. + Integrate support for informal speech.
SEWA	- According to ¹⁶ , it has some limitations regarding social signals and image processing.	+ Needs a multilingual enrichment. + Add support for informal speech.
OPENER	- Lacks multilingual support. - Lacks support for informal speech.	+ Adopt more domains and context support to avoid ambiguities.
MARY TTS	- Lacks multilingual support.	+ boost the expressivity of the generated audio with a domain-based extension.
Sentilo	- Lacks multilingual support. - Lacks support for informal speech.	+ Exploit WordNet for a multilingual support to enlarger the user base, therefore be a reference tool for topic/subtopic detection.

features (gender, age, number of positive and negative tokens) can help improve systems' accuracy, as well as a support for multilingual.

When reading about Linked Data, the research question to ask is how to give a text more semantics and related concepts that could help improve the precision of the classifier? [Araque et al., 2019] provided an answer through the exploitation of four different sentiment lexicon vocabularies (Bing Liu's [Hu and Liu, 2004], SentiWordNet 3.0 [Baccianella et al., 2010], Affective Norms for English Words (ANEW) [Bradley and Lang, 1999], and AFINN [Nielsen, 2011]) and word embeddings (Word2Vec pre-trained English models). Their experiments demonstrate that the proposed method can improve Sentiment Analysis performance over the baseline. Indeed, Linked Data is not the only technology offering semantic features. Recently, new and more complex models emerged in the deep learning discipline. These models offer not only syntactic and semantic data encoding but they also take into consideration morphology and context. These dense vectors (See next chapter) are a new promising feature extraction method. Besides, many recent techniques exploit the concept

of distributed semantics, normally through word embeddings.

The Semantic Web is a Web of Data. The collection of Semantic Web technologies (RDF, OWL, SPARQL, etc.) provides an environment where applications can query the data, draw inferences using vocabularies, etc. However, to make use of such technologies, it is of the utmost importance to have a huge amount of data available in a standard format, to make either conversion or on-the-fly access to existing databases (relational, XML, HTML, JSON). It is also important to be able to setup query endpoints (APIs) to access that data more conveniently ¹⁸.

Unfortunately, when it comes to the Arabic language in general and the Algerian Dialect (ALGD) in particular, the serious lack of publicly accessible resources and the rarity of published models hinder the progress in this field.

To bridge the gap, a first step would be to enrich available resources, either corpora, lexicons or classification models, the next step would be the creation of web services that help interact with these models and retrieve results in a standard format.

This work, will focus mainly on providing more resources to the community (datasets, lexicons) as well as more robust models that take into consideration the semantic aspects of the processed text to provide a sentiment/emotion classification.

We were keen on using LD as it offers the possibility to discover new relationships and "inference" new knowledge. LD, or more specifically the process of publishing Linked Data or creating links between already published data is time consuming knowing that the biggest part must be done manually. Additionally, in most of the existing works, the metrics were often not explicitly defined or did not consist of precise statistical measures. Moreover, only a few approaches are accompanied by an implemented tool, and none of these tools were able to ensure overall quality of the dataset [Rula et al., 2016]. However, when the strengths of embeddings and the simplicity of their creation process were discovered, we were attracted by their power to represent semantics in a one hot encoded vector and their ability to be generalised to cover unseen words, as well as the interoperability between models.

In the light of such trends, it is reasonable to think that in order to improve sentiment analysis systems, words semantics exploitation methods should be improved, based on this conclusion and the review of recent works, such as [Araque et al., 2019] who relied on a representation that incorporates a semantic similarity metric and taxonomy sources like WordNet, which are more promising. Therefore, in the next chapter, there will be further analysis conducted on these semantic-enriched models providing weaknesses and strengths and why we believe that they provide stronger fundamentals for OEA classifiers. Indeed, this approach helps present atoms (words, sub-words, characters) in a dense vectorial space as detailed in the next chapter.

¹⁸<https://www.w3.org/standards/semanticweb/data>

Chapter 4

Vectorial Representation

Recent techniques exploit the concept of distributed semantics, normally through word embeddings as new data representations. This is Where a semantic similarity metric is computed between text words over a large corpus. Sentiment lexemes can display variations in their polarity values depending on the contextual domain [Turney, 2002], language [Perez-Rosas et al., 2012] or even context [Ding et al., 2008], causing a decrease in lexicon-based approaches performances. Distributed representations, or embeddings, are considered a key characteristic of state-of-the-art Sentiment Analysis systems. Embedding models can be used to overcome lexicon limitations [Araque et al., 2019] since they provide semantic aspects that can help with ambiguities. Indeed, this approach help presents atoms at different granularities (words, sub-words, characters) in a dense vectorial space. This chapter details the aspects and characterisations of these models and presents works in the literature targeting sentiment analysis, with an emphasis on Arabic and its varieties.

Keywords Embedding, dense vectors, multi-level embeddings, Convolutional Neural Networks, Arabic dialects, sentiment analysis, emotion detection, irony detection.

4.1 Introduction

Machine learning algorithms and deep learning architectures are unable to process strings and plain text in their raw format. Machines process data only as numerical or mathematical representations of different models' inputs for different tasks (classification, regression or others in a broad sense). When an enormous amount of data is present in text format, it is imperative to extract knowledge to build daily use applications, necessary to help humans save time. As traditional representations such as word frequency or TF*IDF do not count for semantic and contextual aspects. Dense vectors in Euclidean space, also known as embeddings, have become one of the most popular representations in natural language processing. Based on the distributional hypothesis [Harris, 1954], these representations are capable of capturing meaning in context and thus allowing to compute semantic and syntactic similarities between linguistic units from text co-occurrence matrix [Baroni and Lenci, 2010]. Early models were proposed in 2013, with the aim at representing the data in such a way to include both the word's syntax and semantics [Mikolov et al., 2013c]. Embeddings can be learned from text data and reused within machine learning algorithms. They can be produced using different neural networks, shallow (Word2Vec [Mikolov et al., 2013c]) or more deepen (CNN [Dahou et al., 2019]). A neural network's training is based on tweaking its hyperparameters, while word vector representations follow the notion that similar words are closer to each other [Mikolov et al., 2013c]. Recently, the community is more attracted to building embeddings using deep learning techniques such as RNN, which takes a sequence of n-grams as inputs. This type of model effectively learns etymological representations (roots,

affixes) and is more tolerant with out of vocabulary words [Pinter et al., 2017]. Another representation emerged in late 2018, offering more context to such representations [Devlin et al., 2019], where a word is represented differently in different contexts, which ultimately can help overcome ambiguity problems. Indeed, deep learning has been widely used in sentiment analysis in recent years relying on several Arabic pre-trained language models such as Word2Vec [Altowayan and Tao, 2016, Soliman et al., 2017, Al-Azani and El-Alfy, 2017], global representation of word vectors (GloVe) [Fouad et al., 2020] and FastText [Kaibi et al., 2019]. Others proposed to learn their language models, like [Mulki et al., 2019] who trained a neural network with syntax-ignorant n-gram embeddings and [ElJundi et al., 2019] who used the Arabic ULMFit. Recently, [Dahou et al., 2019] propose a multi-channel CNN architecture for combining FastText and domain-specific word embeddings.

Embeddings provide a dense representation of entities at different linguistic levels ranging from character n-grams [Bojanowski et al., 2017], word [Pennington et al., 2014, Mikolov et al., 2013c], sentence [Conneau et al., 2017, Cer et al., 2018] to document [Le and Mikolov, 2014, Medhaffar et al., 2017], including concept [Yamada et al., 2018, Camacho-Collados et al., 2016]. These representations can either be monolingual or multilingual where units from different languages are mapped to the same area in the latent space [Ruder et al., 2019]. While monolingual embeddings model syntax and morphology, multilingual embeddings are vector representations of different languages mapped into a shared space, such that translated word pairs have similar vectors [Gouws et al., 2015, Luong et al., 2015]. They facilitate applications from parallel sentence extraction [Grover and Mitra, 2017] to machine translation [Zou et al., 2013, Cholakov and Kordoni, 2016, Artetxe et al., 2017] and can help improve monolingual embeddings [Faruqui and Dyer, 2014]. They can be learned via one of three methods [Erdmann et al., 2018]: mapping the different spaces into a shared space [Mikolov et al., 2013b], monolingual adaptation of one language’s embedding space into another’s [Zou et al., 2013], or multilingually training both embeddings simultaneously [AP et al., 2014, Pham et al., 2015].

This chapter will go beyond embeddings at the word level and pre-trained language models by exploring embeddings at different levels when combined with different composition functions. Therefore, sections 4.2, 4.3 and 4.4 respectively overview main existing approaches in word, morpheme and subword level representations, focusing on morphology-based models, embedding models as well as task-oriented models and contextual approaches, while sections 4.5 detail approaches targeting sentiment analysis, emotion detection and irony detection. Section 4.6 concludes the chapter with a reminder of this work’s contributions.

4.2 Word-Level Models

Neural word embeddings are prediction-based models, which have the capability of capturing the context of a word in a given text, semantic and syntactic similarity, as well as the relationship with other words. Word embeddings are vector representations of words, learned via the training of a neural network. To generate these distributed representations for words, the network learns its parameters by predicting the correct word (or its context) in a text window over the training corpus. While the point of network training is to learn good parameters, word vector representations follow the notion that similar words are closer to each other [Mikolov et al., 2013c]. This representation approach was first proposed in 2013, which aims at representing data in such a way to include both a word’s syntax and semantics [Mikolov et al., 2013c]. Several language models were proposed for learning word embeddings from raw unlabelled texts. The most popular models are GloVe [Pennington et al., 2014] and dependency-based word embedding [Levy and Goldberg, 2014]. The most commonly known and used is the Word2Vec model introduced by Google [Mikolov et al., 2013a, Mikolov et al.,

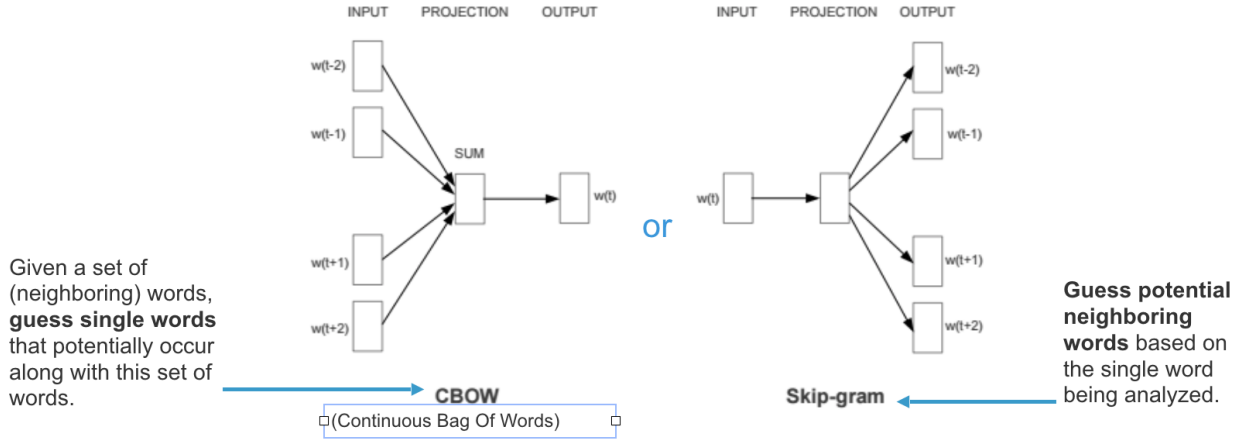


Figure 4.1: CBOW and Skip-gram models architecture.

2013c]. Word2Vec has two architectures for computing continuous vectors representations, the Skip-Gram (SG) and Continuous Bag-Of-Words (CBOW). The former predicts the target context-words from a given source word, while the latter does the inverse and predicts a target word given its context window. Both architectures have their own advantages and disadvantages (see Figure 4.1¹). Skip Gram is the best when working with a small amount of data and when tested gave the best representation of rare words. However, CBOW is faster and has better representations for more frequent words. Embedding vectors created using the Word2Vec algorithm have some advantages compared to earlier algorithms such as the latent semantic analysis [Mikolov et al., 2013a].

During training, the embeddings are gradually adjusted via backpropagation. The learned vectors will roughly encode similarities between words as learned for a specific problem on which the model is being trained (predicting the next sequence for Word2Vec). Passing an entity to an embedding layer substitutes each entity with the vector of the embedding matrix. Selecting the vector of a specific word can be translated to a matrix product in the following way:

$$\underbrace{Object_{vector}}_{dim \ 1} = \underbrace{M}_{dim \ words} \cdot \underbrace{\vec{v}}_{words} \quad (4.1)$$

Where v is the one-hot vector representing a text, determining which word needs to be transformed, and M is the embedding matrix.

In addition to these language models offering pre-trained embeddings, a word embedding can be learned as part of a deep learning model. Although this approach may be slower, it adapts, however the model to a specific training data set or task. Not only with just words, recently researchers have focused on generating embeddings for different word granularities to limit search space and yield low-dimensional representations that count for different morphological features also reducing the size of the output look-up table [Labeau and Allauzen, 2017] (see Sections 4.3 and 4.4). Others, although, have tried to exploit mutual information offered by the word itself and its components [Bojanowski et al., 2017], or ensemble embeddings [Al-Azani and El-Alfy, 2017].

[Almansor and Al-Ani, 2017] exploited both of Word2Vec's architectures for MSA and

¹Source <http://bit.do/idli-Word2Vec-architectures>

Egyptian text translation. While [Abdul-Mageed et al., 2018] proposed a SG model for MSA and different Arabic dialects, of the Twitter-City embeddings. They used a large corpus of 250M comments collected from 29 different cities of 10 countries (Oman, Egypt, Iraq, Jordan, Kuwait, Palestine, Qatar, KSA, UAE, and Yemen). The model uses a window of size 5 with minimal word frequency set at 100 words, and 300-dimensional vectors. [Elrazzaz et al., 2017] performed a comparison between four models, the first three were created by [Zahran et al., 2015] (CBOW, SG and GloVe) and the fourth, the Arabic Polyglot word embeddings, which was trained on Arabic Wikipedia content [Al-Rfou' et al., 2013]. They experimented with Arabic Word Analogy using a manually created benchmark consisting of nine relations (capital, currency, male-female, opposite, comparative, nationality, past tense, pair, plural), each consisting of over 100 words pairs. They also tested the four models on two NLP tasks using LSTM: Arabic document classification (Wikipedia articles were classified into four different classes: person, organization, location, and miscellaneous) and Arabic Named Entity Recognition (NER). CBOW and SG outperformed GloVe and Polyglot in practically all experiments.

[Erdmann et al., 2018] tried to maximize the informative content leveraged in each training sentence in order to address noisy data. They relied on word embedding created using SG FastText while varying the different parameters mainly "window size". Their multi-dialectal Arabic embedding models (Levantine, Gulf, and Egyptian, Maghrebi) ² uses two multi-lingual embedding methods, i.e. 1. training on one large corpus for all dialects (dialect agnostic) and 2. training on separate smaller corpora for each dialect then mapping them to a shared embedding space. While the first approach leverages fewer data per model, it is subject to more noise and ambiguity, as many words are unique to certain dialects or have disparate meanings in different dialects. The second can be seeded with a bidialectal dictionary or parallel sentences. They evaluated their model via bi-dialectal dictionary induction where the second approach performed better even for low-frequency words. [Elaraby and Abdul-Mageed, 2018] also tested several models for Arabic dialect identification (MSA, Egyptian, Gulf and Levantine) both binary and multi-way classification with different ML approaches (CNN, LSTM, BiLSTM and a convolutional LSTM). They relied on different word embedding models: randomly generated embedding, CBOW created using [Abdul-Mageed et al., 2018]'s corpus, as well as their model "Twitter-City".

[Severyn and Moschitti, 2015] on the other hand, experimented with different methods for initialising the parameter weights of a CNN. The network is initialised using Word2Vec English pre-trained weights, which are fine-tuned using a distant corpus as other parameters, and then are used to initialise the sentiment classifier with the same architecture.

[Nguyen and Le Nguyen, 2018] performed a ternary sentiment classification on a YouTube dataset (SenTube [Uryupina et al., 2014]) which contains comments and videos' titles from two domains (automobile and tablet) and two languages (English and Italian). They relied on a convolutional N-gram BiLSTM with a word embedding layer for predicting the type of comment, its sentiment, and whether the sentiment is directed toward a product or a video.

[Abidi and Smaïli, 2018] similarly used the CBOW approach to constitute an Algerian lexicon containing words from ALGD, French, English and MSA, for text translation using comments from Algerian Youtube channels. While [Qwaider et al., 2019] studied the feasibility of using MSA approaches and apply them directly to a Levantine corpus. The author relied on LSTM and BI-LSTM using the pre-trained AraVec models. During experiments they obtained 60% accuracy, however, when they tested different machine learning algorithms with a Levantine corpus they reached an accuracy of 75.2%, thus, drawing similar conclusions to previous works [Saadane and Habash, 2015].

Distributed word embeddings still treat words as atomic units and ignore any surface-

²<http://bit.do/camelEm>

or morphological similarities between different words, and uses a generic out-of-vocabulary representation shared between all unseen words [Rei et al., 2016]. [Bojanowski et al., 2017] believe that the popular continuous word representations ignore the morphology of words, by assigning a distinct vector to each word. They propose an approach based on the skip-gram model to generate the FastText pre-trained embeddings, where a word is represented as a bag of character n-grams. The model predicts the representation for each unit then calculates the sum of these vectors to obtain the word’s dense representation. The results on both word similarity and analogy tasks in nine languages including Arabic achieved state-of-the-art performances. [Wang et al., 2016] experimented with embedding clustering and convolutional neural network (CNN) to expand short texts. They experimented on two open benchmarks (Google Snippets [Phan et al., 2008] and TREC [Li and Roth, 2002]) using a variable window width to search for restricted nearest embedding neighbours.

The general framework to go beyond word-level vectorial representation follows two main steps [Zhu et al., 2019]: (1) Segment a word into basic units, (2) Compose the representation of these units to form the final representation of a word. Segmentation considers either morpheme as the basic semantic unit or character n-grams as the basic orthography unit. Composition methods vary from simple element-wise operations (like addition, average and multiplication) to complex neural networks that learn to compose embeddings. We review below main existing works splitting them according to the basic segmentation unit.

4.3 Morpheme-Based Models

These models generally rely on morphological analysers to segment the input into morphemes. Well known tools include Byte-Pair Encoding (BPE) [Gage, 1994], Morfessor [Creutz and Lagus, 2007] and CHIPMUNK [Cotterell et al., 2015], a supervised morphological segmentation system. [Luong et al., 2013] used Morfessor and RNN to learn morphologically-aware word embeddings while other works propose to learn morpheme and word embeddings simultaneously using addition as a composition function [Botha and Blunsom, 2014, Qiu et al., 2014]. BPE was used in [?] who experimented in embedding morpheme and character embeddings with CNN and BiLSTM for neural machine translation. This latter task has also been addressed in [Sennrich et al., 2016] who proposed a variant of BPE to handle rare or unseen words.

[Zhu et al., 2019] made a systematic comparison of CHIPMUNK, Morfessor and BPE in learning subword representation while varying the composition functions (addition, single head attention and multi-head attention) and the languages (English, German, Turkish, Finnish, Hebrew) resulting in 60 different configurations. Experiments made on three NLP tasks (word similarity, dependency parsing and entity typing) show that unsupervised tools (i.e., Morfessor and BPE) outperform supervised word segmentation (i.e., CHIPMUNK) for some configurations. The authors also conclude that different languages/tasks require different subword-informed configurations. This work draws similar conclusions in this work (cf. Chapter 7.7).

To limit the dependency to external tools, [Chaudhary et al., 2018] constructed word embeddings by adapting a FastText model to take into consideration phonemes, morphemes and graphemes. Embeddings are trained from a phonemic representation in terms of the International Phonetic Alphabet. Experiments on named entity recognition in four languages (Uyghur, Turkish, Bengali and Hindi) showed better results when compared to Word2Vec and FastText. [Cao and Rei, 2016] jointly learned word morphology and word embeddings based on a character-level model. [Üstün et al., 2018] applied a similar architecture showing that morpheme models are better at estimating word representations of morphologically complex words (with at least 2-3 suffixes).

The impact of morphology has also been investigated for the Arabic language [Baly et al., 2017b]. The authors trained Recursive Neural Tensor Networks (RNTN) with stems, automatic discretisation and marked letter repetition and achieved the best results when evaluated on the Arabic Sentiment Treebank.

4.4 Character N-Grams Models

As stated by several authors [Bojanowski et al., 2017, Li et al., 2018, Zhu et al., 2019, Vania and Lopez, 2017], embeddings at the word level suffer from two main issues. Firstly, they ignore the morphology of words treating different forms of the same word as being unrelated vectors. Secondly, the models are limited to learn vectors for the words present in the vocabulary only making systems fail to generalise when the OOV set of unseen words is large during the training phase. These limitations are particularly problematic for morphologically rich languages such as Arabic. Indeed, Arabic social media is often characterised by a large variation of unstandardised dialectal Arabic among Egyptian, Gulf, Levantine and Maghrebi [Zaidan and Callison-Burch, 2014], where speakers use Arabic letters but also Latin [Yaghan, 2008].

To overcome these limitations and account for the internal structure of words, subword embeddings where words are represented as character n-grams have become a popular solution [Zhao et al., 2018, Zhu et al., 2019, Bojanowski et al., 2017, Vania and Lopez, 2017]. These representations rely either on morphological information as given by external tools [Luong et al., 2013, Qiu et al., 2014] or built via supervised or unsupervised methods which learn the embeddings from text corpora [Cao and Rei, 2016, Bojanowski et al., 2017, Pinter et al., 2017, Li et al., 2018, Heinzerling and Strube, 2018]. Subword models have shown to be quite effective outperforming word-level baselines in several NLP tasks such as machine translation [Luong and Manning, 2016, Chung et al., 2016], part of speech tagging [Ling et al., 2015], word similarity [Li et al., 2018] and named entity recognition [Chaudhary et al., 2018]. Character-level on the other hand is viewed as a new paradigm in NLP applications using deep neural networks [Gridach, 2016] since it helped improve the performance of models in many NLP applications. It has been widely used in Neural Machine Translation (NMT), where results from recent works demonstrated that the integration of character-level features improves the translation results for many languages [Ling et al., 2015, Luong and Manning, 2016, Chung et al., 2016]. Character-based approaches have also been applied to other NLP tasks such as document classification [Xiao and Cho, 2016], language modelling [Kim et al., 2016, Ling et al., 2015], parsing [Ballesteros et al., 2015], as well as NER where it was used in text classification for learning representations of words from their characters [Zhang et al., 2015]. As far as we know, we are the first to use multi-level representations to build a system for Arabic OEA for social media.

4.4.1 Pre-trained Models

FastText [Bojanowski et al., 2017] is probably the most well-known model. It is based on skip-gram where a word is represented as a bag of character n-grams and its representation is obtained by summing the representation of its n-grams. Pre-trained word vectors exist for 157 languages (including Arabic), all trained on Common Crawl and Wikipedia. BPEmb [Heinzerling and Strube, 2018] is another collection of pre-trained subword embeddings based on BPE and available in 275 languages. The models are trained with a subword vocabulary (containing the word itself) with different sizes using a Word2Vec model. For the

Arabic language, eight vocabulary sizes ranging from 1,000 to 200,000 are provided³. These models were trained on texts collected from Wikipedia, meaning that the Arabic models are trained on MSA texts only. [Heinzerling and Strube, 2018] showed that BPEmb outperforms subword-based approaches (i.e., FastText) for some languages when evaluated on entity typing task.

A new embedding representation emerged in late 2018 as an alternative to "static embeddings" to better capture word meaning in contexts. Known as "Transformers" or "contextual embeddings", which gained popularity rapidly. Emphasising the contextual features of words in texts, where more than one representation is provided for each word based on the context in which it appears [Smith, 2019], which ultimately can help overcome ambiguity problems [Devlin et al., 2019]. BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019] and ELMo (Embeddings from Language Models) [Peters et al., 2018b] are the best examples of such language models. These word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on large text corpora.

Bert, a language model introduced by Google and recently added to the Tensorflow hub, simplifying its integration in Keras models [Devlin et al., 2019], is based on the idea of training a separate language model to produce better contextual word representation. BERT is based on Transformers [Vaswani et al., 2017], a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection which is called attention, a concept which helps with alignment in seq2seq architectures for translation to capture the relation between words of a sentence. BERT transforms a text into a set of token embeddings, segmentation embeddings and position embeddings and then outputs the sum of these three embeddings. It uses character-based tokenisation for Chinese, and WordPiece tokenisation for all other languages. The same idea has been adopted in [Akbik et al., 2018] who introduced the Flair framework, a contextual string embeddings leveraging pre-trained character-level language models from which they extract hidden states at the beginning and end characters positions of each word to produce embeddings for any string of characters in a sentential context of an RNN architecture. As for BPEmb, Flair character embeddings for the Arabic language were trained on Wikipedia. Contextual embeddings have been used in several NLP tasks including question answering, NER and language inference [Devlin et al., 2019], as well as irony detection [Kayalvizhi et al., 2019, Moudjari and Akli-Astouati, 2019b, Ranasinghe et al., 2019, Zhang and Abdul-Mageed, 2019], dialect identification [Zhang and Abdul-Mageed, 2019], and sentiment analysis [Al-Twairish and Al-Negheimish, 2019].

Another language model worth mentioning is the newly published AraBert language model [Antoun et al., 2020] as well as the Universal Language Model in Arabic hUL-MonA - [ElJundi et al., 2019], where the AWD-LSTM state-of-the-art language model [Merity et al., 2017] is pre-trained on a large Wikipedia Arabic corpus to provide the distributional representations of each word. Using transfer learning, the model is further tuned on different sentiment corpora. Experimental results show that hULMonA achieves comparable results with BERT multilingual language model when dealing with both MSA and dialects for sentiment classification.

4.4.2 Learning Subword Embeddings

Recently, based on the assumption that a word's meaning can be reconstructed from its parts, several subword-based methods have been proposed to deal with the unseen words

³<https://nlp.h-its.org/bpemb/ar/>

problem such as character-based RNN [Luong and Manning, 2016], character-based LSTM-CNN [Chiu and Nichols, 2016], word embeddings enriched with subword information (Fast-Text) [Bojanowski et al., 2017], and byte-pair encoding [Sennrich et al., 2016]. Early models exploited stochastic neural networks such as restricted Boltzmann machines, which consists of a visible and a hidden layer of neurons with unidirectional connections between the layers and no inner layers. [Sperr et al., 2013] exploited such architecture to improve German-English translation. The input to their model is characters index and different character n-grams (1-4), where the character unigram model helps to map several words represented using character index vectors into similar character embedding vectors. Others used the distributed language models (i.e. Word2Vec) to create different word units representations such as radical [Shi et al., 2015].

The use of character level has been shown to help improve model performance in many NLP applications where it is considered as a new paradigm using such features. It is widely used in NMT where recent works show that adding character-level features improves translation results for many languages [Costa-jussà and Fonollosa, 2016, Chung et al., 2016, Luong and Manning, 2016]. As a result, it has been applied to other tasks, such as document classification [Xiao and Cho, 2016]. However, the character-level for OEA was not explored.

Others exploited embeddings jointly learned from sub-words and words units, defining simple compositional architectures (often based on the addition) to create word embeddings from sub-word embeddings [Botha and Blunsom, 2014, Bojanowski et al., 2017], or from characters associated with pre-trained word representations [Santos and Zadrozny, 2014], who used a CNN architecture for POS tagging.

Table 4.1 summarises the main state of the art approaches to learning subword models in different languages (first half of the table) highlighting in particular works targeting the Arabic language (second half). For each reference, the table details commonly used units to segment a word, the functions used to compose these units as well as the NLP tasks on which the learned models have been evaluated.

From Table 4.1, it can be seen that segmentation unit ranges from characters [Wieting et al., 2016, Heigold et al., 2017, Pinter et al., 2017, Paetzold, 2018, Zhai et al., 2018, Kim et al., 2016, Xu et al., 2019] to the word itself, including character n-grams [Sperr et al., 2013, Santos and Zadrozny, 2014, Miyamoto and Cho, 2016, Rei et al., 2016, Lal et al., 2019, Li et al., 2018, Joshi et al., 2016, Zhao et al., 2018, Wu et al., 2016] and radicals [Shi et al., 2015]. Several units are often combined for better performances. For example, [Chen et al., 2015] jointly learn character and word embeddings based on the CBOW model where multiple-prototype character embeddings (position-based and cluster-based) are proposed to deal with the ambiguities of Chinese characters.

[Miyamoto and Cho, 2016] used an RNN-LSTM based architecture to create word embeddings using words and their constituent characters. Their model generates two vectors for each word which are combined into a single word representation through a gated layer, the latter affects the balance between the word-level and character-level inputs. These representations are created using character and word embeddings from a Bi-LSTM model that uses a look-up table initialised using a hyperbolic tangent function proposed in [Glorot and Bengio, 2010].

[Rei et al., 2016] exploited a similar Bi-LSTM architecture to create word representations, where the two vectors of a word are combined through an attention layer using a weighted sum, while [Joshi et al., 2016] relied on a convolutional-LSTM architecture that generates subword representations from their characters using convolutional filters.

[Papay et al., 2018] experiment with four models among which a Skip-Gram model with a CNN architecture to learn representations for each character which are then summed to obtain a word embedding, and a concatenation of pre-trained word and character embeddings

Table 4.1: Summary of main existing subword models. D means that a given reference treats Arabic dialects without targeting a specific dialect while A stands for MSA.

Models	Segmentation units	Composition functions	Evaluation tasks	Languages
[Sperr et al., 2013]	char n-grams (length 1-4)	RBM ⁴	Machine translation	German-English
[Santos and Zadrozny, 2014]	Char, Word	CNN	POS tagging	English, Portuguese
[Shi et al., 2015]	Radical, word, Char	Concat.	Text categorisation, word segmentation, search ranking	Chinese
[Chen et al., 2015]	Char, word	Sum, clustering	Word relatedness, analogical reasoning	Chinese
[Miyamoto and Cho, 2016]	Char, word	GRU-Bi-LSTM	Language modelling	English
[Rei et al., 2016]	Char, word	Sum attention	Sequence labelling	English
[Wieting et al., 2016]	Char	Sum	Word/sentence similarity, POS tagging	English
[Joshi et al., 2016]	Char n-grams	CNN	Sentiment analysis	Hindi, English
[Wu et al., 2016]	Wordpiece, Word/Character	Concat.	Machine translation	4 languages
[?]	Chars, word	max pooling	Machine translation	Estonian, Russian, English
[Labeau and Allauzen, 2017]	Word, Chars, 3-gram Chars	Concat, LSTM, Bi-LSTM, CNN	Language modelling	Czech
[Li et al., 2018]	Chars n-grams	Concat., CNN, RNN,	Word similarity/Analogy, Affix Prediction, POS, Chunking, NER, Sentiment analysis	English
[Papay et al., 2018]	Word, Char	Sum, Concat.	Word relatedness	English, German
[Zhai et al., 2018]	Char	LSTM, CNN	NER	English
[Lal et al., 2019]	Char n-grams	CNN	Sentiment analysis	Eng. Hind.
[Xu et al., 2019]	Char, type, POS	Bi-GRU RNN	Time normalisation	English
[Gridach, 2016]	Char, word	Bi-LSTM	NER	Arabic ^D
[Vania and Lopez, 2017]	Word, Char, 3-grams, BPE, Morfessors	Sum, Bi-LSTM	Language modelling	10 languages ^A
[Heigold et al., 2017]	Char	CNN, LSTM	Morphology tagging	14 language ^A
[Pinter et al., 2017]	Char	BiLSTM	POS/Morpho-syntactic attribute	23 language ^A
[Kim et al., 2018]	Char	CNN	Word similarity, language modelling	6 languages ^A
[Zhao et al., 2018]	Char n-grams	LSTM	Word similarity, POS/morphosyntax	23 languages ^A
[Paetzold, 2018]	Char	RNN	Emotion detection	3 languages ^D
[Watson et al., 2018]	Char, word	Concat.	Text normalisation	Arabic ^A
[Adouane et al., 2018]	Char	None	Language identification	Algerian dialect
[Erdmann et al., 2018]	Char n-grams (length 2-6)	Sum	Dictionary induction	Levantine, Gulf, Egyptian
[Shapiro and Duh, 2018]	Lemma, Chars n-grams	Sum, dot product	Machine translation	Arabic ^A , English
[Dahou et al., 2019]	Word, Char n-grams (length 3-6)	Sum	Sentiment analysis	Arabic ^{A,D}
[Elaraby and Zahran, 2019]	Char	None	Dialect identification	Arabic ^D

with those learned with CNN. Concatenation performed the best when evaluated on English and German small and large corpora. [Labeau and Allauzen, 2017] created word embedding using three neural architectures (CNN, LSTM, BiLSTM) relying on different subword decompositions for Czech (word, character, 3-gram characters, and lemma + tags). While [Zhai et al., 2018] compared the use of LSTM-based and CNN-based character-level word embeddings and POS features in a Bi-directional Long-Short Term Memory Conditional Random Field models (BiLSTM-CRF) model to predict chemical and disease named entities. The BiLSTM-CRF models gave the best results compared to BiLSTM, BiLSTM-CNN and BiLSTM-LSTM as well as BiLSTM-CRF-CNN and BiLSTM-CRF-LSTM.

The representation obtained from units composition can also be augmented with additional extra-features, as in [Lal et al., 2019] who add surface features (word frequency, count of positive/negative sentiments, presence of exclamation, aggregated sentiments from SentiWordNet) to the subword representation of a sentence to predict sentiment in code-mixing Facebook comments (English-Hindi) using a CNN-BiLSTM. Overall, composition via deep learning architectures is widely used compared with addition (sum) or concatenation.

Most efforts target one or two languages at the most. A notable exception is the Mimick model by [Pinter et al., 2017] which is an LSTM-based word representation that learns a function from characters to word embeddings to handle OOV words in 23 languages.

[Zhao et al., 2018] propose a generalisation of word embedding using a bag of words (containing characters and subwords) to tackle rare or unseen words in 23 languages including Arabic. The approach is similar to FastText although instead of creating the embeddings from unlabeled data, they are created using an LSTM architecture relying on pre-trained word embeddings from the Polyglot vectors [Al-Rfou’ et al., 2013] and Word2Vec. The model predicts word vectors from the learning set of vectors then generalises the process for unseen words as part of the word similarity task and in the joint prediction of POS tag and morphosyntactic attributes.

[Kim et al., 2016] proposed a character-aware language model⁵ combining CNN and LSTM, which predicts the next word in a sentence giving a target word in input. The model gave encouraging results for different morphologically rich languages: English, Arabic, Czech, French, German, Spanish, and Russian.

[Kim et al., 2018] relied on a similar architecture, where they concatenate the vectors of each character obtained through CNN to construct vectors for words which are then optimised by applying a Euclidean distance with Word2Vec pre-trained vectors for 6 languages (Arabic, German, English, Spanish, French, and Russian). However, when compared with FastText, the model fails to handle morphological rich languages such as Arabic and Russian. [Li et al., 2018] obtained similar results, where FastText outperformed the CNN and the RNN-based models. [Heigold et al., 2017] exploited both CNN and LSTM to create character embeddings for morphology tagging in 14 languages from 9 language families, including Semitic (Arabic). The results indicated that the LSTM’s architecture is slightly better than CNN’s for creating character embeddings. However, when the two architectures are assembled by the geometric mean of the scores, small gains were observed.

Finally, a comparison of different segmentation units (word, character, tri-grams, BPE and Morfessors) has been carried out in [Vania and Lopez, 2017]. They experimented with several models based on LSTM while varying the units and the composition functions (addition, Bi-LSTM and CNN). The results show that the model based on tri-grams learned with Bi-LSTM performs the best for most languages including Arabic. In some languages, however, the models with explicit knowledge of morphology achieved better than character-level models.

Indeed, for the Arabic language, FastText [Bojanowski et al., 2017] where each word

⁵<https://github.com/yoonkim/lstm-char-cnn>

is represented as a bag-of-character n-gram, has been widely used: [Erdmann et al., 2018] proposed a multi-dialectal word embeddings (Levantine, Gulf and Egyptian) and use it for dictionary induction, [Kim et al., 2018] experimented with word similarity while [Shapiro and Duh, 2018] neural machine translation from Arabic to English.

Other studies developed their neural architectures learning a combination of character and word embeddings. These models are generally evaluated on a single task and mostly dialect-agnostic. Among them, we cite [Watson et al., 2018] who used a sequence-to-sequence attention model with gated recurrent units to learn character embeddings and concatenate them with pre-trained word embeddings (FastText, Word2Vec) for text normalisation, [Gridach, 2016] experimented with named entity recognition and [Elaraby and Zahran, 2019] dialect identification. We also cite [Paetzold, 2018] and [Dahou et al., 2019] that proposed character-level embeddings for classifying short texts according respectively to their polarity and emotions. Most studies target Middle-East dialects, although few investigate character embeddings for North African dialects, such as Algerian [Adouane et al., 2018].

Prior work in the field targeted Arabic named entity recognition in a social media dataset showing that the use of Bi-LSTM with character, word embedding and Conditional Random Field outperforms Bi-LSTM with the word or character embeddings alone [Gridach, 2016]. Since then, both character-based [Elaraby and Zahran, 2019, Adouane et al., 2018] and character n-grams models [Dahou et al., 2019, Erdmann et al., 2018, Shapiro and Duh, 2018, Watson et al., 2018] were explored, all focusing on social media texts, except [Shapiro and Duh, 2018] who worked on MSA texts taken from TED subtitles and Wikipedia. Few works addressed Arabic dialects. Among them, [Adouane et al., 2018] performed language identification in mixed-language Algerian dataset, while [Elaraby and Zahran, 2019] built character embeddings via CNN and Bi-LSTM for Arabic dialect identification. Finally, [Erdmann et al., 2018] proposed multi-dialectal embeddings for dictionary induction using a monolingual adaptation from one language embedding space into another for each pair of dialects (Levantine, Gulf and Egyptian). They relied on the FastText Skip-Gram model, where a word’s vector is the sum of its vector and all its component character n-grams of length $[2 - 6]$.

In addition to the sum, [Shapiro and Duh, 2018] experimented with a modified SG objective using the dot product of words and their lemma representations.

4.5 Arabic Opinion and Emotion Analysis

Sentiment analysis is an omnipresent challenge in NLP. It has been applied to many tasks where several methods have been proposed. In recent years dense representations became very popular, where the distributed word representations to substitute the traditional bag-of-words approach has been widely utilised very efficiently for many NLP tasks including OEA. Where word embeddings were used effectively in several languages, English being the first [Maas et al., 2011, Poria et al., 2016]. Other European languages have taken advantage also, to improve their systems such as Spanish [Rosas et al., 2013], French [Saroufim et al., 2018, Htaït, 2018] and Greek [Giatsoglou et al., 2017].

These representations help map similar words into similar regions based on their context. Therefore, morphologically rich languages such as Arabic (both MSA and DA) also exploited such distributional vector representations that encode syntactic and semantic properties of words into dense vectors. The most popular representations in Arabic are word embeddings, where several published works relied on Word2Vec’s both architectures [Alayba et al., 2018b, Dahou et al., 2016], where the CBOW was mostly used [Altowayan and Tao, 2016, Elaraby and Abdul-Mageed, 2018]. Therefore, the AraVec project [Soliman et al., 2017] was very utilised, since it offers different models using both Word2Vec architectures with different

embedding sizes. AraVec [Soliman et al., 2017] is a set of 12 pre-trained models built on top of three different Arabic content domains; Tweets, World Wide Web pages and Wikipedia Arabic articles. The models were tested on several NLP tasks (sentiment words clustering, NER clustering, semantic textual similarity).

Opinion and emotion analysis are tasks that have intrigued researchers over the last two decades, during which several approaches have been proposed and tested. In this section, we describe the methods used for the analysis of opinions and emotions in Arabic texts in general and dialects in particular, focusing on the different forms of vectorial representations and discuss their utility for capturing semantic features from different sources.

Early work in the field targeted MSA, where [Altowayan and Tao, 2016] used news articles and consumer reviews to build word representations relying on the CBOW architecture of Word2Vec to perform binary SSA (positive/negative and subjective/neutral), while [Alayba et al., 2018b] used both architectures CBOW and SG. [Dahou et al., 2016] also created word vectors using Word2Vec CBOW and SG from collected Arabic web pages, targeting both MSA and DA). In a recent work, [Dahou et al., 2019], exploited a multi-channel CNN to combine pre-trained (AraVec and FastText) and domain-specific word embeddings for Arabic sentiment classification (MSA and dialectal) which gave encouraging results, specifically for imbalanced datasets. [Al-Khatib and El-Beltagy, 2019] however tried to fine-tune pre-trained models (AraVec and Google’s English news embeddings⁶) for text classification tasks, by providing text’s labels as an additional contextual variable, assuming that values of individual word vectors are then conditioned by classes in which they frequently appear, as well as the document representation given by a Doc2Vec CBOW model. They experimented with different sentiment and emotion classification tasks using an LSTM classifier. The results showed some improvements for single-label classification tasks, however, a decrease in its performances was noticed when applied on a multi-label classification task (SemEval 2018 task 1 [Mohammad et al., 2018]). Indeed, recently shared tasks such as the 2017 SemEval Task 4 and SemEval 2018 task 1 for Arabic OEA provided substantial dialect-agnostic datasets, a step towards supporting Arabic OEA research field.

More recently, researchers were more interested in dialectal Arabic, where [Al-Azani and El-Alfy, 2017] performed a binary SA on a Syrian dataset [Mohammad et al., 2016b] using ensemble learning with CBOW pre-trained embeddings⁷. While [Qwaider et al., 2019] experimented with Levantine relying on AraVec using LSTM and Bi-LSTM, outperforming BOW with SVM. [Fouad et al., 2020] on the other hand proposed the "ArWordVec" models, where they investigated different language models for building word embeddings of Arabic tweets including Word2Vec both architectures and GloVe. Results on word similarity and SA respectively showed that CBOW and SG-based models recorded the highest scores. [Kaibi et al., 2019] also performed a comparison between different language models (FastText, GloVe and Word2Vec) using web pages MSA content from the OSAC corpus [Saad and Ashour, 2010], and experimented with tweets binary SA using the ASTD [Nabil et al., 2015] and ArTwitter [Abdulla et al., 2013] corpora with six ML algorithms, namely: GaussianNB, LinearSVC, NuSVC, LR, stochastic gradient descent and Random Forest (RF). The use of FastText with NuSVC (a type of SVM classifier) outperformed other combinations in terms of accuracy. In [Al-Twairesh and Al-Negheimish, 2019], the authors relied on surface features (tweet-length, emoticons, count of positive and negative words, intensifiers, tweet score using the intensity values of the Arabic sentiment lexicon and Arabic translations of the Bing Liu’s and MPQA) which are manually extracted, as well as generic word embeddings and sentiment-specific word embeddings, for Arabic tweets SA. To construct word embeddings, they collected 10M Arabic tweets using positive and negative seeds, and thus

⁶<https://code.google.com/archive/p/Word2Vec/> accessed 09/09/2020.

⁷<https://github.com/iamaziz/ar-embeddings>

automatically labelling the tweets. The experimental results using an SVM classifier demonstrate that their sentiment embeddings and generic word embeddings models surpass the reported results of the three used datasets (semEval2017, AraSenTi and ASTD). However, tests using BERT showed a decrease in the system’s performances. While [Alnawas and Arici, 2019] explored Doc2Vec embeddings, based on the paragraph and distributed memory model of paragraph vectors architecture, using different data (tweets, reviews). They explored the representation features using four binary classifiers (LR, DT, SVM and NB) to detect sentiment in Iraqi texts. Finally, integrated embeddings with the classifier learning were tested on different corpora. Where [Mulki et al., 2019] presented syntax-ignorant n-gram embeddings for sentiment analysis of different dialects namely: Jordanian (ArTwitter [Abdulla et al., 2013]), Tunisian (TEC [Sayadi et al., 2016], TSAC [Medhaffar et al., 2017]), Moroccan (MEC [Elouardighi et al., 2017]) and mixed dialects (MDT [Altowayan and Tao, 2016]). While [Oussous et al., 2019] proposed a framework for pre-processing and SA-2 of Moroccan texts at the sentence-level.

The SemEval2018 task 1 corpus [Mohammad et al., 2018] annotated manually for 12 emotion classes was used in different works [Mohammad and Kiritchenko, 2018b, Alhuzali et al., 2018, Abdullah et al., 2018, Paetzold, 2018, Baali and Ghneim, 2019]. Where word embeddings, affect lexicon features, and word n-grams were some of the most used features. Indeed, [Alhuzali et al., 2018] used a deep gated RNN with integrated word embeddings from the task set. [Abdullah et al., 2018] used a CNN-LSTM fully connected architecture to classify the tweets in their respective sentiment and emotion classes relying on word-vectors. While [Paetzold, 2018] relied on character-to-word embeddings as part of an RNN classifier to create word embeddings for both emotion regression and classification of Arabic, English and Spanish, as well as Word2Vec CBOW to create word embeddings. [Baali and Ghneim, 2019] however presented a CNN-based model which uses word embedding trained on the same dataset with Word2Vec. The experiments showed that the CNN model outperformed SVM, NB and multi-layer perceptron (MLP).

IDAT (Irony Detection in Arabic Tweets) participants used either feature-based methods with traditional surface features machine learning approaches (SVM, Multimodel Naive Bayes, Logistic Regression with TF-IDF, and n-gram frequency) and deep learning methods (CNN, RNN, LSTM, Gated Recurrent Unit, Transformers) relying on different language models such as Word2Vec, FastText and BERT [Moudjari and Akli-Astouati, 2019b, Ranasinghe et al., 2019, Kayalvizhi et al., 2019]. The best two systems used an ensemble model (Gradient Boosting, RF and MLP) with surface features (bag of words, TF-IDF, topic modelling) [Khalifa and Hussein, 2019] and BERT in a multi-task learning configuration, BERT being pre-trained on a dialectal Twitter dataset [Zhang and Abdul-Mageed, 2019]. Such works encourage the idea of exploring sentiment analysis approaches and apply them to irony detection. Therefore, the models proposed in this work are evaluated in the IDAT2019 Task. In a previous work, [Moudjari and Akli-Astouati, 2019b], two models were used, incorporating various sentiment features to capture ironic statements. The results revealed good classification performance on the training dataset. The results were very encouraging. As far as we know, multi-level embeddings for Arabic irony detection is explored here for the first time.

4.6 Conclusion

Continuous word dense vector representations encode syntactic as well as semantic regularities of the language, that are presented as relation offsets in the vector space [Mikolov et al., 2013a]. Conclusions obtained in other studies for Arabic and other Arabic dialects confirm that DL models substantially improves the performance of sentiment classification (MSA

[Alayba et al., 2018a], Egyptian [Alayba et al., 2018a], Levantine [Elnagar et al., 2018], Tunisian [Mulki et al., 2019], and Moroccan [Oussous et al., 2019]). [El-Halees, 2017] showcased that word-embedding representations performed better in terms of accuracy on four different datasets compared to n-grams with the same classification model. [Badaro et al., 2018a] also found that AraVec word embeddings [Soliman et al., 2017] perform better than other features with ensemble classification. The same conclusions were drawn in [Altowayan and Tao, 2016, Altowayan and Elnagar, 2017]. Similarly, [Al-Azani and El-Alfy, 2018] observed that simpler models consisting of word embeddings and emojis performed better than typical n-grams. [Qwaider et al., 2019] evaluated the use of the Arabic word embeddings (AraVec) with LSTM and Bi-LSTM. Among these pre-trained word embeddings, very limited number are publicly available, among [Soliman et al., 2017, Altowayan and Tao, 2016], which outperformed classical ML models and statistical BOW. These distributed word representations are limited to learn vectors for the words present in the vocabulary set, making systems fail to generalise when the set of unseen words is large during the training phase, and, thus vulnerable against the Out-Of-Vocabulary problem. For they still treat words as atomic units and ignore any surface- or morphological similarities between different words, by treating different forms of the same word as being unrelated vectors and assigning a generic out-of-vocabulary representation, which is shared between all other unseen words [Rei et al., 2016, Bojanowski et al., 2017]. These limitations are particularly problematic for morphologically rich languages such as Arabic. Therefore, different word granularities representations (morpheme, character n-grams) were explored, either pre-trained or learned and have shown to be quite effective outperforming word-level baselines and overcoming such shortcomings in several NLP tasks such as machine translation, part of speech tagging and named entity recognition. Recently, the use of fine-grained features to perform text classification showed that the use of character-level embeddings improves the performance of the models, by enhancing their ability in recognising unseen words. These models discard the segmentation into words and aim at learning language representations directly from characters [Bojanowski et al., 2017, Papay et al., 2018]. They do not have the limitation of vocabulary, hence can freely learn to generate new words. Although others believe that research into character-level models is still in fairly early stages, and models that operate exclusively on characters are not yet competitive to word-level models on most tasks [Rei et al., 2016]. Others, relied on different character n-grams composition functions (addition, mean, DL) [Bojanowski et al., 2017, Zhao et al., 2018, Dahou et al., 2019]. Overall, composition via deep learning architectures is widely used compared with addition or concatenation. Contextual embeddings, on the other hand, have been proposed as an alternative to "static embeddings" to better capture word meaning in contexts, and thus, tackled contextual ambiguities by offering different vector representations of a given word depending on its contextual use [Smith, 2019].

These models are generally evaluated on a single task and mostly dialect-agnostic and most efforts target one or two languages at the most. Moreover, even though research in Algerian opinion analysis is recently getting more attention from the research community, most of the existing work using embeddings for the Algerian dialects (ALGD) is dedicated to text translation and language identification [Almansor and Al-Ani, 2017, Abidi and Smaïli, 2018].

As far as we know, no one has conducted an analysis of the impact of different unit information across different Arabic dialects and different tasks. Here it is proposed for the first time in Arabic NLP an in-depth study of the impact of several units configurations and explore different composition functions to obtain the final representation of a given text. To evaluate the effectiveness of this approach, extrinsic evaluations were performed on different text classification tasks while varying the dialects.

We end this chapter with a reminder of this research’s main contributions to this line of

works.

- **A multi-level representation of Arabic texts.** This work proposes various neural architectures to learn character, subword and word embeddings, as well as a combination of these three levels.
- **Compare these representations with state of the art pre-trained models.** Comparison with several pre-trained embeddings at different levels (word, character, subword and contextual) was performed to further validate the findings in this work.
- **A systematic evaluation of these representations across *different dialects*.** MSA was targeted as well as Levantine and Maghrebi dialects using both dialect-agnostic and dialect-aware embeddings to investigate whether embeddings learned from a given dialect can be successfully used to classify texts written in other dialects.
- **A systematic evaluation of these representations in *three tasks*,** namely sentiment analysis, emotion detection and irony detection. For each task, experiments with well-known datasets were conducted, including those that have been recently used in shared tasks to better compare the results with those reported in previous studies on the same datasets.

Part II

Contribution

Chapter 5

An Algerian Corpus and an Annotation Platform for Opinion and Emotion Analysis

Most of the recent research undertaken in Arabic is not yet releasing their resources, where most of them lack sentiment analysis' features and contain domain-dependent social media content (news, product reviews). In addition, the largest share of publicly available resources is dedicated to MSA only and excludes dialects and their complexity. Few corpora have been made available for specific dialects, where Algerian dialect resources are scarce. Therefore, it is of utmost importance to build more resources to help advance the research on ALGD opinion and emotion analysis.

In this chapter, we address the lack of resources for opinion and emotion analysis related to North African dialects, targeting Algerian dialect. We present TWIFIL (TWitter proFILing) a collaborative annotation platform for crowdsourcing annotation of tweets at different levels of granularity. This platform allowed the creation of the largest Algerian dialect dataset annotated for both sentiment (9,000 tweets), emotion (about 5,000 tweets) and extra-linguistic information including author profiling (age and gender). The annotation also resulted in the creation of the largest Algerian dialect subjectivity lexicon of about 21,000 entries which can constitute a valuable resource for the development of future NLP applications for the Algerian dialect.

Keywords Crowdsourcing annotation platform, Algerian dialect, NLP, sentiment analysis, emotion detection, subjectivity lexicon, corpus creation, lexicon, annotation platform.

5.1 Introduction

North African dialects, including the Algerian dialect, are less normalised compared to MSA. They have been enriched by many languages over the years, which resulted in a complex linguistic situation and scarce resources. The linguistic diversities of ALGD call for special attention, which is why the spoken and written dialects are very rich and varied languages.

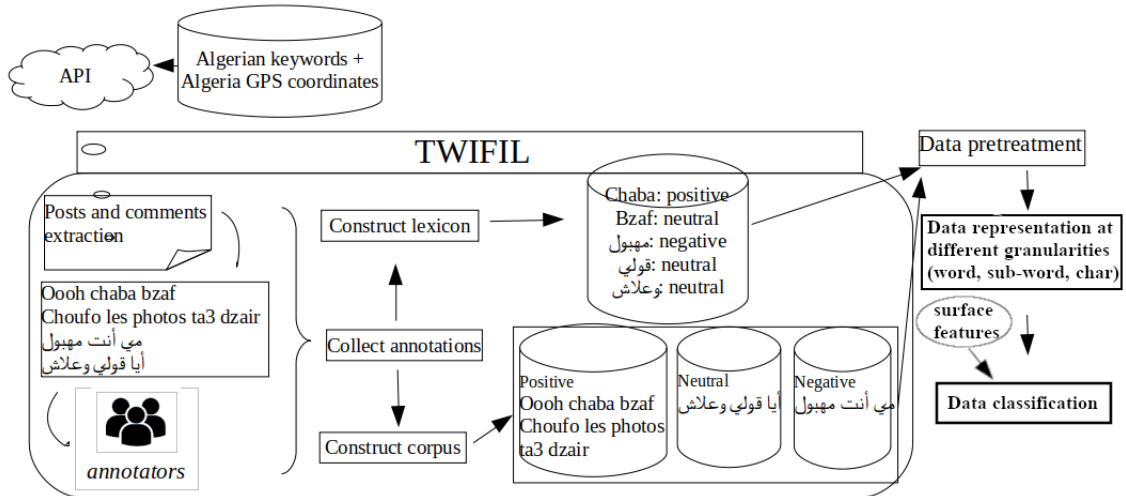
Publicly available corpora contain predominantly social media content depending on the domain (news, product reviews), which resulted in Arabic resources mostly of MSA with also a mix of MSA and dialectal Arabic. Few rare corpora were made available for one dialect, especially for the ALGD. Also, the Arabic language remains under-resourced as to these polarity repositories compared to the English language. It is clear from studying related works (see Chapter 1.6), publicly available resources for sentiment analysis in ALGD are scarce. Those available such as [Mataoui et al., 2016], gives only the polarity of the comments collected, without any information on the emotion expressed or its author, and

are generally constructed manually by the researcher or by a small number of annotators. Tools and resources are essential if progress is to be made in this field of research. In this chapter, the aim is to bridge the gap by addressing the lack of resources for opinion and emotion analysis related to North African dialects, targeting specifically, the Algerian dialect. To ensure the credibility of resources, using crowd-sourcing was considered. Hence, an open collaborative platform was created for manual annotation of tweets at different levels of granularity (word and document level), which we called "TWIFIL" (TWitter proFILing). Thus, this research proposes the first and largest Algerian corpus annotated both at the level of sentiments and emotions as well as at the level of extra-linguistic information (age, gender, topic). Indeed, the platform allowed the creation of the largest Algerian dialect dataset annotated for both sentiment (9,000 tweets), emotion (about 5,000 tweets) and extra-linguistic information including author profiling (age and gender). The annotation also resulted in the creation of the largest Algerian dialect subjectivity lexicon of about 21,000 entries which can constitute valuable resources for the development of future NLP applications for Algerian dialect. The three main contributions devoted to the development of resources are:

- A crowd-sourcing annotation platform.
- Multi-grained annotations were done at both word and tweet level.
- Multi-level annotations including sentiment, emotion and extra-linguistic information (age, gender and topic).

Fig. 5.1 presents a schema of the work detailed in this chapter. As shown the posts collected through the Twitter API are annotated where the annotators provide the annotation at both word and tweet level. This helps create a lexicon and a corpus. These resources are then exploited to perform different opinion and emotion classification tasks. The three main resources developed are: 1) "*Twilex*". An Algerian sentiment Lexicon (containing words written in both Arabic and Latin letters with their polarity/emotion/different spellings). 2) "*AlgSet*". A corpus of the Algerian spoken languages and dialects (Arabic, French, English). 3) "*Twifil*". An annotated Algerian corpus for opinion and emotion detection.

Figure 5.1: A general architecture of our work for OEA of the DALG.



In order to layout the first contribution this chapter is organised as follows: in section 5.2 we present the process of collecting our data, we then describe the annotation platform in section 5.3 as well as the challenges and rules of annotations. Sections 5.4 and 5.5 provide

details regarding the proposed annotated corpus and lexicon respectively. Finally, section 5.6 concludes the chapter and provides some perspectives regarding the enhancement of the proposed tools and resources.

5.2 Data Collection

Many recent NLP researchers targeting social media analysis and applications have used Twitter as the main data resource for carrying out their work. This could be attributed to several reasons including the fact that Twitter can be easily queried within specific periods. Tweets are often tagged with geo locations and provide users' information.

Therefore, Twitter was chosen as the source of data for this research since tweets are self-contained, widely used, public posts, and tend to be rich in emotions [Mohammad and Kiritchenko, 2018a]. When addressing Arabic within the context of social media, it could be said that it is facing many different linguistic domains as seen previously (see Chapter 1.2). This is largely due to how social media users express themselves using a variety of dialects and sub-dialects. Examples include MSA, Egyptian or more specifically Cairene, Gulf, Moroccan, Tunisian, Algerian, and Levantine dialects. Capturing as many of these linguistics domains within the context of daily social interactions was one of the main motivations for creating and using dense representation models. To gain access to a wider range of tweets than those provided by the standard Twitter API, a crawler was designed to collect tweets. To respect users' control and privacy, the crawler fully complies with the rules which specify what can be queried from Twitter¹. The crawler was used to query tweets based on four main parameters: (1) the language which was set to Arabic, (2) the time span permitted to conduct the search, (3) geolocalisation set to Algeria coordinates, and (4) a set of keywords about popular Algerian politicians (e.g. بوتفليقة (Bouteflika)), artists (e.g. منال حدلي (Manel Hadli)), athletics (e.g., عادل عمروش (Adel Amrouche)), locations (e.g. دالي إبراهيم (Dely Ibrahim)) and some important events that occurred in Algeria (e.g. الصّالون الدولي للكتاب بالجزائر (Algiers International Book Fair)). When available, the keyword search was coupled with tweet geolocalisation as given by the metadata to target tweets posted from Algeria only. More than 10,000 search queries were used to collect our final Twitter dataset. In the end, more than 10,000,000 tweets were collected (textual contents as well as tweets' metadata) posted between 2015 and 2019 obtained from different random geo locations using the Twitter API and the Twitter stream API. After cleaning and removing redundant tweets, the final corpus counted around 3,000,000 tweets. Which were then divided to two-part, one for annotation (~500,000) and one for embedding creation (~2,300,000). More details are provided in the next sections.

5.3 The Annotation Platform: TWIFIL

This section presents the platform as well as the approach followed to annotate the *Twifil* corpus. It starts with a specification of the annotation goals, followed by a discussion of the general approach. Then discusses in more detail the important challenges that need to be addressed for annotating the new dialectal lexicon *Twilex*.

TWIFIL (TWIter representing the social media and FIL from the profile, meaning giving a profile to the published data, age of the author, gender) is a public platform accessible to

¹<https://twitter.com/robots.txt>

everyone through the web² or mobile³. It was created to facilitate the generation of Algerian dialect resources (corpus, lexicon) but also to help researchers annotate their data. The annotation task was presented to the platform users as a game, where they gain 1 point for each text they annotate and 2 points when they give the same annotation as other users.

Figure 5.2: TWIFIL platform.

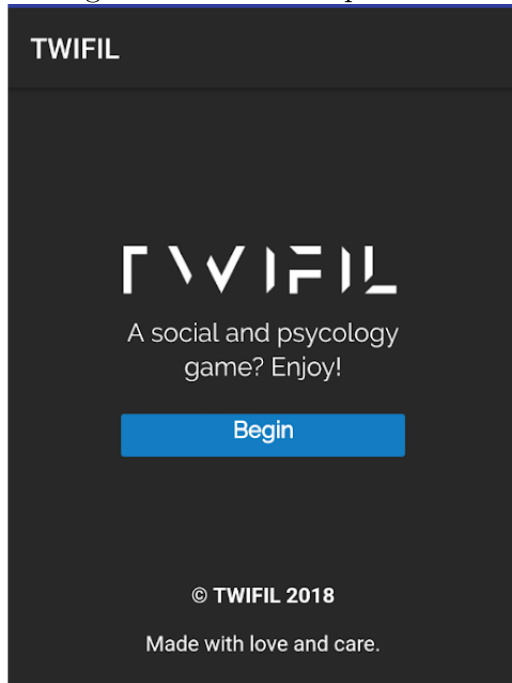


Figure 5.3: Login to TWIFIL.

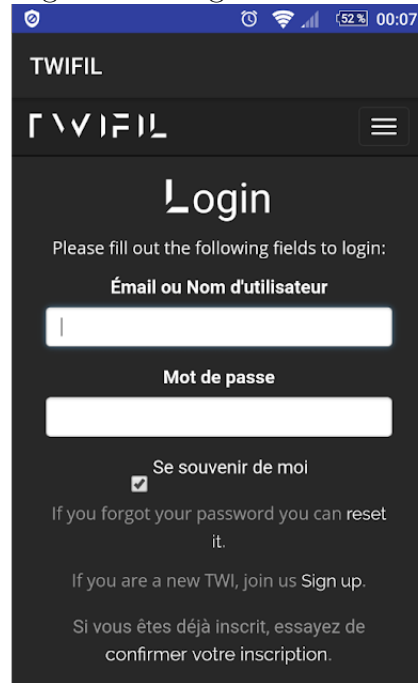


Figure 5.4: Choose a category.

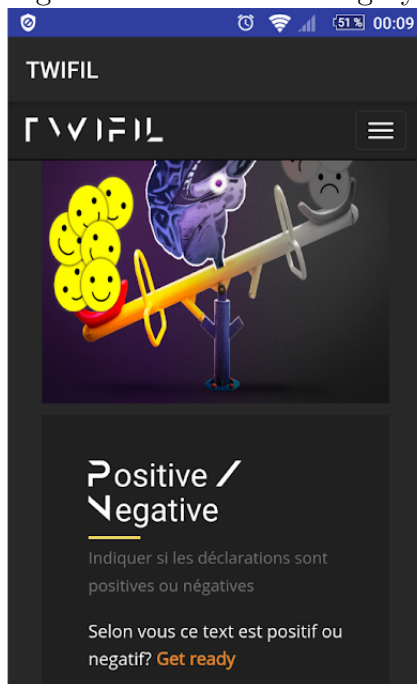


Figure 5.5: Annotate: choose a label.



The annotators were given guidelines on how to annotate each text. Along with, description of each category (polarity, emotion, gender) as well as examples of already annotated

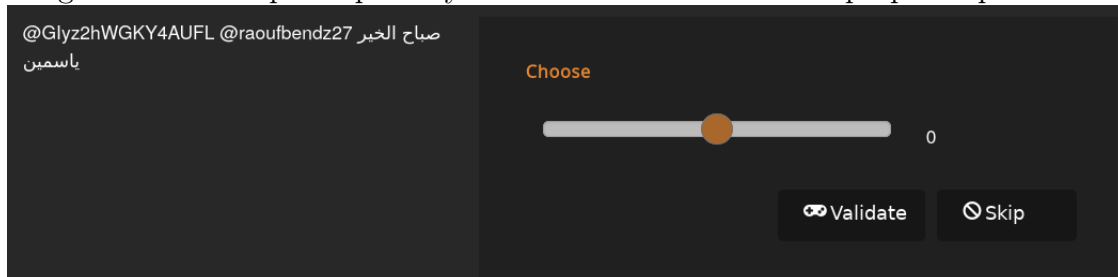
²<https://twifil.com>

³<http://bit.do/twifil>

texts of the same category. In order to manually check the validity of the annotations, administrators or corpus holders can access annotations and accept or ignore each one based on its consistency with regards to OEA and annotation guidelines, this was implemented to help recheck the annotations relevance. The annotation guidelines for each category are as follows:

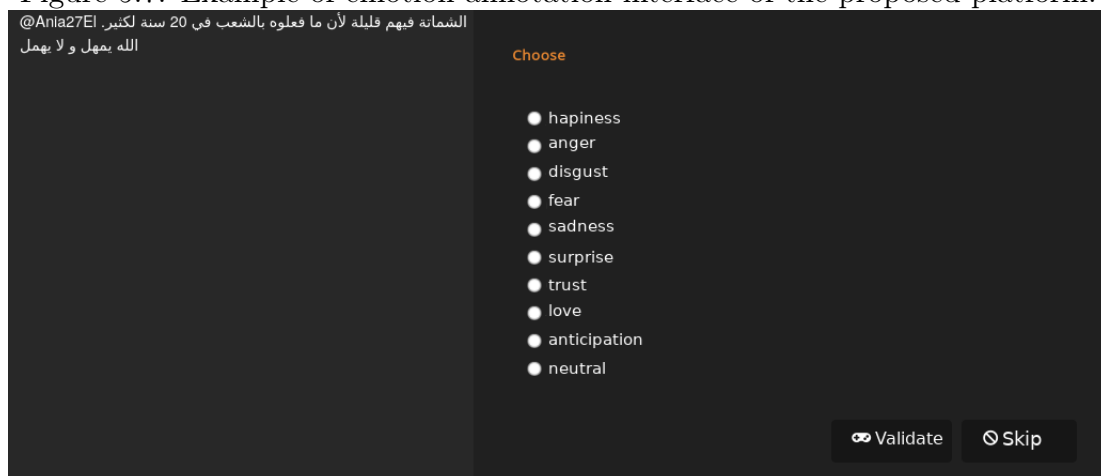
- the sentiment polarity of the shared text labelled between [-10; +10];

Figure 5.6: Example of polarity annotation interface of the proposed platform.



- the opinion class (positive, negative, neutral);
- the emotion felt by the reader of the text labelled as (joy, anger, disgust, fear, sadness, surprise, trust, anticipation, love and neutral); we followed the Plutchik's eight emotion set [Plutchik, 1984] to which we added love and the neutral class to account for factual tweets;

Figure 5.7: Example of emotion annotation interface of the proposed platform.



- the topic of the text (politics, sports, diverse, and others);
- the age of the author, labeled using age classes ([12-20], [21-30], [31-40], [41-50], [51-60], [61 and older]);
- the gender of the author (male, female, other).

The dialect lexicon is practically the same without age or gender. Annotators provide their impressions regarding the polarity and the emotion of a word or an idiom from the ALGD. New words can be added to the dictionary (which must be validated by an administrator), as well as different spellings of the words and different related words. Annotators can also add idioms and known Algerian expressions with their descriptions to facilitate the comprehension and the use of the idiom. The platform also allows users to upload their own corpora to be annotated, and generates a personalised link for each corpus.

5.4 The Generated Annotated Corpus: Twifil

In this section, the process of generating a new Arabic Algerian dataset for opinion and emotion analysis will be described. The new dataset was named "Twifil", using the platform's name. As mentioned in the previous section, the data was collected from the Twitter platform. The data displayed on the TWIFIL platform are tweets collected through the Twitter API using both standard and stream. TWIFIL has more than 500k collected tweets between 2015 and 2019 obtained from different random geo-locations in Algeria, using geo-tagging and keywords. The set of keywords contained names of known figures from politics to arts and sports, as well as the name of some places and local events, etc. With the help of 26 annotators, it was possible to generate a corpus and a lexicon, which were validated by the administrators of the platform. Considering tweets that were at least annotated by three different annotators, the labels of the corpus were assigned according to a majority vote, where a label had been used more than once otherwise the tweet would not be selected to be part of the corpus. Table 5.1 shows samples of these tweets along with their English translation, sentiment polarity and class as well as emotion class, user age and gender.

Table 5.1: An excerpt of the generated corpus via TWIFIL.

Post	Polarity	Polarity class	Emotion	Age	Gender
<i>Wa3lash tdirolna hakda khlonna trankil</i> (Why are you doing this, leave us alone)	-7	Negative	Anger	26	Male
<i>Ch7al rahi lsa3a</i> (what time is it)	0	Neutral	Neutral	30	male
<i>Piii khtito kounti hayla</i> (sister you were awesome)	5	Positive	Joy	28	female

Proposing the TWIFIL platform aims to bridge the gap by developing the first Algerian dataset of social media messages annotated at different levels. This dataset is the largest Algerian dataset for multiple opinion and emotion analysis tasks with multi-labels and multi-regional Algerian dialects as well as extra-linguistic information (age, gender, topic). It is also the first Algerian dataset annotated with multi-grained annotations (word and document level).

Data statistics

As mentioned, a corpus was built of more than 9,000 annotated and validated tweets for sentiments, Table 5.2 shows the number of tweets annotated for each polarity class. Where, the corpus has 4,350 positive, 2,615 negative and 2,191 neutral tweets. While Table 5.3 shows the details for the tweets annotated for emotion analysis (more than 5,000). Regarding age and topic, about 300 annotated tweets were collected and for gender, we have more than 700 (413 male, 255 female and 36 others) annotated tweets.

Table 5.2: Sentiment characteristics of the corpus.

Positive tweets	Neutral tweets	Negative tweets	Total
4,350	2,615	2,191	9,156

Table 5.3: Emotion characteristics of the corpus.

Joy	Anger	Disgust	Fear	Sadness	Surprise	Trust	Love	Anticipation	Neutral	Total
1,170	298	227	60	366	175	282	239	12	2,224	5,054

5.5 The Generated Annotated Lexicon: Twilex

Our approach constructs a lexicon containing words in both Arabic and Latin letters with their polarity/emotion/different spellings. The aim is to provide more NLP tools for the ALGD dialect, by enriching existing lexicons, however, the publicly available lexicons are the lexicons published only by [Mataoui et al., 2016]. Thus, to construct Twilex it was necessary to exploit the sentiment lexicon of Mataoui which contains a total of 5,027 words, without considering their polarity, since a different scale was used. It was enriched by the TWIFIL users and now counts about 21,000 terms and expressions of the ALGD (6,350 positives, 5,160 negatives and 10,090 neutrals), examples can be found in Table 5.4. The words are annotated in context hence, the same word may receive different annotations in different contexts. Therefore, the same approach was followed as used during the generation of the Twifil corpus, where the labels of the words were chosen following the dominant vote.

Table 5.4: An excerpt of the generated lexicon via TWIFIL.

Word/expression and different spellings	Polarity	Polarity class	Emotion
Hayla == great/هَآيَلَا, هَآيَلَى, هَآيَلَة	5	Positive	joy
T3ayi == boring/تَعِي, تَعِي, t3ay	-5	Negative	disgust
Ch7al == how much/شَحَال	0	Neutral	Neutral

5.6 Conclusion

The Arabic language is characterised by a wide number of varieties in dialects. With the emergence of the social web, it enables users to express their opinions using these dialects.

North African dialects, including the Algerian dialect, differ from MSA on all levels of linguistic representation, from phonology and morphology to lexicon and syntax.

Opinion and emotion analysis of the ALGD is challenging due to the rich and complex morphology of the language. Extracting and analysing the enormous volume of comments and reviews presented on the social web requires consideration to be given towards the peculiarities of the Algerian dialect and its characteristics (arabizi, code-switching). However, publicly available resources for OEA of the DALG are scarce.

In this chapter, an open platform for public annotations was presented called "TWIFIL". It helped create a quite large annotated corpus as well as a dialectal lexicon. These tools can be exploited for different opinion and emotion analysis tasks at a relatively low cost, and are now available to the community. It will provide a useful benchmark for those developing opinion and emotion analysis tools for the Algerian dialect.

Chapter 6

Multi-Level Embeddings for Processing Arabic Social Media Contents

Embeddings are very popular representations that allow computing semantic and syntactic similarities between linguistic units from text co-occurrence matrix. Units can vary from character n-grams to words, including more coarse-grained units such as sentences and documents. Recently, multi-level embeddings combining representations from different units have been proposed as an alternative to single-level embeddings to account for the internal structure of words (i.e., morphology) and help systems to generalise well over out of vocabulary words. These representations, either pre-trained or learned, have shown to be quite effective outperforming word-level baselines in several NLP tasks such as machine translation, part of speech tagging and named entity recognition.

The aim here is to contribute to this line of research, by proposing for the first time in Arabic NLP an in-depth study of the impact of several subwords configurations ranging from character to character n-grams (including word) for social media text classification. In order to address the lack of semantics that classical machine learning and lexicon-based methods suffer from, proposed below are several neural architectures to learn character, subword and word embeddings, as well as a combination of these three levels, exploring different composition functions to obtain the final representation of a given text.

Keywords Embedding, dense vectors, multi-level vectorial representations, contextual embedding, Convolutional Neural Networks, Arabic dialects, sentiment analysis, emotion detection, irony detection.

6.1 Introduction

Embeddings can operate at multiple linguistic levels ranging from subword [Bojanowski et al., 2017], word [Pennington et al., 2014, Mikolov et al., 2013c], sentence [Conneau et al., 2017, Cer et al., 2018] to document [Le and Mikolov, 2014], including concept [Yamada et al., 2018, Camacho-Collados et al., 2016]. Representations can either be monolingual or multilingual where units from different languages are mapped to the same area in the latent space [Ruder et al., 2019]. Like others in this field, this research was also attracted by the strength of embeddings and their ability in encoding the different features of a word, including semantics, into a single hot-encoded vector, along with the fact that character and sub-word embedding can overcome the problem of OOV in word embeddings, thus creating embedding vectors for unseen words.

In this work, we focus on vector embeddings at different granularity levels targeting both MSA and its different dialects. As shown in the previous chapter, the literature in learning different representations is vast. Several models have been proposed targeting different lan-

guages and different NLP tasks, offering embeddings at different levels using both standard models and deep learning models. However, few models were specifically designed to handle the Arabic language and its dialects, and as far as we know, existing models have been evaluated on a single task and without any systematic comparison of their performances across different dialects. This work aims to bridge the gap by studying for the first time:

- The impact of multi-level embeddings on supervised classification of Arabic social media texts written both in MSA and Arabic dialect (Algerian, Tunisian and Levantine). The study proposed here continues the same line of research previously initiated in [Vania and Lopez, 2017] and [Zhu et al., 2019] who proposed a systematic comparison of subword models while varying the basic unit of representation and their compositions. However, this work goes beyond MSA and a single task evaluation [Vania and Lopez, 2017] and considers orthography segmentation unit instead of morpheme [Zhu et al., 2019]. More importantly, it proposes to identify the efficient representations which are the most stable across tasks and dialects.
- Different segmentation units (character, character n-grams, word) by taking a sequence of n-grams of characters where $n \in [3, L]$ and L being the word length¹. Therefore this research goes beyond [Vania and Lopez, 2017, Labeau and Allauzen, 2017] and [Dahou et al., 2019, Erdmann et al., 2018] who respectively investigated with $n = 3$ and $n \in [2, 6]$.
- Several ways to compose unit embeddings, ranging from concatenation to learning the embeddings using different neural architectures. Evaluate here are the proposed models on sentiment analysis, emotion detection and for the first time irony detection. For each task, experiments using well-known datasets, including ones that have been used in shared tasks, are undertaken and these models are then compared with the best-performing state of the art models.
- Impact of dialect-specific embeddings on text classification. Compared to MSA embeddings, few dialect embeddings exist and are freely available [Soliman et al., 2017, Al-towayan and Tao, 2016]. Most existing works do not target specific dialects [Vania and Lopez, 2017, Pinter et al., 2017, Zhao et al., 2018], making them agnostic to inter-dialect variations [Qwaider et al., 2018] (cf. examples in the Introduction 1.2). A notable exception is the multidialectal word embeddings by [Erdmann et al., 2018] who modelled dialects independently and then mapped them into one embedding space using supervised and unsupervised techniques. In this work, a different strategy is adopted: Learn embeddings at different granularity levels from a single dialect dataset and then use it to classify texts in other dialects. For the first time, the Algerian dialect which lacks available vectorial representations, as far as it is known is targeted². ALGD is a North African dialect known for its diversity due to cultural mix and the heritage bequeathed by all civilisations who have passed through Algeria. It is, therefore, less normalised compared to MSA and other Arabic dialects as its vocabulary contain terms from several other languages among French, Spanish, Italian, but also Tamazight the original language of the first inhabitants of the region [Harrat et al., 2014, Mataoui et al., 2016, Harrat et al., 2017]. These new pre-trained models will be made available to the community.

¹Tests were carried out with $n \in [i, L]$, $1 \leq i \leq 5$ and the best performances were obtained with $i = 3$.

²[Abidi and Smaïli, 2018] used Word2Vec on an Algerian dialect (ALGD) corpus for learning a dialect lexicon, however, these embeddings were not made publicly available.

There are basically three methods to better represent the text input, and all three approaches were used in this work in order to determine the best representation for text classification.

1. *Use pre-trained models and fine-tune them*: such as AraVec [Soliman et al., 2017] or BPEmb [Heinzerling and Strube, 2018]. Pre-trained models were fine-tuned for each task’s datasets as part of a CNN learning process. The used models are as follows:
 - + *Word-level embeddings*: AraVec was used [Soliman et al., 2017]. The four models built over tweets were tested (Twitter-CBOW-100, Twitter-CBOW-300, Twitter-Skipgram-300, Twitter-Skipgram-100). Only best results which were obtained using Twitter-CBOW-100 are reported.
 - + *Character n -grams embeddings*: we used BPEmb [Heinzerling and Strube, 2018].
 - + *Multi-level embeddings*: we used BERT multilingual model and Flair character-based models (both backward and forward Arabic embeddings³), which are stacked and exploited through an RNN.
2. *Build own embeddings relying on standard techniques*: such as Word2Vec, glove and FastText from a large corpus(cf. section 6.3). The dataset of 2,236,351 tweets was used containing 44,577,172 words. The used techniques are as follows: own word embeddings with Word2Vec (SG and CBOW), own word embeddings with FastText (referred to as ARFT). The best results have been obtained with Word2Vec CBOW of 300 embedding size, referred to as ALG2VEC_w.
3. Construct the embeddings while fitting a neural network. This approach relies mainly on either randomly generated weights (following [Elaraby and Zahran, 2019] who used the task’s dataset to create the embeddings) or weights as given by pre-trained models. It has been shown that relying on pre-trained weights outperform models with randomly generated weights [Wang and Xu, 2017]. However, using the task’s dataset helps avoid the common limitation of the previous approach which usually require large datasets for the trained models to perform well. Following [Kim, 2014], we use a CNN architecture for text classification, targeting both binary (sentiment analysis and irony detection) and multi-class (emotion detection) tasks⁴. Our model takes as input vectorial representation of different units (character, subwords and words) composed using different composition functions and consists of a single 1D Convolutional layers using 50 different filters of different sizes (2, 3 and 7) and a stride of 1, in order to capture different scales of correlation between input units, with a ReLU activation function. We further downsample the output of these layers by a 1D max-pooling layer to obtain a fixed-dimensional representation of the input. The output layer is composed of a fully connected layer, followed by a softmax classifier.

We studied all three, and we applied the last method on different vectorial representations going from character to words including subwords embeddings. We also experimented with multi-level embeddings (character+word embeddings ($ch + w$), character+subword embeddings ($ch + sw$), subword+word embeddings ($sw + w$), character+subword+word embeddings ($ch + sw + w$)). This approach help produce two different forms of embeddings, one using the task-specific dataset as part of the classifier learning process (referred to as RAND_{LEVEL})

³https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md

⁴We also experimented with LSTM, however the results were lower.

and another using our large corpus either grouped (referred to as $\text{MGCNN}_{\text{LEVEL}}$) or concatenated (referred to as $\text{COCNN}_{\text{LEVEL}}$). In order to construct these different models, we used a CNN-based architecture⁵ that is schematised in Figure 6.1. The experiments undertaken include the following models:

- Constructing word embedding using the two methods, the task corpora (RAND_w) and our large corpus (EMBCNN_w) lookup table.
- Constructing character embedding using both, the task corpora (RAND_{ch}) and our large corpus (EMBCNN_{ch}) lookup table.
- Constructing subword embedding using both methods RAND_{sw} and EMBCNN_{sw} .
- Following the same method multi-level embeddings ($ch+w$, $ch+sw$, $sw+w$, $ch+sw+w$) were constructed.
- For all the above, the aim was to experiment with task-specific features to enrich the embeddings with such features.

As the aim is to identify what are the best input representations to improve classification performances, experiments were carried out with three main sources of embeddings: (1) existing pre-trained embeddings, (2) dialect-specific embeddings build from scratch from a large unlabelled dataset, and (3) embeddings learned as part of a neural network. In the next sections, each of these three approaches is detailed. In the following, we first detail each of these three approaches then provide a summary of the different embeddings proposed in this study along with their abbreviations (cf. Section 6.4.6).

The remainder of this chapter details the approach adopted for generating different embedding models at different levels, starting by exploiting existing pre-trained models in section 6.2, dialect-specific embeddings in section 6.3, and embeddings learned as part of a neural network 6.4, while section 6.5 concludes this chapter.

6.2 Leveraging Pre-trained Embeddings at Different Unit Levels

This research relies on pre-trained embeddings and fine-tune them while training the models on the task datasets. These are strong baselines. We recall that these embeddings were either pre-trained on MSA texts (like Wikipedia and blog posts) or social media (e.g. tweets) without targeting a specific dialect:

- ARAVEC. It uses the AraVec word-level embeddings [Soliman et al., 2017]. The four available models trained on tweets with different dimensions were tested (Twitter-CBOW-100, Twitter-CBOW-300, Twitter-Skipgram-300, Twitter-Skipgram-100) and the best results were obtained using Twitter-CBOW-100.
- MAZAJAK⁶. It uses the Mazajak word embeddings [Abu Farha and Magdy, 2019] as input to the CNN architecture. Mazajak provides four pre-trained language models trained on two dialect-agnostic tweet datasets of size 250M and 100M following either continuous bag of words (CBOW) or Skip-Gram (SG) learning procedures. In our

⁵Regarding the construction of embeddings using a CNN architecture we tested the use of Arw2v300 and a random lookup table, where the latter outperformed the Arw2v300. We relied on a CNN architecture, which compared to LSTM, gave the best results.

⁶<http://mazajak.inf.ed.ac.uk:8000/#embedding-page>

experiments, the best results were obtained with CBOW_100. Note that both MAZAJAK and ARAVEC are word-level embeddings, the main difference being the training dataset, which is bigger for MAZAJAK.

- **FASTEXT.** It uses the FastText pre-trained Arabic model (cc.ar.300.bin) of embedding size equal to 300, character n-grams between 3 and 6 and a maxlen of 200. The CNN used was trained for 10 epochs with a learning rate of 0.3.
- **BPEMB.** It uses the BPEmb character n-grams embeddings. Firstly, the input words are split into n-grams using the BPE segmentation which depends on the vocabulary size (a smaller size yields a segmentation into many subwords, while a large size results in frequent words not being split [Heinzerling and Strube, 2018]). BPEmb provides eight models for the Arabic language of different embeddings dimensions. The experiments conducted show that the best results were obtained with a dimension of 100.
- **FLAIR.** It is based on Flair subword embeddings that proposed two models for Arabic: forward (*ar-forward*) and backward (*ar-backward*). We experiment with each model separately but also by stacking the two models as part of the Flair RNN-based TextClassifier⁷. The best results were obtained by the stacked architecture with 512 hidden states and after setting the output dimension of reprojecting token embeddings to 256⁸. The model was trained over 8 epochs with a learning rate of 0.1 and a batch size of 32.
- **BERT.** It uses BERT pre-trained multilingual language model on top of which was added an untrained layer of neurons. It was trained for 2 epochs with a learning rate of $2e - 5$ and a maxlen of 200⁹ ¹⁰.

6.3 Dialect Specific Language Models

As most state of the art language models for the Arabic language are dialect agnostic, we propose to build a dialect specific embeddings, focusing on the Algerian dialect. To this end, we built **AlgSet**, a new dataset composed of tweets using a set of keywords about popular Algerian figures, locations and some important events that occurred in Algeria. Finally, after removing duplicates and tweets containing less than three words, we collected 2,236,351 tweets containing 44,577,172 words. Table 6.1 provides some statistics for the dataset, including the vocabulary size. Among the collected tweets, 6.2% were collected using only geotagging, the remaining 93.8% were selected with both keywords and geolocalisation.

Table 6.1: **AlgSet**: The Algerian dialect dataset for building word embeddings.

Period	# words	vocabulary size	# tweets	% geo-tagging	% keywords + geotag
01/01/2015 - 12/04/2019	44,577,172	4,858,988	2,236,351	6.2	93.8

It is important to note that the keywords used for collecting tweets may bring some bias in the data. To reduce it, we relied on a set of keywords about various hot topics

⁷<http://bit.do/flairTextClassifier>

⁸This parameter indicates whether the token embeddings is reprojected in a separate linear layer before injecting them into the RNN.

⁹Maxlen specifies the maximum number of words in a text, where longer texts are truncated to this length.

¹⁰hUL-MonA, the Arabic universal language model was not experimented with as their results were comparable to BERT.

discussed in the Algerian media as well as some other generic topics (e.g., health). Our intuition behind choosing such media-friendly topics is to capture vocabulary diversity while reducing the number of tweets that relate to personal content. This data collection procedure has already been used in many studies and shared tasks, see for example [Ghanem et al., 2019, Benamara et al., 2017a]. Another important issue concerns the highly dialectal nature of tweets contents (i.e., they can also include MSA). Of course, there is no guarantee to collect only Algerian dialect but we believe that the amount of MSA tweets is relatively small as already pointed out by several studies showing that Algerian dialect is the main language used in Algerian social media (see for example [Meftouh et al., 2012, Abidi and Smaïli, 2017, Abidi and Smaïli, 2018]).

The dataset has been pre-processed using standard preprocessing techniques: removing URLs, user mentions, punctuation marks, as well as long and repeated characters which denote exaggerations (e.g., خوياءااا, which will be transformed to "خويا" removing the extra Alif). We also removed all diacritics and normalised Alif, Ya and Hamza. Finally, as stopwords removal increased the results, we relied on a list of standard MSA stopwords (like the articles "و" (*wa/and*), and the pronoun "أنتم" (*Aantm/you*)) and extended it by translating lexical entries into Algerian dialect (e.g., "تؤمأ" (*ntwmA/you*))¹¹.

The dialect-specific language models are as follows:

- **ALG2VEC_w**. We experimented with different models using two Word2Vec architectures, namely continuous bag of words (CBOW) and Skip-Gram (SG). For each architecture, the models' parameters were varied: vector size ($d \in [50, 500]$), context window of w tokens ($w \in [1, 5]$) where w denotes the maximum distance between the current and predicted word within a sentence, and finally the minimum word count $m \in [1, 8]$ which fixes a threshold to prune infrequent words. The best model was obtained with CBOW ($d = 300$, $w = 5$, and $m = 1$).
- **ALGFT**. Following the same workflow, we experimented with FastText using CBOW and SG architectures. We recall that a word's embedding in FastText is obtained by calculating the sum of all its components character n-grams embeddings. The SG implementation outperforms CBOW and best results were given by $d = 300$, $w = 5$, character n-grams between 3 and 6, and word n-gram max length equal to 2. ALGFT has been trained for 10 epochs with a learning rate of 0.3.
- **ALG_{ENS}**. It refers to ensemble embeddings which consist of combining different embedding vectors into a single word vector using different algebraic operators (concatenation, sum and mean). Embedding composition has shown to yield good results in different NLP tasks. For example, [Fares et al., 2019] combined both TF-IDF character vectors and FastText embeddings while [Speer and Chin, 2016] GloVe and Word2Vec with external knowledge. In this work, we experimented with different ways to combine ALG2VEC_w CBOW with ALGFT as well as ALG2VEC_w CBOW with SG models. The latter achieved the best performances using the mean as a composition function.
- **ALG2VEC_{sw}**. Finally, we experimented with the creation of subword embeddings using Word2Vec CBOW. This model has been inspired by [Shi et al., 2015] who generated word embedding from Chinese radicals, achieving good results in various tasks (text categorisation, word segmentation and web search ranking). ALG2VEC_{sw} is similar to ALG2VEC_w but receives as input subwords (of length in $[3, L]$, L being the word's length) instead of words and generates embeddings of a size equal to 100.

¹¹The complete list of stopwords used in this study can be found here: https://github.com/kinmokusu/oea_algd/blob/master/Important%20documents/stopwords.txt.

6.4 CNN Architectures for Learning Multi-Level Embeddings

6.4.1 Methodology

While building embeddings using standard models (Word2Vec and FastText) has shown to be quite effective at capturing semantic similarity, learning embeddings as part of a CNN network, on the other hand, is more focused on morphology and hence can learn representations for OOV words [Li et al., 2018]. Let $U \in \{ch, sw, w\}$ be the segmentation of a given input text t into a unit u which can be either a character (ch), subword (sw) or word (w). We propose the following architectures to learn the embeddings:

- *Single unit embeddings* at U level, relying either on:
 - $RAND_U$ where unit vectors are randomly generated from task-specific labelled dataset following [Mulki et al., 2019, Elaraby and Zahran, 2019, Kalchbrenner et al., 2014]. The initialised weights are further fine-tuned as part of the classifier learning process. Random values were uniformly fixed from an interval centered at zero $[-0.05, 0.05]$, following [Kocmi and Bojar, 2017] who empirically showed that small variance (i.e., < 0.1) enables the network to learn relations between words in a way similar to Word2Vec learning. The embedding size was set to 300 for words, 100 for subwords and 15 for characters. This model uses a straightforward implementation of a CNN architecture, a single convolution layer with 50 filters and ReLU activation function (batch size 100, 8 epochs and Adam as optimiser with 20% dropout) followed by a dense layer with a softmax activation function. The models try to minimise the mean squared error between the predicted class and the gold class.
 - $EMBCNN_U$ which uses a CNN architecture following [Kim et al., 2016] taking the **AlgSet** dataset (cf. Section 6.3) as input without requiring neither pre-trained embeddings nor labelled data.
- *Multi-level embeddings*. Let $level \subseteq U$ be the subset of units used to learn multi-level embeddings. For example if $level = \{ch, sw\}$ means that both characters and subwords which encode different aspects of a language are used to generate the text representation. Given $level$, we extend $RAND_U$ and $EMBCNN_U$ to take as input a combination of several units $u \in level$. The models are:
 - $RAND_{LEVEL}$: it randomly generates weights for each unit u in a given $level$ ¹².
 - $COCNN_{LEVEL}$: this model uses concatenation to merge all vectors $EMBCNN_U$ which will result in a single set of vocabulary containing linguistic features from different levels.
 - $MGCNN_{LEVEL}$: this is an extension of [Zhang et al., 2016] multi-group CNN where unit embeddings as given by $EMBCNN_U$ are treated independently then combined before the classification layer¹³. As far as we know, no one used multi-group CNN for morphologically rich languages like Arabic.
- *Enriched embeddings*. It enriches embeddings obtained by one of the five models described before by additional task-oriented linguistic features, as already done by [Lal

¹²The $RAND_{LEVEL}$ models use the same parameters as $RAND_U$ except for the embedding sizes which are equal to 150 for all the models and 100 for the $Rand_{ch+sw}$ model.

¹³We also experimented with ALG2VEC instead of $EMBCNN_w$ however, the results were lower.

et al., 2019] when predicting sentiment orientation of English-Hindi code-mixed data. These models are referred to by adding $_{+feat}$. For example, $MGCNN_{LEVEL+FEAT}$ means that multi-level embeddings have been generated using the multi-group CNN model, then coupled with extra features for text classification.

For both $RAND_{LEVEL}$ and $CoCNN_{LEVEL}$, either standard concatenation or different unit permutations were considered while varying their positions within the input text, using $+$ and $\oplus$ operators respectively, as illustrated in Table 6.2 taking "*nice food*" as an example. The table shows how the text is decomposed into characters and words and then how each unit is concatenated following distinct positions: all words then all characters, characters of a given word, then the word itself, etc. The same procedure is followed for a three-level decomposition into characters, subwords and words.

Table 6.2: Different unit permutations when learning multi-level embeddings.

Model	Unit Permutations	Examples
$CoCNN_{CH+W}$	all words + all chars	nice + food + n + i + c + e + f + o + o + d
$CoCNN_{CH\oplus W}$	chars w1 + w1 + chars w2 + w2	n + i + c + e + nice + f + o + o + d + food
$CoCNN_{W\oplus CH\oplus SW}$	w1 + chars w1 + subwords w1 + w2 + chars w2 + subwords w2	nice + n + i + c + e + nic + ice + food + f + o + o + d + foo + ood

The $RAND$ models being straightforward, below are detailed the CNN-based architectures.

Each of these three architectures produces a dense representation of each unit which is then passed to a second CNN to perform the text classification tasks. Before detailing the experiments, a description of these three architectures is provided below, starting by a general description of the CNN.

6.4.2 General Architecture of the CNN

In the classifier, the input at time t is an output from a single-layer convolutional neural network with max-over-time pooling and different filters with different widths.

Input and output word embeddings are usually parameters, stored in lookup matrices $\mathbf{Q}$. The word embedding $\mathbf{v}$ of a word w is simply the column of $\mathbf{Q}$ corresponding to its index in the vocabulary V .

Since several word-level representations were tested, the data (word, subword (character 3-n-grams), character) used is referred to as a unit. For notation, vectors are denoted with bold lower-case (e.g. $\mathbf{x}$, $\mathbf{b}$), matrices with bold upper-case (e.g. $\mathbf{W}$, $\mathbf{U}$), scalars with italic lower-case (e.g. x , b), and sets with italic upper-case (e.g. V , C) letters. For notational convenience, it is assumed that words and characters have already been converted into indices.

Let $|V|$ be the fixed size vocabulary of units, d be the dimensionality of the element embeddings, and $\mathbf{Q} \in \mathbb{R}^{d \times |V|}$ be the matrix of the units embeddings. Suppose that a unit $k \in V$. Then the representation of unit k is derived from the matrix $\mathbf{Q}$, where the j -th column corresponds to the unit embedding (i.e. the j -th column of $\mathbf{Q}$). A convolution is applied between $\mathbf{Q}$ and a filter $\mathbf{F} \in \mathbb{R}^{dm}$ of width m , after which a bias is added and then a non-linearity is applied to obtain a feature map $f^k \in \mathbb{R}^{n-m+1}$. Specifically, the i -th unit of f^k is given by: $f^k[i] = ReLU(< Q^k[*, i : i + n - 1], F >)$. where $Q^k[*, i : i + m - 1]$ is the i -to- $(i+m-1)$ -th

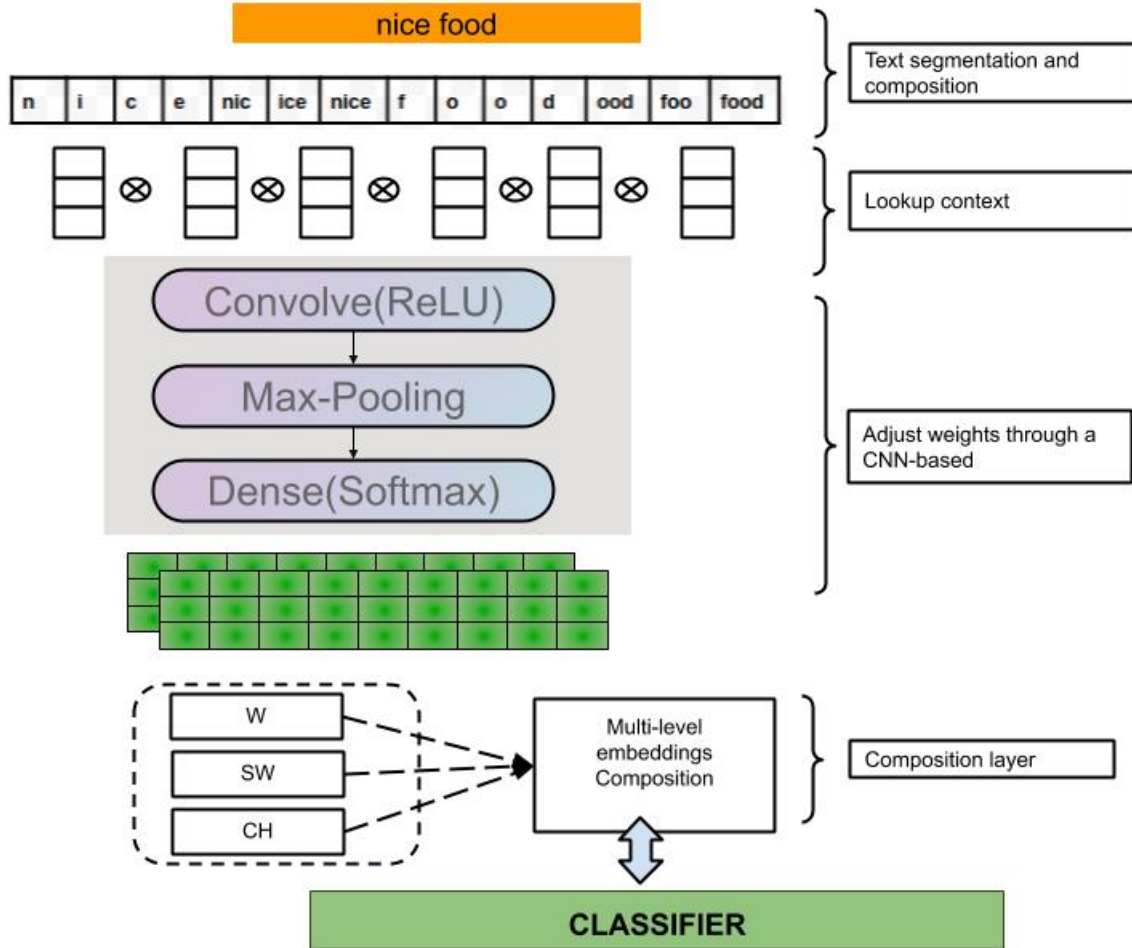
column of Q and $\langle A, B \rangle = P(ABT)$ is the inner product. Finally, the max-over-time $y^k = \max f^k[i]$ is taken as the feature corresponding to the filter $\mathbf{F}$ (when applied to unit k).

The model simply replaces the input embeddings $\mathbf{X}$ with the matrix $\mathbf{Q}$ generated from the unit-level convolutional neural network. If we denote $s_1:T = [u_1, \dots, u_T]$ to be the sequence of units in the training corpus, training involves minimising the negative log-likelihood of the sequence by minimising the loss of the predictions which is typically done by truncated back-propagation through time.

Each of the additional surface features were incorporated into the model via the dense layer for the surface features information to be incorporated to fine-tune the weights and enhance the labelling.

During experiments, each parameter of the model was empirically set. The number of filters was varied from 10 to 200, with a step equal to 10. Also varied were their sizes depending on the embedding size which were tested using a sliding step being 1 (the stride). The validation set size was set to 20%, batch size (200-2000), number of epochs (1-10) and different optimisers such as the Adam optimiser with a regulariser of 0.01. A 20% dropout was used to tighten the vectors to help avoid overfitting and improve error generalisation.

Figure 6.1: Double-CNN architecture for embedding generation and text classification



As previously mentioned we experiment with three main architectures (cf. Figure 6.1):

- Single unit embeddings* (cf. Section 6.4.3) considering either characters, subwords or words as input (hereafter EmbCNN_{ch} , EmbCNN_{sw} , and EmbCNN_w respectively).
- Multi-level embeddings* (cf. Section 6.4.4) which combines two or three units using a set of composition functions, namely $\text{MGCNN}_{\text{LEVEL}}$ and $\text{CoCNN}_{\text{LEVEL}}$.

- (c) *Enriched embeddings* where architectures presented in (a) and (b) are extended using task-specific features that are injected to the dense layer of the classifier during the learning process enriching, therefore, the embeddings with additional sentiment/domain-specific features (cf. Section 6.4.5).

The system created in this work could be regarded as a double CNN architecture, where the dense vectors are retrieved from the first CNN and the second CNN performs a classification task using these embeddings as data representation.

We experiment with single unit embeddings then we move to multi-level embeddings to model more complex interactions between the different units. The embeddings are generated by convolution over the different units within a word prediction context.

Several studies rely on supervised or unsupervised learning when training embeddings. For example, [Santos and Zadrozny, 2014] learns task-specific character-based word embeddings within an end-to-end supervised system. The word embeddings are built by convolution over characters within a part-of-speech tagger. The same approach was followed, which is learning embedding as part of a supervised learning process. Since the large corpus is not annotated, there were two options given. The first is learning the embedding as part of a classification evaluation which labels a given text using one of its surfaces features attributes, extending the morpho-syntactic tagging approach used by [Pinter et al., 2017]. The second is a CBOW-like process, which is predicting a word from its surrounding context words, and by extension predicting a word using its morphological components (character n-grams), enabling the prediction of embeddings for unseen words based on their internal structure. This problem was approached as a generation problem. Unlike prior work [Pinter et al., 2017], pre-trained embeddings were not relied upon. It was assumed that there was a generative wordform-based protocol for creating these embeddings. By training a model over the existing vocabulary, the model can later be used for predicting the embedding of an unseen word.

The second approach was relied upon, where the model takes as input a large corpus and generates a dense vector for each unit, where each unique unit in the used vocabulary is assigned a corresponding vector in the space, where units sharing common contexts in the corpus are positioned close to one another.

Following [Kim et al., 2016] and [Li et al., 2018], a CNN architecture was used to create embeddings. Analysis of the different representations obtained from the results of the second CNN reveals that the model can encode both semantic and orthographic information.

Units encoder (the first CNN) learns to generalise the vectors from a large dataset. The model predicts the following sequence of context units from a window of previous context units. It is faster and more effective on the classification tasks compared to dialect-specific models.

The architecture of the model, shown in Figure 6.1, is straight-forward. Whereas models take pre-trained word embeddings as inputs [Zhao et al., 2018, Dahou et al., 2019], this model creates instead the embeddings from scratch by performing an embedding lookup initialisation, which is fine-tuned through a single-layer convolutional neural network with max-over-time pooling. This architecture was inspired by the model used for language modelling by [Kim et al., 2016]. To assess the quality of the learned dense vectors, it was considered employing the different learned embeddings as continuous features in several NLP classification tasks using the second CNN, including sentiment classification, emotion classification and irony detection. Hereafter, more details regarding each layer are given.

Input Layer A placeholders for sentence-level input are generated at the input-level of the model, which consider both word and character n-grams units depending on the model

(single-unit, multi-units). It is depicted as a sequence of different units consisting of m words, where each word is decomposed into a sequence of n character n-grams, and the value of each unit will be given by its d -dimensional embedding vector. For the simplicity of notations and from the mini-batch training perspective, for a batch of $\mathbf{S}$ sentences, the word- and character-based inputs will be in the form of 3-D tensors, i.e., $U \in \mathbb{R}^{S|U|d}$. Which constitutes the input matrix I .

Embedding Layer The embedding layer performs an initial random lookup of embeddings (of different dimensions depending on desired representations of units and the composition levels see Table 6.2), since initialisation using pre-trained embeddings following [Wang and Xu, 2017], did not render the same results. Our experiments showed that a randomly generated lookup table is better than using pre-trained models for both precision and execution time. These embeddings are then stacked to form the matrix $\mathbf{Q}$ and fill the placeholders in with the corresponding word or character n-grams vectors. As illustrated in Figure 6.1, similarly to [Li et al., 2018], the vectors of units are first extracted from a lookup table. These vectors form the matrix Q of size $d * |V|$, where $|V|$ is the vocabulary size. The matrix $\mathbf{Q}$ is a parameter to be learned, and the size of the embedding d is a hyper-parameter to be chosen or set empirically.

To calculate a unit’s vector, the model considers its surrounding context units, which include the word and its character n-grams- learning while taking into consideration the two help make effective representations. This means, the entire-word and all its subwords are contributing to the network prediction step for every training example. Wherewith every batch $\mathbf{S}$ the wights of matrix Q are adjusted accordingly.

There are two sets of weights in the network, one is generated between the input and the hidden layer, and the second is generated between the embedding layer and the hidden layer. The latter is taken as the units representations. The embedding-hidden weighs matrix has a size of $[d * |V|]$, where d is arbitrary and a hyper-parameter, and it also represents the number of neurons in the hidden layer.

Convolutional Layer Convolution operations are applied between $\mathbf{Q}$ and multiple filter matrices and weights are updated by applying the **ReLU** activation function which stands for rectified linear unit $f(x) = x^+ = \max(0, x)$, where x is the input to a neuron. It has been shown that ReLU compared to other activation functions, allows faster and effective training of deep neural architectures [Hahnloser and Seung, 2001]. During the experiments it was observed that the ReLU function gave better results compared to other activation functions (sigmoid, softmax, tanh). The convolution filters assist in mining local feature where multiple filter sizes can be used.

Max-Pooling Layer A max-over-time pooling operation is applied to obtain a fixed-dimensional representation of each unit.

Max pooling is a sample-based discretisation process. The objective is to down-sample an input representation (hidden-layer output matrix), reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned. This is done in part to help avoid over-fitting by providing an abstracted form of the representation, as well as, decreasing the computational cost by reducing the number of parameters to learn and provides basic translation invariance to the internal representation. Max pooling is applied using a max filter to (usually) non-overlapping subregions of the initial representation [Kulkarni and Satapathy, 2020].

Dense Layer Finally, the affine transformation using a softmax is applied over the hidden representation through a fully connected layer of CNN to obtain the distribution over the next unit sequence. The fully connected layer captures the interactions between the features obtained by the filters. To approximate the conditional log-likelihood the model seeks to minimise the loss of the predictions. The resulting vectors are used for predicting contextual words using a softmax since it is known to work better for infrequent words, the same as word2vec [Google, 2013].

Results Vector embeddings for each unit, where each unit contributes to enriching the embeddings of other units, and each embedding vector is enriched with the surface features.

We anecdotally note that (1) having more convolutional layers before the max-pooling filter did not produce better results, and (2) multi-channel CNN resulted in lower performance and did not improve the classification results.

Selecting the vector of a specific object can be translated to a matrix product in the following way:

$$\begin{cases} 0 & \text{if, } i \neq Object_{index} \\ 1 & \text{if, } i = Object_{index} \end{cases} \quad (6.1)$$

$$\underbrace{Object_{vector}}_{dim \ 1} = \underbrace{\mathbf{Q}}_{dim \ units} \cdot \underbrace{\vec{v}}_{units} \quad (6.2)$$

Where v is the one-hot vector representing a text, which determines which unit needs to be transformed. And $\mathbf{Q}$ is the embedding matrix.

To assess the quality of the learned dense vectors, we considered extrinsic evaluations by employing the different learned embeddings as continuous features in several NLP classification tasks using the second CNN. The second CNN uses the same architecture. It takes the embeddings generated by the first CNN and outputs the class at dense layer level using a softmax function, where cross-entropy loss between the predicted class of the tweet and the actual class is minimised.

6.4.3 EMBCNN_U: Learning Single Unit Embeddings

Let t be an input text composed of a sequence of a given unit u , i.e., $t = \{u_1, u_2, \dots, u_n\}$ where n is the number of units. Segmentation of t into units follows white space for words, and bytes for characters including special characters, such as white-space and punctuation. Subword segmentation is obtained by taking a sequence of n -grams of characters where $n \in [3, L]$ and L being the word length. To learn the representation of a unit, single-layer CNN architecture is relied on, which is composed of one convolution layer and a pooling layer following [Kim et al., 2016]. Given an input sequence of units, firstly each unit vector is extracted from a lookup table which results in a vector v_i of size d then concatenate all unit vectors to compose the matrix lookup context of the sequence $\mathbf{Q} \in \mathbb{R}^{d|S_u|}$, S_u being the unit vocabulary size. The input of the network is thus as follows:

$$\mathbf{Q} = v_1 + v_2 + \dots + v_n$$

$\mathbf{Q}$ is then injected into a 1D-Conv layer and optimised jointly with the model parameters by applying convolution operations that consider windows of consecutive units within the sequence, where a set of filters (to be learned) are applied to these unit windows to generate corresponding feature maps. Finally, a max-pooling operation is used on top of the feature maps to abstract the most salient semantic features, resulting in the final representation. The resulting vector is used for predicting a unit from its surrounding context units as in

CBOW¹⁴, and by extension predicting a word using its morphological components (character n-grams), enabling the prediction of unseen words based on their internal structure. During training, the model minimises the mean squared loss between the predicted distribution of the next sequence and the gold sequence. EMBCNN_u best configurations are shown in Table 6.3 when input vectors are randomly initialised.

Table 6.3: EMBCNN_u experimental settings.

Parameters	Best settings
Embedding size for each unit	ch: 15, sw: 100, w: 300
Number of filters	50
Filter width	2-7
Convolutional layer activation	relu
Dense layer activation	softmax
Optimiser	Adam

To measure the effect of randomly initialised input vectors on embeddings generation, character and subword representations were experimented with, which were derived from ALG2VEC_w following [Zhao et al., 2018]¹⁵ and propose two additional models: EMBCNN_{ALG2VEC_{ch}} and EMBCNN_{ALG2VEC_{sw}}.

6.4.4 Constructing Multi-Level Embeddings

The research into multi-level embeddings was motivated by the fact that distributed word embeddings treat words as atomic units and ignore any surface or morphological similarities between different words, and uses a generic out-of-vocabulary representation, which is shared between all other unseen words [Rei et al., 2016]. The main objective is to define the best representation for texts’ classification and evaluate the influence of subword embedding on NLP tasks since [Watson et al., 2018, Lal et al., 2019, Dahou et al., 2019] showcased that these models yield better results compared to word-level models. Moreover, the fact that integrating character n-grams information takes into account the language morphology [Bojanowski et al., 2017], which would enhance the learning curve of the model and by association the learning of meaningful representation for rare words. With the aim of addressing these challenges, different models were considered as follows.

– **Multi-Group CNN (MGCNN_{LEVEL})**. This model was first proposed by [Zhang et al., 2016] to improve sentence classification using multiple sets of word embeddings. The main idea is to inject to the CNN multiple embeddings matrices from different sources (and hence with different dimensions), where each matrix is treated independently. As in a standard CNN, each matrix passes then to multiple groups of filters with different widths. At the classification layer, the feature vectors for each embedding set are concatenated to form the final vector to be fed into the softmax function.

MGCNN by [Zhang et al., 2016] considers word embeddings (Word2Vec, Glove and syntactic embedding) as input. It is extended here to combine embeddings from multiple levels, as shown in Figure 6.2. Single unit embeddings were considered as provided by EMBCNN_{ch}, EMBCNN_{sw}, and EMBCNN_w, each with its corresponding dimension d_{ch} , d_{sw} and d_w , respectively. The input to the classifier consists of at least two matrices $\mathbf{I}_i$, $2 \leq i \leq 3$, where each $\mathbf{I}_i \in \mathbb{R}^{|S| \times d}$ have its own width $d \in \{d_w, d_{sw}, d_{ch}\}$ and $|S|$ being the unit vocabulary size. The CNN fine-tunes each $\mathbf{I}_i$ to compute the final representations for each unit following the same process. Table 6.4 summarises the hyperparameters used by this model.

¹⁴Compared to SG, CBOW has shown to be slightly faster [Shi et al., 2015].

¹⁵[Zhao et al., 2018] model uses an LSTM instead of a CNN.

Figure 6.2: Multi-group CNN for combining embeddings from different units.

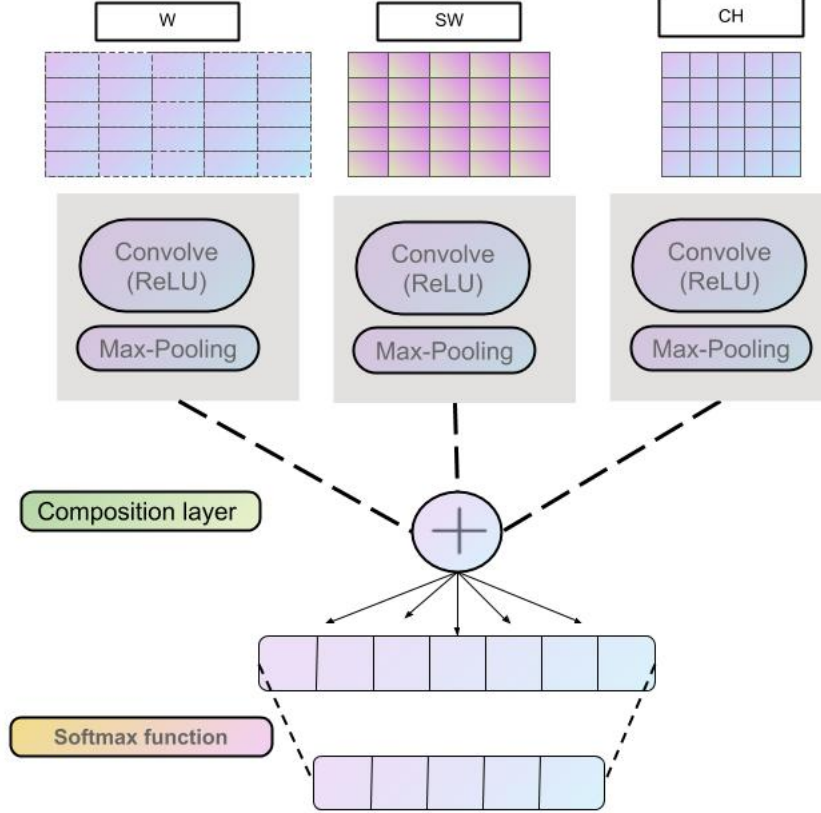


Table 6.4: $\text{MGCNN}_{\text{LEVEL}}$ experimental settings.

Parameters	Best settings
Embeddings size for each unit	ch: 15, sw: 100, w: 300
Number of filters	16
Filter width	2-5
Convolutional layer activation	relu, tanh
Dense layer activation	softmax
Optimizer	Adam

– **Concatenation CNN ($\text{CoCNN}_{\text{LEVEL}}$)**. This model simply concatenates single embeddings as given by EMBCNN_{U} using either standard concatenation or position-based concatenation (cf. Table 6.2). Firstly, each input text t is decomposed into multi-units using a deterministic function that segments each $w \in t$ into an ordered sequence of its constituent characters and characters n-grams, following the same segmentation process already used by the single unit model (cf. Section 6.4.3). The vector corresponding to each unit is then retrieved from the embedding matrix lookup table as given by EMBCNN_{U} . $\text{CoCNN}_{\text{LEVEL}}$ models share the same hyperparameters as $\text{EMBCNN}_{\text{LEVEL}}$ (cf. Table 6.3).

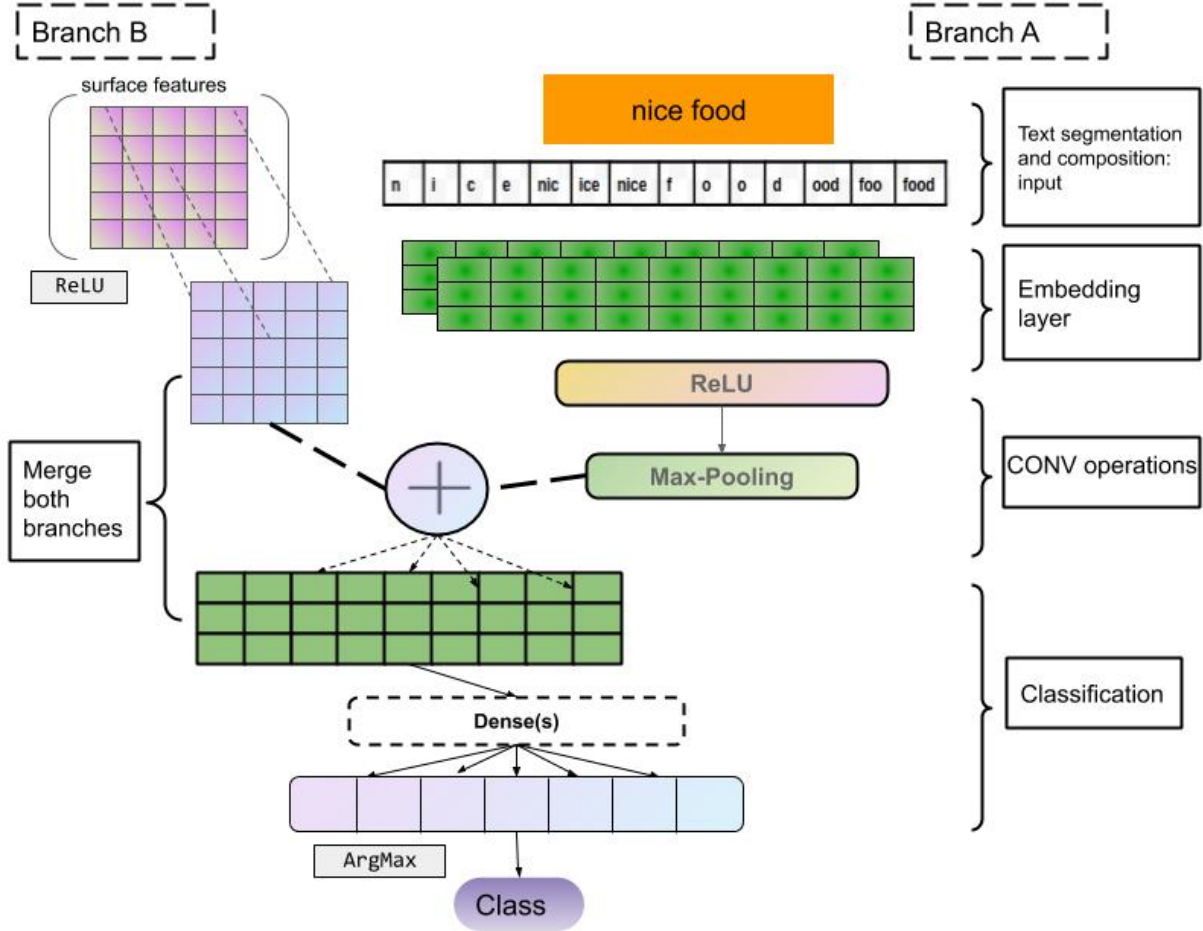
6.4.5 Enriching the Embeddings with Additional Features

Pre-trained word vectors, do not enclose specific sentiment information [Araque et al., 2019], the goal is to augment the semantic features to include sentiment features and task/domain-specific surface features. To address these concerns, we incorporate sentiment and lexical features along with semantic features during the embeddings’ fine-tuning process, which can

provide the classifier more affectual context to help with the Opinion and Emotion Detection (OEA) classification. Following [Lal et al., 2019], we experiment with additional linguistic features (both shallow and semantic) on top of the multi-level representations.

The models described in the last sections are further enriched with task-specific features during the classification process, where features are injected into the CNN classifier at the dense layer without changing its behaviour, as shown in Figure 6.3. As we can see, there are two branches A and B. A performs the convolution operations, where the feature maps are retrieved, then the max-pooling layer is used to abstract the most salient semantic features. The result is a matrix $\mathbf{P} \in \mathbb{R}^{c|S|}$, S being the number of samples and c the number of classes to be predicted. Branch B on the other hand, takes as input the matrix of the additional features $\mathbf{F} \in \mathbb{R}^{f|S|}$, f being the number of features. It then applies the ReLU activation function through a fully connected layer to capture relations between features. $\mathbf{P}$ and $\mathbf{F}$ are merged into $\mathbf{G} \in \mathbb{R}^{(c+f)|S|}$ which is then injected to the final classifier dense layer.

Figure 6.3: Enriching embeddings with additional features at the CNN dense layer.



We rely on features that have already been successfully employed in the sentiment analysis literature for the Arabic language [Mohammad et al., 2016a, Al-Twairsh et al., 2018, Al-Twairsh and Al-Negheimish, 2019]. After performing feature selection using the chi-squared test, we arrive at the following list:

- Surface features: the presence of URLs and user mentions, unigram frequency, number of positive/negative emoticons, number of positive/negative emojis and the sum of positive/negative emoticons and emojis.

- Sentiment features. We make use of several available Arabic sentiment and emotion lexicons ¹⁶: NRC emotion [Mohammad and Turney, 2013b], Arabic NRC hashtags, the dialectal Arabic hashtags, the Arabic translation of Bing Liu’s lexicon, the Arabic MPQA subjectivity [Mohammad et al., 2016a], and EmoLex [Alsmearat et al., 2015]. As we are specifically targeting the ALGD, we finally rely on a manually built lexicon composed of 6,350 positives, 5,160 negatives and 10,090 neutrals (i.e., both positive and negative) Algerian words/expressions [Moudjari et al., 2020]. Given a text, the number of positive/negative words that belong to one of these lexicons is counted.
- Extra-linguistic features: For tweets, we also account for the number of followers and the number of retweets.

The feature-enriched versions of RAND, EMBCNN, CoCNN, and MGCNN models share the same experimental settings with their corresponding non enriched versions, except for the embedding sizes of the $RAND_{LEVEL}$ models, which are empirically set to 100 for $RAND_{CH+SW+FEAT}$ and 150 for all remaining multi-level models.

6.4.6 Summary of the proposed embeddings

All the embeddings proposed in this paper have been designed to account for a specific linguistic unit $U \in \{ch, sw, w\}$, which can be either a character (ch), subword (sw) or a word (w). They have been learned on **AlgSet**, the newly Algerian social media dataset, taking as input various representation levels ranging from a unit $u \in U$ to any combination $level \subseteq U$ of these three units obtained via two different operators, namely concatenation (+) or permutation ($\oplus$). The input representation can also be augmented with task-specific features. Given a model M , the general notation M_{U+Feat} indicates that M has been used to learn the embeddings at U level relying on the additional linguistic features $Feat$. Table 6.5 summarises the models M we proposed in Sections 6.3 and 6.4, along with their abbreviations. These models have been compared to many popular state of the art language models (cf. Section 6.2), as shown in Table 6.6.

Table 6.5: Summary of the different embeddings proposed in this study.

Dialect Specific Language Models	ALG2VEC _w	Word-level Word2Vec embeddings trained on AlgSet .
	ALG2VEC _{sw}	Subword enriched Word2Vec embeddings trained on AlgSet .
	ALGFT	FastText embeddings trained on AlgSet .
	ALG _{ENS}	Ensemble embeddings obtained by combining embeddings from ALG2VEC _w and ALGFT.
CNN-based Language Models	RAND	CNN with randomly generated weights.
	EMBCNN	Single unit embeddings learned using a CNN.
	MGCNN	Multi-group CNN.
	CoCNN	Concatenation of single unit embeddings as given by EMBCNN.

6.5 Conclusion

This chapter presented different dense representations for different word granularities, which help encode syntactic, semantic and morphology-based features in a one-hot vector. They

¹⁶These lexicons can be found here <https://saifmohammad.com/WebPages/ArabicSA.html>.

Table 6.6: Summary of state of the art language models we compare with.

ARAVEC [Soliman et al., 2017]	Arabic word-level embeddings learned on tweet data.
MAZAJAK [Abu Farha and Magdy, 2019]	Arabic word-level embeddings learned on tweet data.
FASTTEXT [Joulin et al., 2016]	Arabic subword level embeddings learned on MSA data collected from Common Crawl and Wikipedia.
BPEMB [Joulin et al., 2016]	Arabic subword level embeddings (BPE segmentation) learned on MSA data collected from Wikipedia.
FLAIR [Joulin et al., 2016]	Arabic character level contextual embeddings learned on MSA data collected from Wikipedia.
BERT [Joulin et al., 2016]	Arabic contextual embeddings learned on MSA data collected from Wikipedia. Word segmentation into subwords/characters is based on BERT WordPiece tokenizer.

can be exploited for different OEA tasks at a relatively low cost and will be available to the community. They will also provide a useful baseline for those working on opinion and emotion analysis in various dialects. These models exploit different neural architectures ranging from standard (Word2Vec) to deep networks (CNN, LSTM). Following [Kim, 2014], a CNN architecture is used to learn embeddings at different granularity levels from a single dialect dataset and use it to classify texts in other dialects. We target for the first time Algerian dialects that lack available vectorial representation.

The proposed models were evaluated on sentiment analysis, emotion detection and for the first time irony detection. For each task, experiments were conducted using well-known datasets, including ones that have been used in shared tasks, to better compare the proposed models with best-performing state of the art models.

The work presented helps analyse the impact of multi-level embeddings on supervised classification of Arabic social media texts written both in MSA and dialectal Arabic (Algerian, Tunisian and Levantine), different segmentation units (character, character n-grams, word) and different composition functions ranging from concatenation to learning the embeddings using different neural architectures. The impact of dialect-specific embeddings on text classification was also studied. Compared to MSA, few dialect embeddings exist and are freely available [Altowayan and Tao, 2016, Soliman et al., 2017, Abu Farha and Magdy, 2019]. Most existing works do not target specific dialects [Vania and Lopez, 2017, Pinter et al., 2017, Zhao et al., 2018], making them agnostic to inter-dialect variations [Qwaider et al., 2018]. This work aims to identify the best input representations to improve classification performances, and in doing so, three main approaches to generate embeddings were studied: (a) fine-tuning existing pre-trained embeddings, (b) creating dialect-specific embeddings from scratch using standard methods with a large unlabelled dataset, and (c) learn embeddings as part of a CNN neural network.

The next chapter provides more details regarding the performances of these models in each task and each dialect and thus answering the following research questions: (1) *What are the best representation levels for processing Arabic social media textual contents?* (2) *To what extent these representations are dialect dependent?* (3) *Are their performances stable across different tasks?*

Chapter 7

Results and Discussion

To evaluate the effectiveness of the models created in this research, extrinsic evaluations were performed on different text classification tasks, namely sentiment analysis, emotion detection and irony detection, while varying the dialects, including Levantine and Maghrebi. For each task, we experiment with well-known datasets, including those used in shared tasks, in order to better compare the results of the proposed models and those reported for the same datasets. This chapter presents the results of this research showing that our multi-level representations outperform word-level representations as well as state of the art models in sentiment and emotion detection and achieve competitive results in irony detection. The proposed models are also the most productive across dialects observing that different dialects require different composition configurations. Finally, it is shown that these performances tend to increase when coupling the multi-level representations with task-specific features.

Keywords Embedding, dense vectors, multi-level vectorial representations, contextual embedding, Convolutional Neural Networks, Arabic dialects, sentiment analysis, emotion detection, irony detection, tests and results.

7.1 Introduction

The literature offers several approaches and models to deal with different textual features. Recently models offering semantic encoding, at different word granularities, have been proposed using both standard models and deep learning models. We, therefore, extended our study to include more salient models ranging from pre-trained to trained language models either standard (i.e. Word2Vec, FastText) or as part of a neural network learning process. However, a direct comparison between those systems is difficult due to various factors. Firstly, some authors report the results on very different datasets, which makes a comparison between the reported results difficult. More importantly, the original authors describe their systems with varying degrees of detail and accuracy, which makes the reported results sometimes not reproducible. With this, even if a method showed excellent results in the authors' evaluation, the lack of details in their publication may render it unusable in practice for the readers.

As far as we know, there has been no analysis conducted on the impact of subword information across different Arabic dialects and different tasks. Here, there is a particular interest in answering the following research questions: (1) *What are the best representation levels for processing Arabic social media textual contents?* (2) *To what extent these representations are dialect dependent?* (3) *Are their performances stable across different tasks?* To this end, it is proposed for the first time in Arabic NLP an in-depth study of the impact of several subwords configurations ranging from character to character n-grams (including word) and explore different composition functions to obtain the final representation of a given text.

The models described in the last chapter were implemented 5.6, relying on a combination of embeddings and the best surface features selected using the chi-squared test. To evaluate the effectiveness of this approach, extrinsic evaluations were performed on different text classification tasks while varying the dialects. The results of the models were compared with strong baselines at different levels, relying on pre-trained embeddings that were fine-tuned while training the said models on the task datasets. We recall that these embeddings are dialect-agnostic and are either standard (Word2Vec, FastText) or contextual (BERT, Flair).

This work target both MSA and dialectal Arabic by considering different orthography and morphology features. The study and the analysis helps identify efficient representations for each dialect and those that are the most stable across tasks and dialects. Where different segmentation units and different composition methods are investigated, either by using algebraic operators or neural networks. The study also covers the impact of dialect-specific embeddings on text classification by targeting the Algerian dialect. Embeddings are learned at different granularity levels from a single dialect dataset and use it to classify texts in other dialects using well-known datasets of sentiment analysis, emotion detection and irony detection. It is demonstrates that models constructed not only achieve state of the art results but also outperform them in practically all use cases for different Arabic corpora in the different tasks, and that extra-linguistic, as well as user-related features, can enhance the accuracy of the systems.

In order to showcase the experiments undertaken, this chapter is organised as follows: section 7.2 details data pre-treatment process. Sections 7.3, 7.4 and 7.5 presents the used datasets as well as the results of the differents models for each dataset in sentiment, emotion and irony detection respectively. Section 7.6 presents our main findings as well as a thorough error analysis for the three tasks, while section 7.7 concludes this chapter.

7.2 Data Pre-treatment and Methodology

Worldwide, expressed opinions and comments constitute a valuable information mine. However, most of the text produced by social websites has an unstructured or noisy nature. This is due to the lack of standardisation, spelling mistakes, missing punctuation, non-standard words, repetitions and more. Indeed, such text needs special treatment.

The purpose at this stage is to prepare the data for the following step, which is the classification of tweets such as "Oooh chaba bzaf" which translates to "ohh it is very beautiful" should be recognised as positive and the sentence "وَعَلَّاش تَدِير هَكَ" meaning "why do you do this?" should be classified as negative. To correctly classify these sentences and others, a set of treatments need to be performed for text pre-processing to transform the texts into a machine-readable format (binary/digital).

- Cleaning: removal of URL links (e.g. <http://example.com>), user mention (e.g. @pseudo - with the symbol @ indicating a user name), all punctuation marks as well as the exaggerations such as "heyyy" replaced by "hey" and the consecutive white spaces were also removed.
- Tokenization: segmentation of texts into BOW.
- Removing stop-words: removal of articles ("و (and)", "نُتُومَا (you)").

In order to analyse the effect of each pre-processing step, each process was tested on a testing corpus of Twifil. Fig 7.1 shows the achievement of the classifier during these experiments, where progress can be seen when each step is applied separately, and when all

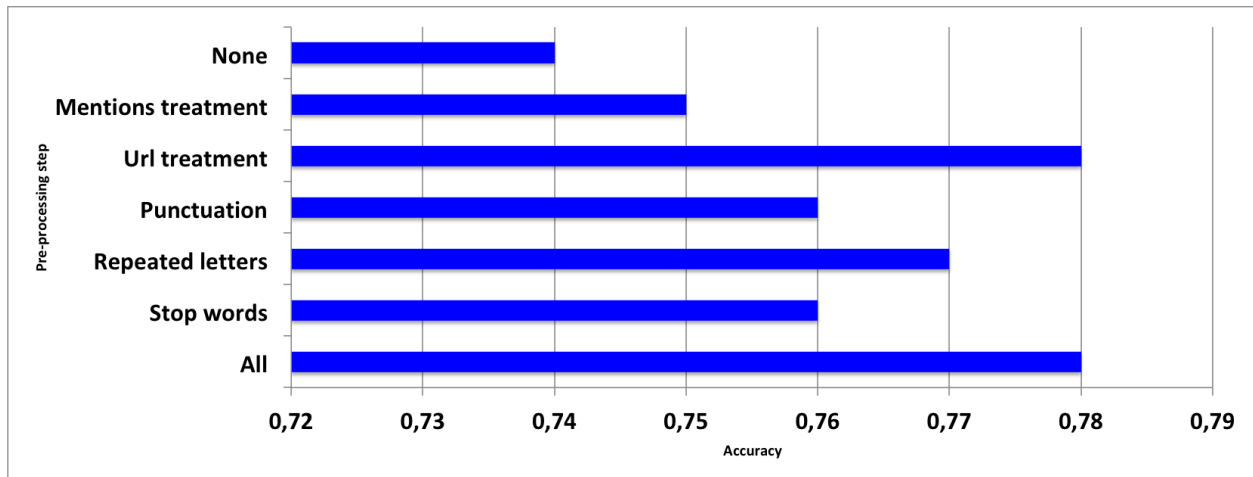


Figure 7.1: The evolution of the classifier’s performance after each pre-processing step

steps are combined. The results demonstrated that pre-processing strategies on the reviews increase the performance of the classifiers.

7.3 Multi-Level Embeddings for Sentiment Analysis

The literature in sentiment analysis being relatively vast, here the focus is on main existing datasets and methods for MSA and its dialects. The reader can refer to [Liu, 2012b, Liu, 2015, Mohammad, 2016, Benamara et al., 2017b, Zhang et al., 2018] for a comprehensive overview of the field.

The analysis of Arabic-speaking opinions and emotions have received much attention in recent years [Baly et al., 2017b, Al-Smadi et al., 2018, Abo et al., 2018]. Most studies focus on MSA [Rushdi-Saleh et al., 2011, Aly and Atiya, 2013], although a few investigated sentiment analysis in Arabic dialects. Among them we cite Gulf [Elnagar et al., 2018], Jordanian [Atoum and Nouman, 2019, Mohammad et al., 2016b], Egyptian [Nabil et al., 2015, Shoukry and Rafea, 2012a], Iraqi [Abdulla et al., 2013, Alnawas and Arici, 2019], Levantine [Baly et al., 2019, Qwaider et al., 2019], Tunisian [Medhaffar et al., 2017, Sayadi et al., 2016], Moroccan mixed with MSA [Oussous et al., 2019], and Algerian [Moudjari and Akli-Astouati, 2019a]. Most of these studies rely on a mixed data written in both MSA and the dialect. Some dialect-agnostic datasets have also been proposed such as the SemEval2017 Task 4 sentiment analysis [Rosenthal et al., 2017] and ArSAS [Elmadany et al., 2018].

Methods vary from standard feature-based techniques [Duwairi and Qarqaz, 2014, Rahab et al., 2017, Nabil et al., 2015, ElSahar and El-Beltagy, 2015, Elouardighi et al., 2017, Baly et al., 2019] to lexicon-based approaches [Boudad et al., 2018b, Mostafa, 2017, Al-Ayyoub et al., 2015]. Deep learning has also been used relying on several Arabic pre-trained language models such as Word2Vec [Altowayan and Tao, 2016, Soliman et al., 2017, Al-Azani and El-Alfy, 2017], Glove [Fouad et al., 2020] and FastText [Kaibi et al., 2019]. Others proposed to learn their own language model, like [Mulki et al., 2019] who trained a neural network with syntax-ignorant n-gram embeddings and [ElJundi et al., 2019] who used the Arabic ULM-Fit. Recently, [Dahou et al., 2019] propose a multi-channel CNN architecture to combining FastText and domain-specific word embeddings. This work goes beyond embeddings at the word level and pre-trained language models by exploring embeddings at different levels when combined with different composition functions.

7.3.1 Datasets

Found here is an evaluation of the performance of the different models using several Arabic sentiment datasets selected according to three main criteria: (a) *availability*, focusing only on freely available corpora, (b) *language variety* in order to target MSA, Maghrebi or Levantine dialects, and (c) *corpus genre* as we wanted to experiment with several social media contents, including tweets, online reviews, blogs and Facebook comments. The datasets are as follows (see Table 7.1 for a general overview):

- **Mataoui** [Mataoui et al., 2016]: It contains 206 Facebook posts and their corresponding 5,042 comments on various topics written in Algerian dialect. Each comment has been manually annotated into positive/negative.
- **Twifil** [Moudjari and Akli-Astouati, 2019a]: This is the largest Algerian dataset of about 11,000 tweets where each tweet has been manually annotated by crowd-sourcing into positive, negative and neutral using a majority vote¹.
- **BBN** [Mohammad et al., 2016b]: It consists of 1,200 Levantine dialectal sentences randomly extracted from the BBN Arabic-Dialect/English Parallel Text, a large dataset of about 3.5 million tokens from different social media posts [Zbib et al., 2012]. Each sentence is provided with a sentiment annotation (positive, negative and neutral).
- **Syria** [Mohammad et al., 2016b]: It is composed of 2,000 tweets originating from Syria, where Levantine dialectal Arabic is commonly spoken. It contains 285 positive and 1,323 negative tweets, collected in May 2014 via the Twitter API and annotated manually.
- **OCA** [Rushdi-Saleh et al., 2011]: It is an MSA movie reviews dataset containing 500 reviews collected from different Arabic web pages and blogs and automatically annotated using the rating as given by the authors of the review (250 are positives and 250 negatives).
- **Sem17** [Rosenthal et al., 2017]: The dataset was used in the SemEval2017 Task 4, subtask A on identifying the overall sentiment of a tweet. The corpus is split into train and test sets containing respectively 3,355 and 6,100 tweets. The corpus does not target a given dialect.
- **TSAC** [Medhaffar et al., 2017]: This is the Tunisian Sentiment Analysis Corpus of about 17k Facebook comments manually annotated into 8,215 positive and 8,845 negative.
- **ArSAS** [Elmadany et al., 2018]: It consists of 21k tweets labelled over four sentiment classes: positive, negative, neutral, and mixed along with their confidence value. Following [Abu Farha and Magdy, 2019], we only kept positive, negative and neutral tweets with a confidence score over 0.5, which results in a total of about 19k tweets.

As one of our objective concerns transferring knowledge from Algerian dialect into other dialects for text classification, it was decided to merge **Mataoui** and **Twifil** to enlarge the Algerian corpus and test the validity of the models on mixed data sources in the same dialect (Facebook comments vs. tweets). This new dataset is referred to as **TwMa**.

In the experiments carried out, the corpora have been divided into train, development and test sets by maintaining the same train/test distribution as the original papers in order to

¹https://github.com/kinmokusu/oea_algd.

Table 7.1: Sentiment analysis datasets used in our experiments. Min(Max) length refers to the minimum(maximum) number of words per instance (tweet, post, review).

	Type	Min/Max length	Positive	Negative	Neutral	Train	Test	Total
Mataoui	FB	4/857	3,032	2,007	–	4,535	504	5,039
Twifil	Tweets	10/34	3,005	2,057	4,704	4,556	506	9,766
TwMa	Tweets+FB	10/857	6,037	4,064	–	9,091	1,010	10,101
BBN	Social Media	11/34	498	575	126	966	107	1,199
Syria	Tweets	21/42	285	1,323	392	1,447	161	2,000
OCA	Movie reviews	394/1,777	250	250	–	450	50	500
Sem17	Tweets	13/35	2,257	3,364	3,834	3,355	6,100	9,455
TSAC	FB	1/154	8,215	8,845	–	13,669	3,400	17,069
ArSAS	Tweets	33/84	4,319	7,450	7,046	16,933	1,882	18,815

better compare with the state-of-the-art results. Table 7.1 provides the number of instances of each class in the datasets as well as the original train/test split. Hyperparameters tuning has been performed using the validation set of each dataset which represents 10% of the train. All the experiments were carried out under the same supervised conditions after standard text pre-processing².

As sentiment annotation scheme differs across datasets (binary vs. three classes) and to better compare the performances when varying the dialects, it was decided to evaluate the models when tested on binary polarity classification, i.e., by using the positive/negative annotations only. Therefore, we do not present our results together with those reported by SOA in the same table. In the following, we first detail the results obtained by our models when evaluated on the Algerian dialect datasets (cf. Section 7.3.2), then show their performances on other dialects (cf. Section 7.3.3). We finally compare with SOA results following this time similar experimental settings by predicting either two or three classes depending on the datasets. The results of these comparaisons are given in Section 7.3.4).

7.3.2 Results for the Algerian Dialect

Baselines

In order to set a general baseline, classical machine learning methods were relied upon including SVM, RF, and ME. Using classical feature-based both statistical and lexicon-based.

The first batch of tests was performed to choose between the different BOW representations (word existence or binary, term frequency, word count, and TF*IDF) using smaller corpora. The results of these experiments showed that term frequency outperformed the others in practically all the tests. Therefore, in the tests, term frequency was used as features for the BOW. Table 7.2 details the experiments performed with classical ML models using uni-grams and bi-grams ("2-model" indicates bi-grams).

Through the experiments, SVM showed a better performance than other models. In fact, the best performance outputs achieved by SVM are 77% as accuracy, 80% as precision, and 77% as F1-score. Therefore, SVM will be referred to as the baseline for the remainder of this chapter.

The results of the lexicon-based approach using different Arabic lexicons are presented in table 7.3. It is seen that the use of the lexicon (8) gave competitive results. It was also noticed that the use of the Arabic Hashtag Lexicon (7) gave good results. However, when the common phrases lexicon was used not all corpora surpassed their baseline. For the

²The same pre-processing steps were used as before in the **AlgSet** unlabelled data (see Section 7.2).

Table 7.2: Baseline for the SA task

	Mataoui		Twifil		Twima	
	A	F	A	F	A	F
SVM	0.77	0.73	0.73	0.73	0.77	0.74
2-SVM	0.65	0.66	0.77	0.77	0.74	0.75
RF	0.58	0.55	0.58	0.58	0.76	0.69
2-RF	0.60	0.65	0.57	0.64	0.76	0.44
ME	0.60	0.65	0.57	0.59	0.78	0.70
2-ME	0.60	0.40	0.60	0.59	0.56	0.64
MLP	0.77	0.73	0.73	0.72	0.76	0.75

remainder, the lexicon will be referred to as the lexicon-based (lex) approach.

Table 7.3: Lexicon-based results of the SA task

		Mataoui corpus		Twifil corpus		Lema	
		A	F	A	F	A	F
1	Bing Liu’s [Mohammad et al., 2016a]	0.75	0.74	0.70	0.71	0.75	0.75
2	MPQA [Mohammad et al., 2016a]	0.76	0.74	0.69	0.70	0.76	0.74
3	EmoLex [Alsmearat et al., 2015]	0.74	0.72	0.71	0.71	0.74	0.73
4	NRC Emoticon [Mohammad and Turney, 2013b]	0.59	0.66	0.32	0.69	0.74	0.75
5	NRC Hashtag	0.52	0.60	0.56	0.46	0.73	0.72
6	Arabic Hashtag	0.78	0.75	0.69	0.68	0.75	0.72
7	Common Phrases [Mataoui et al., 2016]	0.73	0.75	0.77	0.71	0.74	0.73
8	Twilex [Moudjari and Akli-Astouati, 2019a]	0.78	0.75	0.71	0.71	0.73	0.70

Results

All the models were experimented with as described in Chapter 5.6 where each model has at least 54 configurations depending on the input representations: each unit alone (character, subword, word), the combination of several units with different composition functions (concatenation, mean, CNN), and with and without additional features. Due to the high number of configurations (158 in total), presented in this section (and the following sections as well) are the best ones per unit.

Results are shown in Table 7.4, presented in terms of accuracy (Acc) and macro-averaged F1-scores (F1). They are grouped according to the units used to represent the inputs to each model: the word alone, subword (including character) alone and multi-level where several units have been combined. Given a model, the best configurations are followed with *, while for each level, the best results are in bold font. The best model per dataset is highlighted in grey cells. We recall that the baselines are all the pre-trained language models namely: ARAVEC, BPEMB, FASTEXT, FLAIR and BERT. They are marked with a ‡ symbol in the table.

It was observed that ARAVEC performs well compared to word-based models when learned either on task-specific datasets (i.e., $RAND_w$) or on unlabelled large dataset via a CNN (i.e., $EMBCNN_w$). We can also observe that ARAVEC outperforms MAZAJAK in terms of F-score for the three datasets. When embeddings were enriched with additional extra-features, a decrease in performances for all the models was noted except for $RAND_{w+FEAT}$ where results have been boosted for **Twima** by about 8.6% in terms of F-score. The ARAVEC baseline however, was beaten by $ALG2VEC_w$ which shows that dialect specific word embeddings are more effective to capture sentiment information than dialect-agnostic

word embeddings. During the training phase, CBOW models slightly outperformed the SG models. The ensemble model ALG_{ENS} that combines CBOW and SG does not outperform ALG2VEC_w .

Concerning the subword level, the best results were obtained by $\text{EMBCNN}_{\text{sw}}$, $\text{EMBCNN}_{\text{sw}+\text{FEAT}}$ and RAND_{sw} outperforming their character-based counterpart (RAND_{ch} and $\text{EMBCNN}_{\text{ch}}$) as well as FASTTEXT and BPEMB. We note in addition that our FasText-based embeddings ALGFT did not surpass ALG2VEC_w , the Word2Vec variant. For the three corpora, it can be conclude that CNN-based subword representations learned from a large dataset (i.e., **AlgSet**) with randomly initialised weights are more effective than learning character and subword representations derived from ALG2VEC_w (i.e., $\text{EMBCNN}_{\text{ALG2VEC}_{\text{ch}}}$ and $\text{EMBCNN}_{\text{ALG2VEC}_{\text{sw}}}$). The same conclusion can be drawn for task-specific embeddings where RAND_{ch} and RAND_{sw} models achieved lower results compared to CNN but still outperform the baselines by at least 9% in terms of F-score. These experiments show that characters and subwords are less powerful than words for sentiment classification, except for $\text{EMBCNN}_{\text{sw}}$, and that vectors with a size equal to 100 are the most suitable for subword representation.

Table 7.4: Sentiment analysis results on the Algerian dialect datasets when evaluated on a binary classification task (positive vs. negative). ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model per dataset.

		Mataoui		Twifil		Twima	
model		Acc	F1	Acc	F1	Acc	F1
BOW (baseline)		75.12	73.16	73.15	73.07	75.09	74.18
Word-level	ARAVEC ‡	76.13	75.19	73.28	73.02	77.17	73.20
	MAZAJAK	75.09	71.29	66.42	57.86	80.38	66.79
	RAND _w	77.38*	75.31*	72.88*	71.08*	80.91	68.00
	RAND _{w+FEAT}	76.87	75.14	68.28	67.56	81.57	79.39*
	ALG2VEC _w	79.66	78.67	79.72	77.78	80.22	79.12
	ALG _{ENS}	71.17	71.07	72.11	71.03	74.22	72.45
	EMBCNN _w	74.78*	73.72*	75.39*	72.66*	76.57*	74.22*
	EMBCNN _{w+FEAT}	63.76	63.09	60.65	54.01	76.56	65.22
Sub-word-level	FASTTEXT ‡	56.80	56.80	52.17	52.17	62.92	62.92
	BPEMB ‡	62.67	38.30	55.21	39.64	71.11	60.12
	RAND _{ch}	66.87	65.25	69.85	67.97	78.00	56.17
	RAND _{ch+FEAT}	69.84	67.40	60.29	55.29	77.60	70.11
	RAND _{sw}	75.60	73.77	69.73	68.83	82.23	70.47
	RAND _{sw+FEAT}	75.99*	73.95*	74.09*	72.61*	82.70*	81.29
	ALGFT	60.24	59.00	64.32	64.19	66.39	66.34
	ALG2VEC _{sw}	70.23	68.36	70.45	68.52	72.59	71.15
	EMBCNN _{ALG2VEC_{ch}}	57.67	52.78	55.42	51.94	59.72	57.48
	EMBCNN _{ALG2VEC_{sw}}	70.23	65.18	68.15	62.09	74.55	70.12
	EMBCNN _{ch}	67.72	63.74	63.80	60.53	79.77	60.62
	EMBCNN _{ch+FEAT}	67.76	64.28	64.07	58.33	81.44	65.64
	EMBCNN _{sw}	77.55	74.27	76.59	71.41	83.63	78.90*
	EMBCNN _{sw+FEAT}	77.23	77.20	77.65	73.14	80.51	78.63
	BERT ‡	68.52	62.41	63.52	66.58	77.70	70.33
	FLAIR ‡	37.27	42.47	50.64	54.40	52.67	57.31
Multi-level	RAND _{ch⊕w}	77.78	77.96	75.08	74.23	83.03	81.70
	RAND _{sw⊕w}	76.39	75.17	73.61	73.26	81.99	80.14
	RAND _{ch⊕sw⊕FEAT}	79.76	79.63	75.23	74.04	82.35	80.44
	RAND _{w⊕ch⊕sw⊕FEAT}	76.39	75.37	79.03	77.94	82.13	80.61
	MGCNN _{ch+w+FEAT}	63.89*	63.72*	63.34	52.14*	78.48*	68.65
	MGCNN _{sw+w+FEAT}	62.56	58.15	64.96*	51.74	78.19	70.97*
	MGCNN _{ch+sw+FEAT}	63.66	55.72	63.24	49.16	78.22	68.18
	MGCNN _{w+ch+sw+FEAT}	62.90	59.83	62.98	51.25	76.92	68.42
	CoCNN _{ch+w}	66.67*	65.80*	65.69*	52.08*	78.42*	68.94
	CoCNN _{sw⊕w}	62.50	58.74	64.15	49.55	78.04	68.42
	CoCNN _{ch+sw}	61.90	57.34	59.30	43.30	77.15	69.42
	CoCNN _{w+ch+sw}	65.87	65.38	61.46	45.65	78.23	69.61*

With regard to multi-level representations, the results obtained by BERT and FLAIR contextual pre-trained language models are lower than ARAVEC, BERT achieving the best with an F-score of 62.41, 66.58 and 70.33 for respectively **Mataoui**, **Twifil** and **Twima**. Globally, task-specific models are better for polarity classification than multi-group CNN

and concatenation-based CNN architectures. Confirming results reported in previous works where contextual embedding performs the worst when dealing with Arabic contents [Al-Twairesh and Al-Negheimish, 2019]. The best configurations being the combination of either characters with words or characters with subwords. For $\text{RAND}_{\text{CH} \oplus \text{SW}}$ and $\text{RAND}_{\text{CH} \oplus \text{SW} \oplus \text{W}}$. It was observed that the results have been boosted by around 4% in terms of F-score when augmenting the models with extra-features. Also, position-based concatenation of units at different levels outperform standard concatenation for all the configurations. For example, $\text{RAND}_{\text{CH}+\text{SW}+\text{FEAT}}$ obtained an F-score of 74.40 for **Mataoui**, 71.65 for **Twifl** and 75.21 for **Twima**. Finally, MGCNN models performed the best when enriched with surface features and that best level combination is characters coupled with subwords. The same unit combination gave the best results for the COCNN MODELS (i.e., $\text{COCNN}_{\text{CH}+\text{SW}}$).

7.3.3 Results for the Other Dialects

We evaluate here to what extent transferring knowledge from the embeddings trained on Algerian dialect to tasks involving other dialects is feasible. To this end, we compare the performances of our language models with those trained on MSA (FASTTEXT, BPEMB, FLAIR and BERT) and dialect-agnostic large corpora (ARAVEC, MAZAJAK). We recall that we target both dialect-agnostic (**ArSaS**, **Sem17**) and dialect-specific datasets (Tunisian-**TSAC**, Levantine-BBN and **BBN**, **Syria**) which allows for a direct comparison between the models when evaluated on dialect similar datasets (i.e., Mataoui vs. Twifl and BBN vs. Syria). An extensive set of experiments was conducted on these datasets to determine best performing models. Among the tested models, we do not present those that performed badly on the Algerian dialect (cf. ALG_{ENS} , $\text{ALG2VEC}_{\text{SW}}$, $\text{EMBCNN}_{\text{ALG2VEC}_{\text{CH}}}$ and $\text{EMBCNN}_{\text{ALG2VEC}_{\text{SW}}}$, as shown in the previous section).

Results for the word-level are similar to those obtained for the Algerian corpora where both $\text{ALG2VEC}_{\text{W}}$ and $\text{RAND}_{\text{W}+\text{FEAT}}$ were the best. Both models outperform the pre-trained ARAVEC and MAZAJAK word embeddings confirming similar finding reported by previous works where pre-trained models generally fail to deal with Arabic variations [Qwaider et al., 2019, Fouad et al., 2020]. We notice that words embeddings learned on the Algerian dialect outperform all the word-level models for Levantine although not for Tunisian datasets despite it is the closest to the Algerian dialect. For the subword, several models yielded good results depending on the corpus: $\text{EMBCNN}_{\text{SW}+\text{FEAT}}$ for **BBN** and **Syria**, $\text{RAND}_{\text{SW}+\text{FEAT}}$ for **OCA**, $\text{EMBCNN}_{\text{CH}+\text{FEAT}}$ for **Sem17**, $\text{EMBCNN}_{\text{SW}}$ for **TSAC**, and $\text{RAND}_{\text{SW}+\text{W}+\text{FEAT}}$ for **ArSAS** outperforming character-based models including the pre-trained baselines. It is interesting to see that FASTTEXT achieves better results than BPEMB and more importantly, better than ALGFT, the Algerian dialect embeddings built following the FastText model. However, FASTTEXT results are lower compared to $\text{ALG2VEC}_{\text{W}}$. This suggests that creating word embeddings from a specific dialect dataset can be better than leveraging pre-trained dialect agnostic word embeddings enriched with subword information. This can be explained by the fact that Algerian vocabulary shares common words and expressions with other Arabic dialects. This is particularly salient for loan words which are words that have been altered due to code-switching or used in their original form when employing other dialects on social media communications, such as the Egyptian expression *ستين داهية* (*styn dAhyp*/calamity) or the Levantine expression *اهلين* (*Ahlyn*/hello or welcome) that are often employed in Algeria.

On the multi-level side, BERT was far better than FLAIR, especially for **Sem17** and **TSAC** datasets. It is believed this is due to the large vocabulary of both corpora (more than 23k (cf. Table 7.6). As observed in the Algerian corpora, the MGCNN and COCNN models enriched with extra-features were the most effective and that the former is better than the latter. For example, an F-score of 84.46 and 63.85 was obtained for **Syria** and **TSAC**

Table 7.5: Sentiment analysis results on other Arabic dialect datasets when evaluated on a binary classification task (positive vs. negative). ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model per dataset.

Dataset		BBN		Syria		OCA		Sem17		TSAC		ArSAS	
model		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1		
BOW (baseline)		63.16	62.09	78.88	69.57	90.00	89.49	70.26	56.99	53.76	36.97	83.64	83.23
Word-level	ARAVEC ‡	61.21	58.22	81.15	74.51	76.52	76.15	61.22	47.81	51.21	46.22	83.78	82.48
	MAZAJAK	54.42	38.36	78.88	69.57	81.00	80.84	59.15	39.49	54.82	54.76	72.05	70.61
	RAND _w	57.41	47.80*	81.37	73.01	82.00	82.04	64.24	51.87	57.24	57.24	83.75	82.31
	RAND _w +FEAT	57.63*	40.60	86.21	79.00*	98.00	97.80	68.78	52.14*	59.22	59.22	84.07	82.68
	ALG2VEC _w	64.76	63.11	85.68	79.12	81.61	81.16	68.04	51.29	53.35	50.59	85.22	83.95
	EMBCNN _w	60.00	58.72*	80.23	73.36*	79.74*	73.67*	65.72*	51.78	50.72	49.39	77.69	71.90
	EMBCNN _w +FEAT	64.10*	51.22	81.05*	50.13	69.70	67.68	64.15	63.05	55.30*	55.12*	79.46	72.46
Sub-word-level	FASTTEXT ‡	54.96	54.96	69.33	69.33	39.00	39.00	61.91	61.91	45.24	45.24	70.10	63.61
	BPEMB ‡	53.00	35.12	82.05	45.18	54.00	50.43	59.08	44.15	50.44	33.46	36.64	26.82
	RAND _{ch}	52.78	39.46	80.12	44.48	38.57	38.50	59.48	44.36	49.82	43.59	77.58	70.49
	RAND _{ch} +FEAT	52.58	47.93	80.59	46.48	56.00	54.69	60.48	48.36	51.24	51.23	81.18	75.75
	RAND _{sw}	81.37	45.86	61.77	54.43*	54.00	37.87	52.25	52.12	53.85	51.42	87.31	84.18
	RAND _{sw} +FEAT	69.07	68.42*	83.23*	53.29	98.59	92.86	63.48*	54.53*	59.00*	55.88*	88.59	86.77
	ALGFT	55.35	54.40	66.22	64.23	53.02	51.20	57.41	43.06	39.03	29.37	69.53	64.78
	EMBCNN _{ch}	67.59	65.70	80.41	67.42	64.67	60.40	63.38	47.70	57.37	54.33	67.09	63.72
	EMBCNN _{ch} +FEAT	69.22	56.40	83.53	56.70	69.35	67.37	65.84	65.65	57.31	54.35	69.26	68.41
	EMBCNN _{sw}	74.61*	71.67	83.60	80.74	74.39	71.70*	63.74	60.80	63.44	63.35	71.34	66.83
	EMBCNN _{sw} +FEAT	73.41	73.18	83.47	83.46	77.19*	70.40	60.01	60.28	50.76	50.65	69.73	67.82
	BERT ‡	55.35	36.75	65.55	46.39	52.54	35.39	71.47	63.54	61.72	60.48	81.07	80.19
	FLAIR ‡	41.21	47.51	47.64	53.56	50.39	54.57	31.28	47.70	49.18	53.21	54.15	50.23
	RAND _{ch+w}	63.89	61.84	87.58	81.78	92.00	91.89	74.07	54.53	57.38	56.95	89.87	88.43
	RAND _{sw+w} +FEAT	71.30	70.14	85.21	77.36	82.22	78.04	69.31	59.26*	58.43	55.05	91.39	90.25
Multi-level	RAND _{ch+sw} +FEAT	67.67	67.19	88.82*	83.56*	78.00	77.78	73.54	57.97	59.24*	58.66*	91.15	89.88
	RAND _{w+ch+sw} +FEAT	72.22	69.71	85.71	79.12	78.00	76.40	69.84	56.86	55.88	55.87	89.79	88.99
	MGCNN _{ch+w}	61.11*	46.36	80.69	71.28	58.00	42.58	58.48	42.28	60.65	60.44	78.10	76.49
	MGCNN _{sw+w} +FEAT	58.33	48.27*	78.88	69.57	46.00	31.51	57.78	41.36	63.85	63.11	78.50	75.76
	MGCNN _{ch+sw} +FEAT	51.85	35.41	89.44	84.46	60.00*	56.21*	59.44*	44.36*	62.41	62.33	74.69	68.58
	MGCNN _{w+ch+sw}	54.63	38.60	81.37	73.01	42.00	32.21	56.37	40.73	61.18	61.18	77.62	75.68
	CoCNN _{ch+w} +FEAT	59.26*	44.10*	82.61	74.74	62.00	47.46	60.63	47.48*	61.03	60.96	68.97	61.84
	CoCNN _{sw+w} +FEAT	58.33	42.98	84.47	77.36	50.00	33.33	59.63	46.56	61.71	61.21	72.97	65.18
	CoCNN _{ch+sw} +FEAT	56.48	40.77	85.09	78.24	68.00*	55.05*	59.48	44.36	60.62	60.48	70.09	71.39
	CoCNN _{w+ch+sw} +FEAT	52.78	36.46	86.96*	80.89*	60.00	45.00	60.89*	45.58	61.50*	61.40*	68.33	61.19

respectively. For some corpora, such as **BBN** and **ArSAS**, random models performed better. The very good results reported for **OCA** can be explained by the very small number of instances in the test set (around 50 reviews) and the small percentage of unshared words between the train and test sets (39.98%) (see below).

In order to evaluate the impact of the learning procedure when combining the embeddings at different units, we also experimented by concatenating EMBCNN_U with $U \in \{ch, sw\}$ and ALG2VEC_w as done in [Santos and Zadrozny, 2014, Gridach, 2016] who used a CNN architecture to learn character-level representations of words and associate them with Word2Vec word representations. However, the results were lower compared to the MGCNN_{ch+sw} and CoCNN_{ch+sw}. Overall, the most effective models for the different Arabic corpora are the ones relying on subword and multi-level information.

Now that the best models per dataset have been presented, the next step is to measure the ability of these models to handle OOV words. We, therefore, compare the OOV coverage rate of ALG2VEC_w, the best word-level model, with the model given by the best non-random multi-level models across datasets, namely EMBCNN_{sw}. Table 7.6 presents the results. For each corpus, the table gives the vocabulary size $|V|$ for the train/test sets, common words shared by these two sets, as well as the percentage of unseen words in the test set (i.e., present only in the train). The OOV rate is also provided at both the word and subword levels, computed as follows:

$$OOV_{level} = \frac{OOV_{emb}}{Vocab_{level}} \quad (7.1)$$

Where $Vocab_{level}$ is the vocabulary size of a dataset at the given *level* (w for words and sw for subwords) and OOV_{emb} is the number of OOV tokens (words, subwords) from the language model’s vocabulary (ALG2VEC_w and EMBCNN_{sw}). The Table 7.6 highlights that the percentage of unseen words in the test set is quite high surpassing 50% for most datasets except for **OCA** and **TSAC**. OOV_w are relatively high for **Mataoui** (therefore **TwMa**) and **TSAC** probably as they are from another social media platform and that the embeddings learned from tweets fail to capture language variety in Facebook comments. However, subword models succeeded in reducing the OOV rates by at least 3% for these two corpora but also for most other corpora where a decrease of OOV words by around 4% is achieved. The decrease is particularly salient for **ArSAS**. For some datasets (**Twifil**, **Syria** and **OCA**), OOV_{sw} rates have increased compared to OOV_w which is in line with the results reported in Tables 7.4 and 7.5 where EMBCNN_{sw} was not amongst the best performing models for these datasets.

Table 7.6: Percentage of shared vocabulary between the train and test sentiment datasets and out of vocabulary rates as given by ALG2VEC_w (OOV_w) and EMBCNN_{sw} (OOV_{sw}).

Datasets	Train V	Test V	Common words	% of unseen words in the test	OOV_w	OOV_{sw}
Mataoui	15,878	1,152	449	61.02	26.04	23.91
Twifil	14,880	1,732	883	49.02	8.97	11.16
TwMa	34,426	3,177	1,442	54.61	28.61	31.05
BBN	4,723	732	285	61.07	8.69	4.67
Syria	7,401	1,464	864	40.98	7.15	9.28
OCA	36,951	8,752	6,070	30.64	14.59	16.35
Sem17	10,862	17,985	4,237	76.44	13.35	9.63
TSAC	23,066	9,197	5,153	43.97	41.02	35.13
ArSAS	42,809	9,573	6,724	29.76	86.48	30.82

7.3.4 Comparison with State of the Art Results

Table 7.7 presents the results achieved by the models of this work for each dataset along with those reported in previous studies on the same datasets in terms of accuracy and F-scores³. For a fair comparison, the best models were run on the whole datasets, including the neutral class, if available. Hence, a binary classification was performed for **Mataoui**, **TwMa**, **TSAC** and **OCA** and a three-class classification for **Twifil**, **BBN**, **Syria** and **Sem17**. For **ArSAS** and following [Abu Farha and Magdy, 2019], we used the same experimental settings by filtering out annotations with a confidence score less than 0.5 and evaluating the classifier performances on three classes setting (positive, negative, neutral). **ArSAS** F-scores were obtained by averaging the positive and negative class F-scores (excluding the neural class).

From Table 7.7, it is observed that our proposed multi-level models outperform state of the art ones by a significant margin. For the **Mataoui** dataset, [Mataoui et al., 2016] reported an accuracy of 79.13% using a lexicon-based approach relying on a dedicated common phrases sentiment lexicon. With RAND_{CH⊕SW⊕FEAT}, an accuracy of 79.76 was reached and an F1-score of 79.63. The work by [Moudjari et al., 2020] experimented with a three-class classification achieving an accuracy of 76% when tested on the **Twifil** dataset with a CNN architecture and randomly initialised weights (this is similar to RAND_w model). When

³Some studies reported results only in terms of accuracy. In this case, we put – in the F-score column in the table.

Table 7.7: Comparison with state of the art results on the same datasets.

Corpus	Reported		Our best results		Our best models
	Acc	F1	Acc	F1	
Mataoui [Mataoui et al., 2016]	79.13	–	79.76	79.63	RAND _{CH⊕SW⊕FEAT}
Twifil [Moudjari et al., 2020]	76.00	76.00	83.17	79.52	RAND _{W⊕CH⊕SW⊕FEAT}
BBN [Mohammad et al., 2016b]	65.31	–	68.50	58.26	RAND _{SW} , EMBCNN _{SW+FEAT}
Syria [Mohammad et al., 2016b]	79.35	–	81.00	68.00	MGCNN _{CH+SW+FEAT}
OCA [Rushdi-Saleh et al., 2011]	90.60	–	98.59	97.80	RAND _{SW+FEAT} , RAND _{W+FEAT}
Sem17 [Rosenthal et al., 2017]	58.10	61.00	67.61	65.17	RAND _{CH+W} , EMBCNN _{CH+FEAT}
TSAC [Medhaffar et al., 2017]	53.76	36.97	63.85	63.35	MGCNN _{SW+W+FEAT} , EMBCNN _{SW}
ArSAS [Elmadany et al., 2018]	92.00	90.00	90.03	86.99	RAND _{SW} , RAND _{SW+W+FEAT}

using RAND_{W⊕CH⊕SW⊕FEAT} on the same corpus we arrived at 83.17% accuracy and 79.52% F1 surpassing the previous reported results. Finally, subword-based models were also very productive for Tunisian, reaching an accuracy and F-score of about 63% more than 25% over the F-score obtained when using the Multilayer Perceptron, the best performing system reported by [Medhaffar et al., 2017].

For Levantine, [Mohammad et al., 2016b] described a lexicon-based approach using the Arabic Hashtag Lexicon resulting in an accuracy of 65.31% on the **BBN** dataset. As can be seen, RAND_{SW} and EMBCNN_{SW+FEAT} gave better results reaching $acc = 81.37\%$ and $F1 = 73.18\%$ for the binary classification while $acc = 68.50\%$ and $F1 = 58.26$ for the ternary classification (positive, negative and neutral). The same good performances have been obtained by the models on the **Syria** corpus, where a 79.35% accuracy has been reported. The best model was MGCNN_{CH+SW+FEAT} with $acc = 81.00\%$ and $F1 = 68.00\%$.

The presented results on **Sem17**, the dialect-agnostic dataset, were also very encouraging, outperforming best-shared task systems which obtained $acc = 58\%$ and $F = 61.00\%$ using a Naive Base classifier with a combination of lexical and sentiment features trained on the competition train set augmented with about 13K external tweets. Regarding **ArSAS**, the second dialect-agnostic dataset, our results are lower when compared to [Elmadany et al., 2018]. This can be explained by the different neural architecture we used (a CNN rather than a CNN-LSTM). Finally, for **OCA** and **MSA**, both task-specific subword and word models obtained better results than the reported accuracy of 90% using a traditional machine learning model (SVM) with tri-gram TF*IDF as features.

7.4 Multi-Level Embeddings for Emotion Detection

Emotion detection (ED) is generally cast as a multi-class problem that aims at detecting the emotions conveyed in a text from a pre-defined taxonomy of emotions [Mohammad, 2016, Al-Ayyoub et al., 2019, Acheampong et al., 2020]. Well known taxonomies include the Ekman’s classification [Ekman et al., 1980] composed of six basic emotions (*Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise*), and Plutchik’s eight emotions [Plutchik, 1984]. Compared to sentiment analysis, little work has been conducted on Arabic emotion detection. Although, a few, [Rabie and Sturm, 2014, Abdul-Mageed et al., 2016, Almahdawi and Teahan, 2019] attempted to predict Ekman’s basic emotions or [Alhuzali et al., 2018] who followed the Plutchik’s framework. Other studies like [Hussien et al., 2016] propose their own personalised set of emotions. Recently, emotion analysis’ shared tasks have been organised as part of SemEval 2018 [Mohammad et al., 2018]⁴ which helped advance research in the field.

Emotion datasets for the Arabic language resources are scarce. Some of the proposed

⁴<https://competitions.codalab.org/competitions/17751>

datasets include DINA [Abdul-Mageed et al., 2016], LAMA [Alhuzali et al., 2018], the SemEval2018 emotion detection corpus [Mohammad et al., 2018], and Twifil [Moudjari et al., 2020]. The last two are the only publicly available corpora. Regarding automatic detection, different machine learning algorithms were used, such as Naive Base [Rabie and Sturm, 2014, Almahdawi and Teahan, 2019] and SVM [Hussien et al., 2016] relying on the standard bag of words features (term frequency, n-grams and TF*IDF). Other experimented with deep learning models including RNN [Alhuzali et al., 2018], CNN [Baali and Ghneim, 2019], and CNN-LSTM [Abdullah et al., 2018]. Most of these models relied on word embeddings created using task training sets (which are similar to those of this work $RAND_w$ models) [Alhuzali et al., 2018, Abdullah et al., 2018, Baali and Ghneim, 2019] or CBOW [Paetzold, 2018]. To the best of our knowledge, no one has investigated multi-level embeddings on emotion detection. A notable exception is the generation of character-level embeddings as part of an RNN [Paetzold, 2018]. We aim here to go further by exploring different unit combinations, targeting for the first time Algerian dialect together with other Arabic dialects.

7.4.1 Datasets

We rely on the following publicly available datasets (see Table 7.8):

- **Twifil_{emo}** [Moudjari et al., 2020]: the first Algerian corpus composed of tweets annotated manually by crowd-sourcing using a set of 7 emotions (*Happiness, Anger, Disgust, Fear, Sadness, Surprise, and Neutral*).
- **Sem18** [Mohammad et al., 2018]: This is the SemEval-2018 Task 1 Affect in Tweets about predicting emotion intensity. It makes use of the Emotion Classification (E-C) task data ⁵ which contains tweets collected in 2017 and manually annotated into 11 emotions categories (*Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise, and Trust*).

To better compare the performances of the models with those reported in stats or the art, the experiments have been carried out on the same train/test split. **Twifil_{emo}** and **Sem18** annotation schemes were also aligned by removing the neutral class and only keeping the overlapping emotion categories in order to analyse how dialect-specific embeddings can be employed to predict emotions on dialect-agnostic datasets when evaluated on the same task. To measure the impact of neutral emotions on the classification performances of the Algerian dialect dataset, we also experiment with **Twifil_{emo}** without the neutral class, and refer to this dataset as **Twifil_{emo-n}**.

Table 7.8: Emotion datasets used in the experiments carried out.

Corpus	Happiness	Anger	Disgust	Fear	Sadness	Surprise	Neutral	Train	Test	Total
Twifil _{emo}	829	224	175	49	284	161	1,466	2,869	319	3,188
Sem18	605	899	433	391	842	47	–	2,895	322	3,217

7.4.2 Results

Table 7.9 displays the obtained results. We notice that best performing models for sentiment analysis are the same for ED. At the word-level, $ALG2VEC_w$ outperformed $ARAVEC$. Moreover, when enriched with surface features, most models showed improvements compared to their non-enriched version. The best results were obtained with $EMBCNN_{w+FEAT}$ when discarding the neutral class. $EMBCNN_{SW+FEAT}$ and $RAND_{SW}$ were the two models which gave

⁵https://competitions.codalab.org/competitions/17751#learn_the_details-datasets

the best results at the subword level, observing a quite similar scores between $\text{EMBCNN}_{\text{sw}}$ and $\text{RAND}_{\text{sw}+\text{FEAT}}$ for the **Twifl_{emo}** corpus. The use of subword representations with CNN achieves better compared to characters and words. It was also noticed that coupling them with surface features increases the results for **Twifl_{emo-n}**.

Compared to **Twifl_{emo}**, results for **Twifl_{emo-n}** are probably better due to the removal of the dominant class (i.e., neutral) helps the classifier to better focus on tweets that convey emotions. Finally, the results of both BERT and FLAIR contextual embeddings are very low. Again, multi-level models combining either subwords with words ($\text{RAND}_{\text{sw} \oplus \text{w}}$) or characters with words ($\text{CoCNN}_{\text{ch} \oplus \text{w}}$) were the best for **Twifl_{emo}** and **Twifl_{emo-n}** respectively while subwords learned with CNN from dialect-specific embeddings were the best for **Sem18**.

It is important to note that the results reported in Table 7.9 for **Sem18** cannot be compared to SOA scores because our experiments take into account only 6 classes out of 11 and consist of classifying a tweet into one class only. This is different from SemEval 2018 Task 1 which aimed at a multi-label classification where a tweet can be assigned multiple classes [Mohammad et al., 2018].

The results obtained for this last dataset, take into account only 6 classes out of 11 and consists of classifying a tweet into one class. The SemEval 2018 Task 1 however was different since it aimed at a multi-label classification where a tweet can be assigned multiple classes [Mohammad et al., 2018]. The winner system achieved an accuracy of 48.9% and a macro f-score of 46.1% using SVC, a traditional machine learning algorithm outperforming the scores obtained by deep learning approaches with the ARAVEC word embeddings [Badaro et al., 2018a]. When using the same SVC model to predict the whole set of SemEval emotions (i.e., 11 classes), we obtained $\text{acc} = 34.05$ and $F = 16.30$. These results have been improved further by $\text{EMBCNN}_{\text{sw}+\text{FEAT}}$ which achieved $\text{acc} = 39.85\%$ and $F = 23.80$ showing the effectiveness of our CNN-based subword model. One possible explanation for such low results is that the size of the training data per class was very small making deep neural networks fail to generalise well on unseen instances in the test set [Beleites et al., 2013].

Overall, the most effective models for the different corpora are subword-level and multi-level models. These results are also confirmed when dealing with unseen words (see Table 7.10). The table highlights that the percentage of unseen words in the test set is quite high except for the **Sem18** corpus. OOV_w rate is relatively high for **Twifl_{emo}**. However, an improvement can be seen in the OOV coverage rate for most corpora up to 9% for **Twifl_{emo}**. **Sem18** OOV_{sw} rates have increased compared to OOV_w which is in line with the results reported in Table 7.9 where $\text{EMBCNN}_{\text{sw}+\text{FEAT}}$ outperformed $\text{EMBCNN}_{\text{sw}}$.

Table 7.10: Percentage of shared vocabulary between the train and test emotion datasets and out of vocabulary rates as given by ALG2VEC_w (OOV_w) and $\text{EMBCNN}_{\text{sw}}$ (OOV_{sw}).

Datasets	Train V	Test V	Common words	% of unseen words in the test	OOV_w	OOV_{sw}
Twifl _{emo}	6,524	893	346	61.25	16.07	7.49
Twifl _{emo-n}	5,618	729	252	65.43	7.22	6.53
Sem18	12,561	2,989	2,463	17.60	8.48	9.05

7.5 Multi-Level Embeddings for Irony Detection

In addition to sentiment analysis and emotion detection, the models were also tested on irony detection, one of the hardest tasks in NLP. The task is generally binary where a message is classified into ironic vs. non-ironic. It has gained relevance over the years particularly due to its importance in improving sentiment analysis [Zhang et al., 2019, Hernández Farías et al., 2016, Rosso et al., 2018]. Many shared tasks on irony have been proposed in several

languages, such as SemEval 2018 Task 3 for English [Hee et al., 2018], DEFT@TALN2017 for French [Benamara et al., 2017a], IronITA@Evalita 2018 for Italian [Cignarella et al., 2018], IroSvA@IberLEF-2019 for Spanish variants [Ortega-Bueno et al., 2019] (from Spain, Cuba and Mexico), and IDAT@FIRE2019 for Arabic [Ghanem et al., 2019]. This last competition was built on the initial work by [Karoui et al., 2017b] who proposed the first irony detection system for Arabic. Where most popular methods were traditional feature-based with machine learning approaches and/or deep learning methods relying on different language models such as Word2Vec, FastText and BERT [Kayalvizhi et al., 2019, Moudjari and Akli-Astouati, 2019b, Ranasinghe et al., 2019]. The two best systems used an ensemble model (Gradient Boosting, RF and MLP) with surface features (bag of words, TF-IDF, topic modelling) [Khalifa and Hussein, 2019] and BERT in a multi-task learning configuration, BERT being pre-trained on a dialectal Twitter dataset [Zhang and Abdul-Mageed, 2019]. As far as we know, multi-level embeddings for Arabic irony detection is explored here for the first time.

7.5.1 Dataset

Relied upon here was the IDAT dataset⁶ [Ghanem et al., 2019] which contains tweets about different political issues and events related to the Middle East held during 2011 to 2018. Tweets are written in formal Arabic (MSA) and different varieties of the Arabic language: Egypt, Gulf, Levantine, and Maghrebi dialects. Each tweet has been manually annotated into ironic or not ironic. Table 7.11 shows the distribution of tweets in the train and test sets.

The dataset includes tweets related to different political issues and events related to the Middle East held during the years 2011 to 2018. Tweets are written in a formal language (MSA) and in some dialects of Arab countries: Egypt, Gulf, Levantine (the dialect of Syria, Lebanon and Palestine), and Maghrebi dialects.

7.5.2 Results

Table 7.12 presents the results of the experiments conducted for the irony detection task. Once again, the best models are also those which performed the best for sentiment analysis and emotion detection tasks. $ALG2VEC_w$ for word-level surpasses $ARAVEC$ by 6% in both accuracy and F1. $RAND_w$ and $EMBCNN_w$ gave better results compared to their feature-enriched versions. Among subword-level models, $EMBCNN_{sw+FEAT}$ and $RAND_{sw}$ outperform the $FASTTEXT$ and $BPEMB$ baselines, the later achieving 3% F1 over the former. Finally, in the multi-level experiments, $FLAIR$ gave the worst results while $BERT$ performances were similar to the $COCNN$ models (around 65% in terms of F-score). Best models were $RAND_{sw+w+FEAT}$ followed by $COCNN_{sw+w+FEAT}$ and $MGCNN_{w+ch+sw+FEAT}$ with 69.11 and 68.64 F1 respectively. $RAND_{sw+w+FEAT}$ achieved +1.37% accuracy and +1.12% F1 when compared to the first and second IDAT best systems respectively.

With a percentage of 41.43% of unseen words in the IDAT test set and an OOV rate of 9.75% and 11.63% respectively at the word and subword levels, results in Table 7.12 showcase that the models built here are reliable when dealing with irony detection in Arabic content.

⁶<https://www.irit.fr/IDAT2019/>

7.6 Discussions and Error Analysis

7.6.1 Main Conclusions

The main findings of the experiments presented in this chapter are:

- *Transferring knowledge across dialects is feasible which is a first step towards processing dialects that lack pre-trained language models.* It shows that dialects share common vocabulary (see the OOV rate reported in Tables 7.6, 7.10 and 7.13) and that learning embeddings on an unlabelled Algerian large dataset can capture multiple Arabic dialect variations outperforming pre-trained dialect-agnostic subwords and words embeddings for all the dialects and across three different NLP tasks. Therefore, these dialect-aware embeddings can be used even for dialects that lack pre-trained models. These results generalise previous conclusions drawn by [Qwaider et al., 2019] who showed a decrease in performances when training a neural model on an MSA corpus with ARAVEC and applying it directly on a Levantine corpus.
- *Subword and multi-level embeddings are better than word-embeddings for processing Arabic social media contents.* We proposed several ways to learn embeddings at both a single (character, subword and word) and multiple granularity levels using either algebraic composition functions (concatenation, position-based concatenation, mean) and neural networks (1D CNN, multi-group CNN, concatenation CNN). We experimented with randomly initialised weights but also with weights as given by dialect-specific embeddings leading to a total of around 160 different configuration settings. The results suggest that subword and multi-level models were the best outperforming several pre-trained models but also state of the art results on the same datasets. It was observed that different dialect/tasks require different embedding levels and composition functions which is in line with [Zhu et al., 2019] who compared morpheme-based segmentation methods in learning subword representations in different languages (excluding Arabic) and tasks.
- *Randomly generated weights learned either from task-specific datasets or large datasets are more effective than weights given by pre-trained models.* Although it has been shown that weights, as given by pre-trained embeddings slightly outperform models with randomly generated weights [Wang and Xu, 2017, Kocmi and Bojar, 2017], it was displayed in this chapter that random weights can achieve competitive results across dialects and tasks, as already observed by [Watson et al., 2018] for text normalisation. This finding was also confirmed when further experimenting with MGCNN and COCNN by replacing the EMBCNN_w word embeddings by ALG2VEC_w and the results were not conclusive observing a decrease in performances of about 4 to 7% in terms of F-score.
- *Both static (i.e., ARAVEC, MAZAJAK, FASTTEXT, BPEMB) and contextualised (i.e., FLAIR, BERT) state of the art language models fail to fully capture the meaning of Arabic dialects.* The development of contextualised pre-trained models that capture high and low-level language features have substantially improved systems performances on a variety of NLP tasks in many languages including English and French. However, this is not true for morphologically rich languages such as Arabic dialect where both static and contextualised language models fail to account for dialect orthography and variations. Indeed, our experiments clearly showed that state of the art vectorial representations are suboptimal which is in line with conclusions already drawn from previous studies (see [ElJundi et al., 2019, Akyürek et al., 2020]). For example, according to

[ElJundi et al., 2019], this is mainly due to the BERT multilingual tokenizer which fails to tokenize Arabic sentences. On the other hand, the proposed multi-level representations showed better performances when compared to previous results on the same datasets. Although there is still room for improvement, we believe our approach makes a first step towards a better processing of Arabic dialects and hope it will benefit future research in the field.

7.6.2 Error Analysis

In this section, an error analysis is provided of the instances for which the predictions of the best performing models and the manual annotation differ. Table 7.14 presents some examples of misclassified samples (in Arabic along with their English translation) for each dataset, grouped by task. It shows that some classification errors are common across tasks, in particular:

- Presence of ambiguous words such as "أَلَحْم" (dream)" in example (21), and objective words that trigger opinions. For example, in (1), "جعت" (hungry)" can be used to express either positive or negative opinions. Examples (3), (8) and (17) show similar errors.
- Lack of context especially for short texts making it hard to predict the correct class. See examples (2), (5) and (15) in the table.
- Misspellings and grammatical errors due to dialect variation and absence of any standards in writing the dialects, as illustrated by the example (7), where the use of the unknown word "ردورها", can be corrected in two ways and interpreted as "دورها" (turn)" or "ردودها" (answers)", which makes it difficult to understand the full text. As well as example (18), where the word "مجدداً" (again)" was misspelt as "مجداً" (glory)".
- Presence of connotations and idioms, such as "حَاكِيْن كُرُوْشَهْم" (await)" and "مَاسَط لَاسَط" (lame)" in (4) and (15).
- Presence of irony and sarcasm that has an impact on both polarity analysis and emotion detection where positive expressions are usually used to convey negative opinions/emotions, as in (11).

Other errors are task-specific. For sentiment analysis, it was discovered that the most recurrent errors occur when a text contains both positive and negative words, as in (6), (12) and (16), making the model fails to identify the gold class in some cases. For instance in (12), the first segment "خطأ الحكام" (it is the judges' fault)" is negative while "حزام يضيع" (too bad he lost)" positive. Concerning **OCA**, all the wrongly classified reviews from this corpus belonged to the positive class. These reviews generally describe sad movies or extreme crime stories and are characterised by several switches between positive and negative sentiments. **OCA** errors can also be explained by the fact that the reviews are longer than the other corpora (between 394 and 1,777 vs. 857 words per text). Given the imbalanced nature of **Sem17** where the negative class is the majority in both train and test sets, the models fail to generalise over the positive class achieving an $F1 = 38.30$ compared to an $F1 = 93.00$ for the negative class. Finally, error analysis for **TSAC** shows that the lack of context (there are many texts with few words) and the presence of a mixture of posts written in Arabic

and Latin scripts are the main causes of errors. Indeed, the OOV rate of unseen words in this corpus was the highest (cf. 41.02% and 35.13% for word and subword-based models respectively)⁷.

When analysing confusion matrices for emotion detection in the three corpora, it was found that most errors come from the Surprise class followed by Fear and Disgust, due to a lower number of samples per class in the training set. Irony detection on the other hand shares the same errors, where the lack of context was in the majority. Indeed, the interpretation of misclassified tweets requires external contextual knowledge of the text to be understood. For example, in (22) and (23), the reader should be familiar with Egyptian politics. Another possible error is that the additional features shown to be quite effective for sentiment analysis and emotion detection do not necessarily apply for discriminating between ironic and non-ironic tweets.

7.7 Conclusion

In this chapter, the results of the extensive experiments were presented. Indeed, more than 500 tests for each corpus were performed. The first systematic study was proposed here for the first time on the impact of multi-level embeddings for processing Arabic social media contents. The focus was placed on different Arabic content written in both MSA and different Arabic dialects across three text classification tasks (sentiment analysis, emotion detection and irony detection) while varying the corpus genre (tweets, Facebook comments, reviews, blogs) and the dialects targeting both dialect-specific datasets (Algerian, Tunisian, Levantine) and dialect-agnostic datasets containing texts from different Arab regions. As far as we know, no one has conducted such extensive experiments to measure to what extent multi-level representations are dialect/task-dependent. Several deep learning architectures were experimented with (classical machine learning classifiers, basic CNN classifier with randomly initialised weights, own embeddings learned via a CNN from the dialect-specific large unlabelled dataset), and different composition functions (concatenation, position-based concatenation, multi-group CNN, CNN) varying the input vectorial representations (among words, characters and subwords eventually coupled with additional task-specific features) to identify best performing configurations. The results show that multi-level dialect-specific embeddings are the most effective across dialects and tasks outperforming pre-trained language models (AraVec, BPEmb, FastText, Flair, Bert) as well as state of the art results on the same datasets. A thorough manual error analysis shows that language variety, the lack of context and the presence of ambiguous expressions are among the main causes of misclassification.

⁷There was also an attempt to pre-process the corpus by translating Latin words into Arabic and using transliteration. However, the results were not conclusive.

Table 7.9: Results of the emotion detection experiments when evaluated on a multi-class classification task (8 classes). ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model per dataset.

Dataset		Twifil _{emo}		Twifil _{emo-n}		Sem18	
Model		Acc	F1	Acc	F1	Acc	F1
BOW (baseline)		53.60	42.23	66.44	42.24	34.05	16.30
Word-level	ARAVEC ‡	55.04	51.03	61.16	49.32	32.54	33.58
	MAZAJAK	56.22	27.56	59.70	29.55	28.73	19.04
	RAND _w	78.23	51.54*	82.75	54.67*	31.43*	24.20
	RAND _{w+FEAT}	79.01	39.46	80.77	32.32	29.64	26.37*
	ALG2VEC _w	66.41	59.19	70.31	62.49	44.09	39.48
	EMBCNN _w	62.48	51.35	63.27	61.13	40.43	39.29
	EMBCNN _{w+FEAT}	66.36*	54.43*	67.59*	63.78	47.80	44.72
Sub-word-level	FASTTEXT ‡	34.28	34.28	39.04	39.04	34.47	34.47
	BPEMB ‡	52.17	36.18	26.03	22.14	25.19	25.03
	RAND _{ch}	76.86	38.93	79.58	45.48	31.68	25.24
	RAND _{ch+FEAT}	77.47	47.65	82.68	58.76	26.55	20.78
	RAND _{sw}	79.33	52.43	82.21	59.66*	43.17*	47.42*
	RAND _{sw+FEAT}	81.43	52.96*	85.29	57.47	41.72	36.98
	ALGFT	35.18	23.03	41.16	39.13	18.18	14.05
	EMBCNN _{ch}	50.34	42.39	59.58	51.31	45.44	41.42
	EMBCNN _{ch+FEAT}	50.45	42.47	59.72	51.73	45.68	41.74
	EMBCNN _{sw}	64.30	53.18	66.26	57.99	45.59	43.72
	EMBCNN _{sw+FEAT}	65.40*	49.55	67.55*	63.68	49.58	49.53
	BERT ‡	28.04	28.01	39.24	28.14	31.22	33.15
Multi-level	FLAIR ‡	21.29	26.00	33.10	39.00	27.07	30.15
	RAND _{ch⊕w}	80.66	74.72	82.89	50.25	44.41	39.24
	RAND _{sw⊕w}	81.81	77.22	82.59	48.79	41.72	35.48
	RAND _{ch+sw}	81.38	74.64	83.70*	53.31*	44.72	39.30
	RAND _{w⊕ch⊕sw}	80.44	46.39	81.55	48.52	41.61	35.70
	MGCNN _{ch+w+FEAT}	80.26*	69.61*	85.05	67.19	24.84	22.76
	MGCNN _{sw+w+FEAT}	77.89	68.15	82.75	61.69	31.06*	30.12*
	MGCNN _{ch+sw+FEAT}	75.76	65.25	86.78	69.54*	26.90	25.78
	MGCNN _{w+ch+sw+FEAT}	77.53	67.71	81.53	63.25	30.12	28.28
	CoCNN _{ch⊕w}	76.98	66.96	81.02	70.39	34.99*	21.50
	CoCNN _{sw+w}	80.37*	66.49	83.68*	69.31	33.43	17.20
	CoCNN _{ch+sw}	77.53	67.53	80.85	69.49	32.81	20.65
	CoCNN _{w⊕ch⊕sw}	79.08	70.46*	80.42	69.56	33.43	22.42*

Table 7.11: Tweets distribution in the IDAT dataset.

	Ironic	Non Ironic	Train	Test	Total
# of tweets	2,614	2,416	4,024	1,006	5, 030

Table 7.12: Results of the irony detection experiments when evaluated on the IDAT dataset. ‡: baseline models, *: best configuration per model. Bold font: best model per level. Grey cells: best model for the whole dataset.

Model		Acc	F1
BOW (baseline)		68.23	68.09
Best IDAT	Yolo [Khalifa and Hussein, 2019]	83.00	84.40
	UBC [Zhang and Abdul-Mageed, 2019]	81.80	82.40
Word-level	ARAVEC ‡	71.10	71.00
	MAZAJAK	67.20	67.17
	RAND _w	72.40*	72.16*
	RAND _{w+FEAT}	64.62	64.54
	ALG2VEC _w	77.26	77.15
	EMBCNN _w	75.08*	73.26*
	EMBCNN _{w+FEAT}	73.42	72.43
Sub-word-level	FASTTEXT ‡	62.60	62.60
	BPEMB ‡	51.08	48.06
	RAND _{ch}	53.28	53.28
	RAND _{ch+FEAT}	73.42	68.76
	RAND _{sw}	82.69	80.15
	RAND _{sw+FEAT}	79.68	73.69
	ALGFT	51.16	50.33
	EMBCNN _{ch}	63.34	53.51
	EMBCNN _{ch+FEAT}	60.80	56.31
	EMBCNN _{sw}	79.77	76.11
	EMBCNN _{sw+FEAT}	81.34*	77.51*
	BERT ‡	68.50	65.43
Multi-level	FLAIR ‡	28.28	37.16
	RAND _{ch⊕w⊕FEAT}	77.86	52.68
	RAND _{sw+w+FEAT}	84.37	83.52
	RAND _{ch+sw}	76.12	58.35
	RAND _{w+ch+sw+FEAT}	74.88	60.14
	MGCNN _{ch+w+FEAT}	68.59	68.59
	MGCNN _{sw+w}	70.90*	68.11
	MGCNN _{w+ch+sw}	67.79	66.75
	MGCNN _{w+ch+sw+FEAT}	68.69	68.64*
	CoCNN _{ch+w+FEAT}	62.43	59.15
	CoCNN _{sw+w+FEAT}	69.58*	69.11*
	CoCNN _{ch+sw+FEAT}	59.64	58.60
	CoCNN _{w+ch+sw+FEAT}	60.83	58.60

Table 7.13: Percentage of shared vocabulary between the train and test irony datasets and out of vocabulary rates as given by ALG2VEC_w (OOV_w) and EMBCNN_{sw} (OOV_{sw}).

Train V	Test V	Common words	% of unseen words in the test	OOV _w	OOV _{sw}
20,794	7,415	4,343	41.43	9.75	11.63

Table 7.14: Examples of the predicted classes in each dataset as given by the best performing models and their associated gold label.

Sentiment Analysis				
Corpus	Text	Predicted	Gold label	
Mataoui	1 جعت دروك نديرها I am hungry I will prepared it now	Negative	Positive	
	2 مين جيتيها هادي from where did you get this	Positive	Negative	
Twifil	3 أسوأ أحوالنا ربما لكننا واثقون يعيق نهضتنا ولن تضع جهود النهضة سدى فقد أفلحت Our worst maybe!!But we are aware and will not hinder our renaissance and the efforts of the others will not be lost in vain	Positive	Negative	
	4 حنا نفرحوو و المتسولين حاكيمين كزوشهم we rejoice and officials governing await	Positive	Negative	
TwiMa	5 الله يبارك God bless	Negative	Positive	
	6 لصوص محترفين قد لا تربحوا شيئاً سوى خسارة العملاء Professional thieves, you may not win anything but lose customers	Negative	Positive	
BBN	7 مشاركتها روعة وحتى ردورها أحياناً بتخيل عمرها ١٠ سنين برؤدها وأحياناً بتقول هالبت وأعية وعمرها فوق ٢٥ من طريقة كلامها او حتى كتاباتها البسيطة Her posts are amazing, and even her responses, sometimes you imagine Aseel is 10 years old, and sometimes you think she is wiser and over 25 years old from the way she speaks or her simple writings	Positive	Negative	
	8 معايدتي طارت لعندك My greeting flew to you	Negative	Positive	
Syria	9 أفواج الجهاديين اللبان تستمر التدفق سوريا رغم قمع السلطات وقد حث الامام جميل الياس كاستريوت دوكا Albanian jihadi regiments continue to flow Syria despite repression of authorities and Imam Jamil has urged Elias Castriot Doka	Positive	Negative	
	10 سوريتي انت العزف الوحيد ال لمسته اهتر قلبي الما وحرنا My Syria, you are the only sound that I have ever touched my heart shook pain and sorrow	Positive	Negative	
OCA	11 أنا لدي عمل جيد و امرأة تحبني، لكني أترك كل هذا وأني لأدمر حياتي بالشراب، لا أستطيع التغير لأن هذا ما أحبه، إنه الشغف I have a good job and a woman who loves me, but I leave all this and go destroy my life with drinking, I cannot change because this is what I love, it is a passion!!!	Positive	Negative	
Sem17	12 خطأ الحكم حرام يضيع It is the judges' fault. Too bad he lost	Negative	Positive	
	13 الخطيب كان يتباهي الجامعة العربية بمساعدة اميركا The speaker was boasting about the Arab League with America's help	Positive	Negative	
TSAC	14 بعمرى هضمتو وعجبني والنهار الكل ماد وجهو التلفزة ياخي عندوش خدمه وعندوش غايلة يقعد معاها I never liked him, he is everyday on television, he has no work to do, and he has no family to stay with	Positive	Negative	
	15 ماسط لاسط An expression meaning: lame	Positive	Negative	
ArSAS	16 هذا الربيع العربي صحيح وسليم اجثاث روس الفساد !!عبيد حاول تبداء بذارك This Arab Spring is correct and proper, eradicating the Russians of corruption !!	Positive	Negative	
	17 عبد الله بن بخت : الربيع العربي الذي لا يثي بعده خريف Abdullah bin Bakhit: The Arab Spring that will not be followed by an autumn	Negative	Positive	
Emotion Detection				
Twifilemo	18 ألعام جديد ٢٠١٦ كآين مسابقة مضيف طائرة ارجو الرد New Year 2016 is there flight host contest please answer	Joy	Sadness	
	19 أدير لافيتاس لول you are speeding lol	Surprise	Joy	
SemEval-2018-Task-1	20 مرحباً نجمتي يائس ومكتئب ومنزعج جداً صوتي يح شدة حزني خذي بيدي فأنا مشارف الوقوع بتلك الحفرة محبداً Hello my star desperate and depressed and very annoyed my voice with great sadness Take my hand I am the outskirts of falling into that hole glory	Sadness	Anger	
	21 كنزة نجمة لن تكراني بالنسبة الفرح و السعادة انتي كالحلم تتمني ان نستيقظ Kenza is a star will not hate you, you are the joy you are a dream, we hope to wake up	Joy	Surprise	
Irony Detection				
IDAT	22 ألسيسي مكلف نفسه بكره حوار وغدا ده انت طلعك غلاوة شعب Sisi charged himself today and tomorrow discussion you love	Non-ironic	Ironic	
	23 شخص قاله الرئيس مرسي يتصل بحسان يسأله المشوره، كذاب بيتصل بحرامي Someone said the president Morsi calls Hassen to ask advice liar calling a thief	Non-ironic	Ironic	

GENERAL CONCLUSION AND PERSPECTIVES

Conclusion

This work targeted sentiment analysis in social media. The research in this field showed great progress, especially for the English language. However, it was apparent that a gap existed with regard to the Arabic language specifically the dialectal Arabic where North African dialects suffer the most from scarce resources.

Recent advances in the field explore different dense representations in the Euclidean space known as "Embeddings". These representations allow computing different semantic, syntactic and morphologic similarities between different linguistic units from text co-occurrence matrix, and thus became very popular. Units vary from character to character n-grams and words, including more coarse-grained units such as sentences and documents. Several authors [Bojanowski et al., 2017, Li et al., 2018, Zhu et al., 2019, Vania and Lopez, 2017], have highlighted that embeddings at the word level suffer from two main issues. Firstly, they ignore the morphology of words treating different forms of the same word as being unrelated vectors. Secondly, the models are limited to learn vectors for the words present in the vocabulary which only makes systems fail to generalise when the set of unseen words is large during the training phase. These limitations are particularly problematic for morphologically rich languages like Arabic. Indeed, Arabic social media are often characterised by a large variation of unstandardised dialectal Arabic among Egyptian, Gulf, Levantine and Maghrebi [Zaidan and Callison-Burch, 2014], where speakers use both Arabic and Latin letters. More recently, multi-level embeddings combining representations from different units have been proposed as an alternative to single-level embeddings to overcome these limitations and account for the internal structure of words (i.e., morphology) and to help systems overcome the out of vocabulary words problem, subword embeddings where words are represented as character n-grams have become a popular solution [Zhao et al., 2018, Zhu et al., 2019, Bojanowski et al., 2017, Vania and Lopez, 2017]. These representations, either pre-trained or learned, rely either on morphological information as given by external tools [Luong et al., 2013, Qiu et al., 2014] or build via supervised or unsupervised methods that learn the embeddings from text corpora [Cao and Rei, 2016, Bojanowski et al., 2017, Pinter et al., 2017, Li et al., 2018, Heinzerling and Strube, 2018]. Subword models have shown to be quite effective outperforming word-level baselines in several NLP tasks such as machine translation [Luong and Manning, 2016, Chung et al., 2016], part of speech tagging [Ling et al., 2015].

In this work, the focus was given to vector embeddings at different granularity levels targeting both Modern Standard Arabic (MSA) and its different dialects. The aim of the work here is to contribute to this line of research proposing for the first time in Arabic NLP an in-depth study of the impact of several subwords configurations ranging from character to character n-grams (including word) for social media text classification. It proposes several neural architectures to learn character, subword and word embeddings, as well as

a combination of these three levels, exploring different composition functions to obtain the final representation of a given text (both algebraic and neural networks). To evaluate the effectiveness of these representations, extrinsic evaluations were performed on different text classification tasks, namely sentiment analysis, emotion detection and irony detection, while varying the dialects, including Levantine and Maghrebi. For each task, there were experiments conducted with well-known datasets, including those used in recently shared tasks to better compare the results of our models.

The main contributions of this work are thus as follows:

- **A crowd-sourcing annotation platform.** An open platform was created for manual annotation which we called "TWIFIL". Where multi-grained annotations were done at both word and tweet level, as well as multi-level annotations including sentiment, emotion and extra-linguistic information. Resulting in the largest Algerian dataset annotated for both sentiment (about 11,000 tweets), emotion (about 5,000 tweets) and extra-linguistic information including author profiling (age, gender and topic). The annotation resulted also in the creation of the largest Algerien dialect subjectivity lexicon of about 9,000 entries which can constitute a valuable resource for the development of future NLP applications for Algerian dialect. We believe that these resources will help increase research in SA of ALGD and explore new techniques in this field.
- **A multi-level representation of Arabic texts.** We propose various neural architectures to learn character, subword and word embeddings, as well as a combination of these three levels. In addition, we propose enriched models with surface features (text-length, emoticons, count of positive and negative words and intensifiers). All these pre-trained embeddings will be made available for the research community⁸.
- **Comparison of these representations with state of the art pre-trained models.** We first compare with pre-trained subword-agnostic embeddings such as Word2vec [Mikolov et al., 2013c], then with well-known pre-trained character and subword embeddings, namely FastText and BPEmb [Heinzerling and Strube, 2018]. Finally, we compare with contextual embeddings as given by Flair [Akbik et al., 2018] and BERT [Devlin et al., 2019]. We showcase that our models outperform existing pre-trained models.
- **A systematic evaluation of these representations across *different dialects*.** We target MSA along with Levantine and Maghrebi dialects using both dialect-agnostic and dialect-aware embeddings to investigate whether embeddings learned from a given dialect can be successfully used to classify texts written in other dialects. When compared to word-level representations, the results show that our models are the most productive across dialects observing that different dialects require different composition configurations.
- **A systematic evaluation of these representations in *three tasks*,** namely sentiment analysis, emotion detection and irony detection. For each task, experiments with well-known datasets including those which have recently been used in shared tasks are performed to better compare the results here with those reported in previous studies on the same datasets. The results show that multi-level representations carried out in this work outperform state of the art results in sentiment and emotion detection and achieve competitive results in irony detection. More importantly, the results show that performances are stable across the three tasks. Finally, it is shown

⁸https://github.com/kinmokusu/oea_algd

that these performances tend to increase when coupling the multi-level representations with task-specific features.

Indeed, the results conclude that the multi-level representations outperform word-level representation as well as state of the art models in sentiment and emotion detection and achieve competitive results in irony detection. The proposed models are also the most productive across dialects observing that different dialects require different composition configurations. Also, these performances tend to increase when coupling the multi-level representations with task-specific features.

A thorough error analysis was performed, which revealed that the consistency between the predicted and the real classification was generally good, except for borderline texts containing different vocabularies of different classes. Moreover, the models built here performed very well on the irony detection task, which suggests that our models can be used for several classification problems, and also be adapted for other tasks.

The summary of the main findings is as follows:

- *Dialect-specific embeddings are more effective than dialect-agnostic embeddings for Arabic text classification.* Experimental results in the previous chapter show that dialects share a common vocabulary. Hence, learning embeddings on an unlabelled Algerian large dataset helped capture multiple Arabic dialect variations outperforming pre-trained dialect-agnostic subwords and words embeddings for all the dialects and across the three different NLP tasks. Therefore, the dialect-aware embeddings presented in this research can be used even for dialects that lack pre-trained models. These results generalise previous conclusions drawn by [Qwaider et al., 2019] who showed a decrease in performances when training a neural model on an MSA corpus with ARAVEC and applying it directly on a Levantine corpus.
- *Subword and multi-level embeddings are better than word-embeddings for processing Arabic social media contents.* Several ways were proposed to learn embeddings at both a single (character, subword and word) and multiple granularity levels using either algebraic composition functions (concatenation, position-based concatenation, mean) and neural networks (1D CNN, multi-group CNN, concatenation CNN). Randomly initialised weights were experimented with, as well as weights as given by dialect-specific embeddings leading to a total of around 160 different configuration settings. The results suggest that subword and multi-level models were the best, outperforming several pre-trained models and reported state of the art results on the same datasets. It was observed that different dialect/tasks require different embedding levels and composition functions which is in line with [Zhu et al., 2019] who compared morpheme-based segmentation methods in learning subword representations in different languages (excluding Arabic) and tasks.
- *Randomly generated weights learned either from task-specific datasets or large datasets are more effective than weights given by pre-trained models.* Although it has been shown that weights, as given by pre-trained embeddings slightly outperform models with randomly generated weights [Wang and Xu, 2017, Kocmi and Bojar, 2017], it was noted in this chapter that random weights can achieve competitive results across dialects and tasks, as already observed by [Watson et al., 2018] for text normalisation. This finding was confirmed by further experiments with MGCNN and CoCNN when replacing the $EMBCNN_w$ word embeddings by $ALG2VEC_w$ and the results were not conclusive observing a decrease in performances of about 4 to 7% in terms of F-score.

- *Both static (i.e., ARAVEC, MAZAJAK, FASTTEXT, BPEMB) and contextualised (i.e., FLAIR, BERT) state of the art language models fail to fully capture the meaning of Arabic dialects.* The development of contextualised pre-trained models that capture high and low-level language features have substantially improved systems performances on a variety of NLP tasks in many languages including English and French. However, this is not true for morphologically rich languages such as Arabic dialect where both static and contextualised language models fail to account for dialect orthography and variations. Indeed, our experiments clearly showed that state of the art vectorial representations are suboptimal which is in line with conclusions already drawn from previous studies (see [ElJundi et al., 2019, Akyürek et al., 2020]). For example, according to [ElJundi et al., 2019], this is mainly due to the BERT multilingual tokenizer which fails to tokenize Arabic sentences. On the other hand, the proposed multi-level representations showed better performances when compared to previous results on the same datasets. Although there is still room for improvement, we believe our approach makes a first step towards a better processing of Arabic dialects and hope it will benefit future research in the field.

It is hoped that this work will provide useful insights for developing future multi-level language models for Arabic dialects.

Perspectives

Natural language processing is a vast field offering different problematics. This work tackled some of these including text mining, data representation and data analysis targeting opinion and emotion analysis exploring three different tasks (sentiment analysis, emotion detection and irony detection).

In the future, we plan to continue with this research and address the remaining challenges, towards developing additional resources and tools for opinion and emotion analysis of Maghrebian multilingual dialects and use the obtained data to build a multilingual sentiment classifier. As well as implementing and testing other machine learning algorithms and composition functions. It is also planned to complete the development of the TWIFIL platform to allow users to add their own classes and allow the platform to offer part of speech annotations. Although more importantly, to enlarge the corpus and lexicon and explore other features, specifically the use of age and gender. To extend the models to different classification tasks, including the multi-label classification and other languages.

Many other opinion and emotion detection tasks in this domain are also interesting and investigation worthy, e.g., prediction of the fine granularity scores of reviews, construction of semantic models which can provide stronger support for domain and context to avoid ambiguities and improving the multi-lingual opinion and emotion detection.

Also, it is hoped to further explore contextual embeddings and improve their results for Arabic content. An interesting direction is to explore and enrich the newly published AraBert language model [Antoun et al., 2020], as well as hUL-MonA (حُلْمْنَا) [ElJundi et al., 2019] with multi-level information and test their performances on other NLP tasks, such as machine translation.

Bibliography

- [Abdelli et al., 2019] Abdelli, A., Guerrouf, F., Tibermacine, O., and Abdelli, B. (26-27 December 2019). Sentiment analysis of arabic algerian dialect using a supervised method. In *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pages 1–6, Taza - Morocco. IEEE.
- [Abdul-Mageed et al., 2018] Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- [Abdul-Mageed et al., 2016] Abdul-Mageed, M., AlHuzli, H., and Duaa’Abu Elhija, M. D. (2016). Dina: A multidialect dataset for arabic emotion analysis. In *The 2nd Workshop on Arabic Corpora and Processing Tools*, pages 29–37.
- [Abdul-Mageed et al., 2014] Abdul-Mageed, M., Diab, M., and Kübler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.
- [Abdul-Mageed and Diab, 2012] Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914. Citeseer.
- [Abdulla et al., 2013] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., and Al-Ayyoub, M. (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE.
- [Abdulla et al., 2014] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M., Al-Kabi, M. N., and Al-rifai, S. (2014). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, 9(3):55–71.
- [Abdullah et al., 2018] Abdullah, M., Hadzikadicy, M., and Shaikhz, S. (2018). Sedat: sentiment and emotion detection in arabic text using cnn-lstm deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 835–840. IEEE.
- [Abidi and Smaïli, 2017] Abidi, K. and Smaïli, K. (2017). An empirical study of the algerian dialect of social network. In *ICNLSSP 2017-International Conference on Natural Language, Signal and Speech Processing*, pages 1–5.
- [Abidi and Smaïli, 2018] Abidi, K. and Smaïli, K. (2018). An automatic learning of an algerian dialect lexicon by using multilingual word embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 832–838, Miyazaki, Japan. European Language Resources Association (ELRA).

- [Abo et al., 2018] Abo, M. E. M., Ahmed, N., and Balakrishnan, V. (2018). Arabic sentiment analysis: An overview of the ml algorithms. In *Data Science Research Symposium 2018*, page 63.
- [Abu Farha and Magdy, 2019] Abu Farha, I. and Magdy, W. (2019). Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.
- [Abu-Melhim, 1991] Abu-Melhim, A.-R. (1991). Code-switching and linguistic accommodation in arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.
- [Acheampong et al., 2020] Acheampong, F. A., Wenyu, C., and Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*.
- [Adouane et al., 2018] Adouane, W., Dobnik, S., Bernardy, J.-P., and Semmar, N. (2018). A comparison of character neural language model and bootstrapping for language identification in multilingual noisy texts. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 22–31.
- [Adouane and Johansson, 2016] Adouane, W. and Johansson, R. (2016). Gulf arabic linguistic resource building for sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2710–2715.
- [Agerri et al., 2013] Agerri, R., Cuadros, M., Gaines, S., and Rigau, G. (2013). Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51).
- [Aghaei et al., 2012] Aghaei, S., Nematbakhsh, M. A., and Farsani, H. K. (2012). Evolution of the world wide web: From web 1.0 to web 4.0. *International Journal of Web & Semantic Technology*, 3(1):1–10.
- [Agrawal et al., 2020] Agrawal, A., An, A., and Papagelis, M. (2020). Leveraging transitions of emotions for sarcasm detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’20*, pages 1505–1508, New York, NY, USA. Association for Computing Machinery.
- [Akbik et al., 2018] Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.
- [Akyürek et al., 2020] Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., and Wijaya, D. T. (2020). Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.
- [Al-Ayyoub et al., 2015] Al-Ayyoub, M., Essa, S. B., and Alsmadi, I. (2015). Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining (IJSNM)*, 2(2):101–114.
- [Al-Ayyoub et al., 2019] Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., and Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information processing & management*, 56(2):320–342.

- [Al-Azani and El-Alfy, 2017] Al-Azani, S. and El-Alfy, E.-S. M. (2017). Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. *Procedia Computer Science*, 109:359–366.
- [Al-Azani and El-Alfy, 2018] Al-Azani, S. and El-Alfy, E.-S. M. (2018). Combining emojis with arabic textual features for sentiment classification. In *9th International Conference on Information and Communication Systems (ICICS)*, pages 139–144, Irbid, Jordan. IEEE.
- [Al-Aziz et al., 2015] Al-Aziz, A. M. A., Gheith, M., and Eldin, A. S. (2015). Lexicon based and multi-criteria decision making (mcdm) approach for detecting emotions from arabic microblog text. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, pages 100–105. IEEE.
- [Al-Kabi et al., 2016] Al-Kabi, M., Al-Ayyoub, M., Alsmadi, I., and Wahsheh, H. (2016). A prototype for a standard arabic sentiment analysis corpus. *Int. Arab J. Inf. Technol.*, 13(1A):163–170.
- [Al-Khatib and El-Beltagy, 2017] Al-Khatib, A. and El-Beltagy, S. R. (2017). Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 105–114. Springer.
- [Al-Khatib and El-Beltagy, 2019] Al-Khatib, A. and El-Beltagy, S. R. (2019). A simple and effective approach for fine tuning pre-trained word embeddings for improved text classification. *CoRR*, abs/1908.02579.
- [Al-Moslmi et al., 2018] Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., and Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44(3):345–362.
- [Al-Radaideh and Al-Qudah, 2017] Al-Radaideh, Q. A. and Al-Qudah, G. Y. (2017). Application of rough set-based feature selection for arabic sentiment analysis. *Cognitive Computation*, 9(4):436–445.
- [Al-Rfou’ et al., 2013] Al-Rfou’, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.
- [Al Sallab et al., 2015] Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El Hajj, W., and Shaban, K. B. (2015). Deep learning models for sentiment analysis in arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, volume 9, pages 9–17.
- [Al-Smadi et al., 2018] Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., and Gupta, B. (2018). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels reviews. *Journal of computational science*, 27:386–393.
- [Al-Twairish et al., 2017] Al-Twairish, N., Al-Khalifa, H., Al-Salman, A., and Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72.
- [Al-Twairish et al., 2018] Al-Twairish, N., Al-Khalifa, H. S., Al-Salman, A., and Al-Ohali, Y. (2018). Sentiment analysis of arabic tweets: Feature engineering and A hybrid approach. *CoRR*, abs/1805.08533.

- [Al-Twairesh and Al-Negheimish, 2019] Al-Twairesh, N. and Al-Negheimish, H. (2019). Surface and deep features ensemble for sentiment analysis of arabic tweets. *IEEE Access*, 7:84122–84131.
- [Alayba et al., 2018a] Alayba, A. M., Palade, V., England, M., and Iqbal, R. (2018a). A combined cnn and lstm model for arabic sentiment analysis. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 179–191. Springer.
- [Alayba et al., 2018b] Alayba, A. M., Palade, V., England, M., and Iqbal, R. (2018b). Improving sentiment analysis in arabic using word representation. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 13–18. IEEE.
- [Alhuzali et al., 2018] Alhuzali, H., Abdul-Mageed, M., and Ungar, L. (2018). Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, pages 25–35.
- [Almahdawi and Teahan, 2019] Almahdawi, A. J. and Teahan, W. J. (2019). A new arabic dataset for emotion recognition. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 200–216. Springer.
- [Almansor and Al-Ani, 2017] Almansor, E. H. and Al-Ani, A. (2017). Translating dialectal arabic as low resource language using word embedding. In *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 3653–3659, Varna, Bulgaria.
- [Alnawas and Arici, 2019] Alnawas, A. and Arici, N. (2019). Sentiment analysis of iraqi arabic dialect on facebook based on distributed representations of documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–17.
- [Alsmearat et al., 2015] Alsmearat, K., Shehab, M., Al-Ayyoub, M., Al-Shalabi, R., and Kanaan, G. (2015). Emotion analysis of arabic articles and its impact on identifying the author’s gender. In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE.
- [Altowayan and Elnagar, 2017] Altowayan, A. A. and Elnagar, A. (2017). Improving arabic sentiment analysis with sentiment-specific embeddings. In *IEEE International Conference on Big Data (Big Data)*, pages 4314–4320, Boston, MA, USA. IEEE.
- [Altowayan and Tao, 2016] Altowayan, A. A. and Tao, L. (2016). Word embeddings for arabic sentiment analysis. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3820–3825. IEEE.
- [Aly and Atiya, 2013] Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 494–498.
- [Ammar et al., 2016] Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.
- [Antoun et al., 2020] Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding.

- [AP et al., 2014] AP, S. C., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Advances in neural information processing systems*, pages 1853–1861.
- [Araque et al., 2019] Araque, O., Zhu, G., and Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346–359.
- [Artetxe et al., 2017] Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.
- [Arti Buche, 2013] Arti Buche, Dr. M. B. Chandak, A. Z. (June 2013). Opinion mining and analysis: A survey. *International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3*.
- [Asur and Huberman, 2010] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.
- [Atoum and Nouman, 2019] Atoum, J. O. and Nouman, M. (2019). Sentiment analysis of arabic jordanian dialect tweets. *International Journal of Advanced Computer Science and Applications*, 10(2):256–262.
- [Baali and Ghneim, 2019] Baali, M. and Ghneim, N. (2019). Emotion analysis of arabic tweets using deep learning approach. *Journal of Big Data*, 6(1):89–101.
- [Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.
- [Badaro et al., 2019] Badaro, G., Baly, R., Hajj, H., El-Hajj, W., Shaban, K. B., Habash, N., Al-Sallab, A., and Hamdi, A. (2019). A survey of opinion mining in arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):27.
- [Badaro et al., 2018a] Badaro, G., El Jundi, O., Khaddaj, A., Maarouf, A., Kain, R., Hajj, H., and El-Hajj, W. (2018a). EMA at SemEval-2018 task 1: Emotion mining for Arabic. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 236–244, New Orleans, Louisiana. Association for Computational Linguistics.
- [Badaro et al., 2018b] Badaro, G., Jundi, H., Hajj, H., El-Hajj, W., and Habash, N. (2018b). Arsel: A large scale arabic sentiment and emotion lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA)*.
- [Ballesteros et al., 2015] Ballesteros, M., Dyer, C., and Smith, N. A. (2015). Improved transition-based parsing by modeling characters instead of words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal. Association for Computational Linguistics.
- [Baly et al., 2017a] Baly, R., Badaro, G., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., El-Hajj, W., Habash, N., and Shaban, K. (2017a). A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models. In *Proceedings of the third Arabic natural language processing workshop*, pages 110–118.

- [Baly et al., 2017b] Baly, R., Hajj, H., Habash, N., Shaban, K. B., and El-Hajj, W. (2017b). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–21.
- [Baly et al., 2019] Baly, R., Khaddaj, A., Hajj, H., El-Hajj, W., and Shaban, K. B. (2019). Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *CoRR*, abs/1906.01830.
- [Baroni and Lenci, 2010] Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.
- [Beleites et al., 2013] Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., and Popp, J. (2013). Sample size planning for classification models. *Analytica chimica acta*, 760:25–33.
- [Benamara et al., 2017a] Benamara, F., Grouin, C., Karoui, J., Moriceau, V., and Robba, I. (2017a). Analyse d’opinion et langage figuratif dans des tweets présentation et résultats du Défi Fouille de Textes DEFT2017. In *Actes de DEFT@TALN2017*, pages 1–12.
- [Benamara et al., 2017b] Benamara, F., Taboada, M., and Mathieu, Y. (2017b). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.
- [Benito-Osorio et al., 2013] Benito-Osorio, D., Peris-Ortiz, M., Armengot, C. R., and Colino, A. (2013). Web 5.0: the future of emotional competences in higher education. *Global Business Perspectives*, 1(3):274–287.
- [Bentahila and Davies, 1983] Bentahila, A. and Davies, E. E. (1983). The syntax of arabic-french code-switching. *Lingua*, 59(4):301–330.
- [Berners-Lee et al., 1992] Berners-Lee, T., Cailliau, R., Groff, J.-F., and Pollermann, B. (1992). World-wide web: the information universe. *Internet Research*.
- [Bilal et al., 2016] Bilal, M., Israr, H., Shahid, M., and Khan, A. (2016). Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques. *Journal of King Saud University-Computer and Information Sciences*, 28(3):330–344.
- [Bizer, 2009] Bizer, C. (2009). The emerging web of linked data. *IEEE intelligent systems*, 24(5).
- [Bizer et al., 2009] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227.
- [Bizer et al., 2008] Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T. (2008). Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pages 1265–1266. ACM.
- [Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- [Bosma et al., 2009] Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, pages 1–8.

- [Botha and Blunsom, 2014] Botha, J. and Blunsom, P. (2014). Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, pages 1899–1907.
- [Boudad et al., 2018a] Boudad, N., Faizi, R., Thami, R. O. H., and Chiheb, R. (2018a). Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4):2479–2490.
- [Boudad et al., 2018b] Boudad, N., Faizi, R., Thami, R. O. H., and Chiheb, R. (2018b). Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4):2479–2490.
- [Bradley and Lang, 1999] Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology.
- [Bravo-Marquez et al., 2014] Bravo-Marquez, F., Mendoza, M., and Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-based systems*, 69:86–99.
- [Brooke et al., 2009] Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54.
- [Buitelaar et al., 2013] Buitelaar, P., Arcan, M., Iglesias, C., Sanchez-Rada, F., and Strapparava, C. (2013). Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 1–8.
- [Camacho-Collados et al., 2016] Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.
- [Cambria et al., 2013] Cambria, E., Mazzocco, T., and Hussain, A. (2013). Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining. *Biologically Inspired Cognitive Architectures*, 4:41–53.
- [Cao and Rei, 2016] Cao, K. and Rei, M. (2016). A joint model for word embedding and word morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 18–26, Berlin, Germany. Association for Computational Linguistics.
- [Cer et al., 2018] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.
- [Charfuelan and Steiner, 2013] Charfuelan, M. and Steiner, I. (2013). Expressive speech synthesis in mary tts using audiobook data and emotionml. In *INTERSPEECH*, pages 1564–1568.
- [Chaudhary et al., 2018] Chaudhary, A., Zhou, C., Levin, L. S., Neubig, G., Mortensen, D. R., and Carbonell, J. G. (2018). Adapting word embeddings to new languages with

- morphological and phonological subword representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3285–3295. Association for Computational Linguistics.
- [Chen et al., 2015] Chen, X., Xu, L., Liu, Z., Sun, M., and Luan, H. (2015). Joint learning of character and word embeddings. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 953–959, Buenos Aires, Argentina.
- [Cheng et al., 2017] Cheng, K., Li, J., Tang, J., and Liu, H. (2017). Unsupervised sentiment analysis with signed social networks. In *AAAI*, pages 3429–3435.
- [Chiu and Nichols, 2016] Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.
- [Cholakov and Kordoni, 2016] Cholakov, K. and Kordoni, V. (2016). Using word embeddings for improving statistical machine translation of phrasal verbs. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 56–60.
- [Chung et al., 2016] Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.
- [Cignarella et al., 2018] Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., and Rosso, P. (2018). Overview of the EVALITA 2018 task on irony detection in italian tweets (ironita). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*, volume 2263, pages 1–6. CEUR-WS.
- [Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
- [Costa-jussà and Fonollosa, 2016] Costa-jussà, M. R. and Fonollosa, J. A. R. (2016). Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 357–361, Berlin, Germany. Association for Computational Linguistics.
- [Cotterell et al., 2015] Cotterell, R., Müller, T., Fraser, A., and Schütze, H. (2015). Labeled morphological segmentation with semi-markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China.
- [Creutz and Lagus, 2007] Creutz, M. and Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.
- [Dadvar et al., 2012] Dadvar, M., Jong, d. F., Ordelman, R., and Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

- [Dahou et al., 2019] Dahou, A., Xiong, S., Zhou, J., and Elaziz, M. A. (2019). Multi-channel embedding convolutional neural network model for arabic sentiment classification. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):41.
- [Dahou et al., 2016] Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., and Duan, P. (2016). Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of the 26th international conference on computational linguistics: Technical papers*, pages 2418–2427, Osaka, Japan. The COLING 2016 Organizing Committee.
- [Dashtipour et al., 2016] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.
- [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- [Diab et al., 2010] Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74.
- [Ding et al., 2008] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *International conference on web search and data mining*, pages 231–240.
- [Dixit et al., 2017] Dixit, A., Pal, A. K., Temghare, S., and Mapari, V. (2017). Emotion detection using decision tree. *Development*, 4(2).
- [DoniaGamal et al., 2018] DoniaGamal, M. A., El-Horbaty, E.-S. M., and Salem, A.-B. (2018). Opinion mining for arabic dialects on twitter. *Egyptian Computer Science Journal*, 42(4).
- [Duwairi, 2015] Duwairi, R. M. (2015). Sentiment analysis for dialectical arabic. In *2015 6th International Conference on Information and Communication Systems (ICICS)*, pages 166–170. IEEE.
- [Duwairi and Qarqaz, 2014] Duwairi, R. M. and Qarqaz, I. (2014). Arabic sentiment analysis using supervised classification. In *2014 International Conference on Future Internet of Things and Cloud*, pages 579–583, Barcelona, Spain. IEEE.
- [Ekman, 1999a] Ekman, P. (1999a). Basic emotions. *Handbook of cognition and emotion*, 5(6).
- [Ekman, 1999b] Ekman, P. (1999b). Facial expressions. *Handbook of cognition and emotion*, 16(301):e320.
- [Ekman et al., 1980] Ekman, P., Freisen, W. V., and Ancoli, S. (1980). Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125–1134.
- [El-Halees, 2017] El-Halees, A. M. (2017). Arabic opinion mining using distributed representations of documents. In *2017 Palestinian International Conference on Information and Communication Technology (PICICT)*, pages 28–33. IEEE.

- [Elaraby and Abdul-Mageed, 2018] Elaraby, M. and Abdul-Mageed, M. (2018). Deep models for arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274.
- [Elaraby and Zahran, 2019] Elaraby, M. and Zahran, A. (2019). A character level convolutional bilstm for arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 274–278.
- [ElJundi et al., 2019] ElJundi, O., Antoun, W., Droubi, N. E., Hajj, H. M., El-Hajj, W., and Shaban, K. B. (2019). hulmona: The universal language model in arabic. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., and Zitouni, I., editors, *Proceedings of the Fourth Arabic Natural Language Processing Workshop, WANLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 68–77. Association for Computational Linguistics.
- [Elmadany et al., 2018] Elmadany, A., Mubarak, H., and Magdy, W. (2018). Arsas: An arabic speech-act and sentiment corpus of tweets. *OSACT*, 3:20.
- [Elnagar et al., 2018] Elnagar, A., Lulu, L., and Einea, O. (2018). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia computer science*, 142:182–189.
- [Elouardighi et al., 2017] Elouardighi, A., Maghfour, M., Hammia, H., and Aazi, F.-z. (2017). A machine learning approach for sentiment analysis in the standard or dialectal arabic facebook comments. In *3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, pages 1–8, Rabat, Morocco. IEEE.
- [Elrazzaz et al., 2017] Elrazzaz, M., Elbassuoni, S., Shaban, K., and Helwe, C. (2017). Methodical evaluation of arabic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 454–458.
- [ElSahar and El-Beltagy, 2015] ElSahar, H. and El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34, Cairo, Egypt. Springer.
- [Erdmann et al., 2018] Erdmann, A., Zalmout, N., and Habash, N. (2018). Addressing noise in multidialectal word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 558–565, Melbourne, Australia. Association for Computational Linguistics.
- [Fares et al., 2019] Fares, Y., El-Zanaty, Z., Abdel-Salam, K., Ezzeldin, M., Mohamed, A., El-Awaad, K., and Torki, M. (2019). Arabic dialect identification with deep learning and hybrid frequency based features. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 224–228.
- [Farra et al., 2010] Farra, N., Challita, E., Assi, R. A., and Hajj, H. (2010). Sentence-level and document-level sentiment mining for arabic texts. In *2010 IEEE international conference on data mining workshops*, pages 1114–1119. IEEE.
- [Faruqui and Dyer, 2014] Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

- [Fouad et al., 2020] Fouad, M. M., Mahany, A., Aljohani, N., Abbasi, R. A., and Hassan, S.-U. (2020). Arwordvec: efficient word embedding models for arabic tweets. *Soft Computing*, 24(11):8061–8068.
- [Gage, 1994] Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2):23–38.
- [Ghanem et al., 2019] Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., and Rosso, P. (2019). Idat@fire2019: Overview of the track on irony detection in arabic tweets. In *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15.*
- [Ghosh and Veale, 2018] Ghosh, A. and Veale, T. (2018). Ironymagnet at semeval-2018 task 3: A siamese network for irony detection in social media. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 570–575.
- [Giatsoglou et al., 2017] Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sargiannidis, G., and Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.
- [Glorot and Bengio, 2010] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.
- [Glorot et al., 2011] Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.
- [Google, 2013] Google, C. A. (2013). Word2vec.
- [Gouws et al., 2015] Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments.
- [Gridach, 2016] Gridach, M. (2016). Character-aware neural networks for arabic named entity recognition for social media. In *Proceedings of the 6th workshop on South and Southeast Asian natural language processing (WSSANLP2016)*, pages 23–32.
- [Grover and Mitra, 2017] Grover, J. and Mitra, P. (2017). Bilingual word embeddings with bucketed cnn for parallel sentence extraction. In *Proceedings of ACL 2017, Student Research Workshop*, pages 11–16.
- [Guellil et al., 2018] Guellil, I., Adeel, A., Azouaou, F., and Hussain, A. (2018). Sentialg: Automated corpus annotation for algerian sentiment analysis. In *International Conference on Brain Inspired Cognitive Systems*, pages 557–567. Springer.
- [Guellil and Boukhalfa, 2015] Guellil, I. and Boukhalfa, K. (2015). Social big data mining: a survey focused on opinion mining and sentiments analysis. In *Programming and Systems (ISPS), 2015 12th International Symposium on*, pages 1–10. IEEE.
- [Guimaraes et al., 2017] Guimaraes, R. G., Rosa, R. L., De Gaetano, D., Rodriguez, D. Z., and Bressan, G. (2017). Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816.

- [Habash et al., 2012] Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.
- [Habash et al., 2018] Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghoulani, W., Bouamor, H., Zalmout, N., et al. (2018). Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- [Habash et al., 2009] Habash, N., Rambow, O., and Roth, R. (2009). Mada+ token: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, volume 41, page 62.
- [Habash, 2010] Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.
- [Hahnloser and Seung, 2001] Hahnloser, R. H. and Seung, H. S. (2001). Permitted and forbidden sets in symmetric threshold-linear networks. In *Advances in neural information processing systems*, pages 217–223.
- [Harrat et al., 2016] Harrat, S., Meftouh, K., Abbas, M., Hidouci, W.-K., and Smaili, K. (2016). An algerian dialect: Study and resources. *International journal of advanced computer science and applications (IJACSA)*, 7(3):384–396.
- [Harrat et al., 2014] Harrat, S., Meftouh, K., Abbas, M., and Smaïli, K. (2014). Building resources for algerian arabic dialects. In Li, H., Meng, H. M., Ma, B., Chng, E., and Xie, L., editors, *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2123–2127. ISCA.
- [Harrat et al., 2017] Harrat, S., Meftouh, K., and Smaïli, K. (2017). Maghrebi arabic dialect processing: an overview. In *ICNLSSP 2017-International Conference on Natural Language, Signal and Speech Processing*, pages 90–94, ISGA - Casablanca.
- [Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.
- [Heath and Bizer, 2011] Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.
- [Hee et al., 2018] Hee, C. V., Lefever, E., and Hoste, V. (2018). SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018*, pages 39–50.
- [Heigold et al., 2017] Heigold, G., Neumann, G., and van Genabith, J. (2017). An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 505–513, Valencia, Spain.
- [Heinzerling and Strube, 2018] Heinzerling, B. and Strube, M. (2018). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors,

- Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2989–2993, Miyazaki, Japan. European Language Resources Association (ELRA).
- [Hellmann et al., 2013] Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating nlp using linked data. In *International semantic web conference*, pages 98–113. Springer.
- [Hernández Farías et al., 2016] Hernández Farías, D. I., Patti, V., and Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):1–24.
- [Hernández-Farías et al., 2015] Hernández-Farías, I., Benedí, J.-M., and Rosso, P. (2015). Applying basic features from sentiment analysis for automatic irony detection. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 337–344. Springer.
- [Htait, 2018] Htait, A. (2018). Adapted sentiment similarity seed words for french tweets’ polarity classification. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, pages 323–328.
- [Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.
- [Hu et al., 2013] Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM.
- [Hussien et al., 2016] Hussien, W. A., Tashtoush, Y. M., Al-Ayyoub, M., and Al-Kabi, M. N. (2016). Are emoticons good enough to train emotion classifiers of arabic tweets? In *7th International Conference on Computer Science and Information Technology (CSIT)*, pages 1–6, vienna, Austria. IEEE.
- [Ibrahim et al., 2015a] Ibrahim, H. S., Abdou, S. M., and Gheith, M. (2015a). MIKA: A tagged corpus for modern standard arabic and colloquial sentiment analysis. In *2nd IEEE International Conference on Recent Trends in Information Systems, ReTIS 2015, Kolkata, India, July 9-11, 2015*, pages 353–358. IEEE.
- [Ibrahim et al., 2015b] Ibrahim, H. S., Abdou, S. M., and Gheith, M. (2015b). Mika: A tagged corpus for modern standard arabic and colloquial sentiment analysis. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pages 353–358. IEEE.
- [Iglesias and al, 2017] Iglesias, S.-R. and al (2017). Linked data models for sentiment and emotion analysis in social networks. In *Sentiment Analysis in Social Networks*, pages 49–69. Elsevier.
- [Ismail et al., 2018] Ismail, R., Omer, M., Tabir, M., Mahadi, N., and Amin, I. (2018). Sentiment analysis for arabic dialect using supervised learning. In *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–6. IEEE.
- [Jarrar et al., 2014] Jarrar, M., Habash, N., Akra, D. F., and Zalmout, N. (2014). Building a corpus for palestinian arabic: a preliminary study.

- [Jia et al., 2019] Jia, X., Deng, Z., Min, F., and Liu, D. (2019). Three-way decisions based feature fusion for chinese irony detection. *International Journal of Approximate Reasoning*.
- [Joshi et al., 2016] Joshi, A., Prabhu, A., Shrivastava, M., and Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, Osaka, Japan. The COLING 2016 Organizing Committee.
- [Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.
- [Kaibi et al., 2019] Kaibi, I., Nfaoui, E. H., and Satori, H. (2019). A comparative evaluation of word embeddings techniques for twitter sentiment analysis. In *International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, pages 1–4, FEZ, MOROCCO. IEEE.
- [Kalchbrenner et al., 2014] Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.
- [Karoui et al., 2017a] Karoui, J., Farah, B., Moriceau, V., Patti, V., Bosco, C., and Aussenac-Gilles, N. (2017a). Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 262–272.
- [Karoui et al., 2017b] Karoui, J., Zitoune, F. B., and Moriceau, V. (2017b). Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168.
- [Kayalvizhi et al., 2019] Kayalvizhi, S., Thenmozhi, D., Kumar, B. S., and Aravindan, C. (2019). Ssn nlp@ idatfire-2019: Irony detection in arabic tweets using deep learning and featuresbased approaches. In *Proceedings of the IDAT@ FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M.(Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings*, pages 439–444.
- [Kermanidis and Maragoudakis, 2013] Kermanidis, K. L. and Maragoudakis, M. (2013). Political sentiment analysis of tweets before and after the greek elections of may 2012. *International Journal of Social Network Mining*, 1(3-4):298–317.
- [Khalifa and Hussein, 2019] Khalifa, M. and Hussein, N. (2019). Ensemble learning for irony detection in arabic tweets. In *FIRE (Working Notes)*, pages 433–438.
- [Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
- [Kim et al., 2016] Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-aware neural language models. In Schuurmans, D. and Wellman, M. P., editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2741–2749. AAAI Press.

- [Kim et al., 2018] Kim, Y., Kim, K.-M., Lee, J.-M., and Lee, S. (2018). Learning to generate word representations using subword information. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2551–2561, Santa Fe, New Mexico, USA.
- [Kiritchenko et al., 2016] Kiritchenko, S., Mohammad, S. M., and Salameh, M. (2016). Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval ’16*, San Diego, California.
- [Klinger et al., 2018] Klinger, R., Clercq, O. D., Mohammad, S. M., and Balahur, A. (2018). IEST: WASSA-2018 implicit emotions shared task. *CoRR*, abs/1809.01083.
- [Klouche et al., 2020] Klouche, B., Benslimane, S. M., and Bennabi, S. R. (2020). Ooredoo rayek: A business decision support system based on multi-language sentiment analysis of algerian operator telephones. *International Journal of Technology Diffusion (IJTD)*, 11(2):66–81.
- [Kocmi and Bojar, 2017] Kocmi, T. and Bojar, O. (2017). An exploration of word embedding initialization in deep-learning tasks. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 56–64, Kolkata, India. NLP Association of India.
- [Kulkarni and Satapathy, 2020] Kulkarni, A. J. and Satapathy, S. C. (2020). *Optimization in Machine Learning and Applications*. Springer.
- [Labeau and Allauzen, 2017] Labeau, M. and Allauzen, A. (2017). Character and subword-based word representation for neural language modeling prediction. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 1–13.
- [Lal et al., 2019] Lal, Y. K., Kumar, V., Dhar, M., Shrivastava, M., and Koehn, P. (2019). De-mixing sentiment from code-mixed text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 371–377.
- [Lazarus, 1999] Lazarus, R. S. (1999). The cognition-emotion debate: A bit of history. *Handbook of cognition and emotion*, 5(6):3–19.
- [Le and Mikolov, 2014] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, Beijing, China.
- [Levy and Goldberg, 2014] Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.
- [Li et al., 2018] Li, B., Drozd, A., Liu, T., and Du, X. (2018). Subword-level composition functions for learning word embeddings. In *Proceedings of the second workshop on subword/character level models*, pages 38–48.
- [Li and Roth, 2002] Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

- [Ling et al., 2015] Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fernandez, R., Amir, S., Marujo, L., and Luís, T. (2015). Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.
- [Liu, 2012a] Liu, B. (2012a). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- [Liu, 2012b] Liu, B. (2012b). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- [Liu, 2015] Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- [Luong and Manning, 2016] Luong, M.-T. and Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.
- [Luong et al., 2015] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.
- [Luong et al., 2013] Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria.
- [Maamouri and Cieri, 2002] Maamouri, M. and Cieri, C. (2002). Resources for arabic natural language processing. In *International Symposium on Processing Arabic*, volume 1.
- [Maas et al., 2011] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1*, pages 142–150. Association for Computational Linguistics.
- [Manning et al., 2014] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.
- [Mataoui et al., 2016] Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research in Computing Science*, 110:55–70.
- [Medhaffar et al., 2017] Medhaffar, S., Bougares, F., Estève, Y., and Hadrich-Belguith, L. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the third Arabic natural language processing workshop*, pages 55–61.
- [Medhat et al., 2014] Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

- [Meftouh et al., 2012] Meftouh, K., Bouchemal, N., and Smaïli, K. (2012). A study of a non-resourced language: an algerian dialect. In *Spoken Language Technologies for Under-Resourced Languages*.
- [Merity et al., 2017] Merity, S., Keskar, N. S., and Socher, R. (2017). Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.
- [Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- [Mikolov et al., 2013b] Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- [Mikolov et al., 2013c] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems 26*, pages 3111–3119.
- [Miyamoto and Cho, 2016] Miyamoto, Y. and Cho, K. (2016). Gated word-character recurrent language model. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1992–1997, Austin, Texas. Association for Computational Linguistics.
- [Mohammad et al., 2018] Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.
- [Mohammad and Kiritchenko, 2018a] Mohammad, S. and Kiritchenko, S. (2018a). Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- [Mohammad et al., 2016a] Mohammad, S., Salameh, M., and Kiritchenko, S. (2016a). Sentiment lexicons for arabic social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 33–37, Portorož, Slovenia.
- [Mohammad, 2016] Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.
- [Mohammad and Kiritchenko, 2018b] Mohammad, S. M. and Kiritchenko, S. (2018b). Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.
- [Mohammad et al., 2016b] Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016b). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.
- [Mohammad and Turney, 2013a] Mohammad, S. M. and Turney, P. D. (2013a). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.
- [Mohammad and Turney, 2013b] Mohammad, S. M. and Turney, P. D. (2013b). Nrc emotion lexicon. *National Research Council, Canada*, 2.

- [Mohammad et al., 2015] Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.
- [Mostafa, 2017] Mostafa, A. M. (2017). An evaluation of sentiment analysis and classification algorithms for arabic textual data. *International Journal of Computer Applications in Technology*, 158(3):1–8.
- [Moudjari and Akli-Astouati, 2018] Moudjari, L. and Akli-Astouati, K. (2018). Opinion and emotion analysis through the linked data lens. In *ASD 2018: Big data & Applications 12th edition of the Conference on Advances of Decisional Systems*, pages 125–137, Marrakech, Morocco. Intelligente Transportation Proceeding : Big Data, Machine Learning & Cloud Computing, Published by FST Mohammedia / IOSIS, ISBN : 978-9920-35-679-4.
- [Moudjari and Akli-Astouati, 2019a] Moudjari, L. and Akli-Astouati, K. (2019a). Construction et exploitation d’un corpus multilingue algérien pour l’analyse des opinions et des émotions. In *Extraction et Gestion des connaissances (EGC), RNTI-E-35*, pages 321–326, METZ, France.
- [Moudjari and Akli-Astouati, 2019b] Moudjari, L. and Akli-Astouati, K. (December 12-15, 2019b). An embedding-based approach for irony detection in arabic tweets. In Mehta, P., Rosso, P., Majumder, P., and Mitra, M., editors, *Proceedings of the IDAT@ FIRE2019. Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings*, volume 2517 of *CEUR Workshop Proceedings*, pages 409–415, Kolkata, India. CEUR-WS.org.
- [Moudjari et al., 2020] Moudjari, L., Akli-Astouati, K., and Benamara, F. (2020). An algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1195–1203, Marseille, France. European Language Resources Association.
- [Moudjari et al., 2021] Moudjari, L., Benamara, F., and Akli-Astouati, K. (2021). Multi-level embeddings for processing arabic social media contents. *Computer Speech & Language*, 70:101240.
- [Mulki et al., 2019] Mulki, H., Haddad, H., Gridach, M., and Babaoğlu, I. (2019). Syntax-ignorant n-gram embeddings for sentiment analysis of arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 30–39.
- [Nabil et al., 2015] Nabil, M., Aly, M., and Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal.
- [Nasser et al., 2016] Nasser, A., Dinger, K., and Sever, H. (2016). Investigation of the feature selection problem for sentiment analysis in arabic language. *Research in Computing Science*, 110:41–54.
- [Nguyen and Le Nguyen, 2018] Nguyen, H. T. and Le Nguyen, M. (2018). Multilingual opinion mining on youtube—a convolutional n-gram bilstm word embedding. *Information Processing & Management*, 54(3):451–462.
- [Nielsen, 2011] Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

- [Ortega-Bueno et al., 2019] Ortega-Bueno, R., Rangel, F., Hernández Farias, D., Rosso, P., Montes-y Gómez, M., and Medina Pagola, J. E. (2019). Overview of the Task on Irony Detection in Spanish Variants. In Cumberras, M. Á. G., Gonzalo, J., Cámara, E. M., Martínez-Unanue, R., Rosso, P., Carrillo-de-Albornoz, J., Montalvo, S., Chiruzzo, L., Collovini, S., Gutiérrez, Y., Zafra, S. M. J., Krallinger, M., Montes-y-Gómez, M., Ortega-Bueno, R., and Rosá, A., editors, *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421, pages 229–256. CEUR-WS.org.
- [Oussous et al., 2019] Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2019). Asa: A framework for arabic sentiment analysis. *Journal of Information Science*, pages 544–559.
- [Paetzold, 2018] Paetzold, G. (2018). Utfpr at iest 2018: Exploring character-to-word composition for emotion analysis. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 176–181.
- [Pak and Paroubek, 2010] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.
- [Pang et al., 2008] Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135.
- [Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.
- [Papay et al., 2018] Papay, S., Padó, S., and Vu, N. T. (2018). Addressing low-resource scenarios with character-aware embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 32–37.
- [Patel, 2013] Patel, K. (2013). Incremental journey for world wide web: introduced with web 1.0 to recent web 5.0—a survey paper. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(10).
- [Patti et al., 2015] Patti, V., Bertola, F., and Lieto, A. (2015). Arsemotica for arsmeteo.org: Emotion-driven exploration of online art collections. In *FLAIRS Conference*, pages 288–293.
- [Peng and Zhong, 2014] Peng, Q. and Zhong, M. (2014). Detecting spam review through sentiment analysis. *JSW*, 9(8):2065–2072.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar.
- [Perez-Rosas et al., 2012] Perez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning sentiment lexicons in spanish. In *LREC*, volume 12, page 73.
- [Peters et al., 2018a] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- [Peters et al., 2018b] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018b). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- [Pham et al., 2015] Pham, H., Luong, M.-T., and Manning, C. D. (2015). Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94.
- [Phan et al., 2008] Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.
- [Pinter et al., 2017] Pinter, Y., Guthrie, R., and Eisenstein, J. (2017). Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark. Association for Computational Linguistics.
- [Plutchik, 1984] Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.
- [Poria et al., 2016] Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.
- [Prinz, 2004] Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford University Press.
- [Qiu et al., 2014] Qiu, S., Cui, Q., Bian, J., Gao, B., and Liu, T.-Y. (2014). Co-learning of word representations and morpheme representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 141–150, Dublin, Ireland. ACL.
- [Qwaider et al., 2019] Qwaider, C., Chatzikyriakidis, S., and Dobnik, S. (2019). Can modern standard Arabic approaches be used for Arabic dialects? sentiment analysis as a case study. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 40–50, Cardiff, United Kingdom. Association for Computational Linguistics.
- [Qwaider et al., 2018] Qwaider, C., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). Shami: A corpus of levantine arabic dialects. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, pages 3645–3652. European Language Resources Association (ELRA).
- [Rabie and Sturm, 2014] Rabie, O. and Sturm, C. (2014). Feel the heat: Emotion detection in arabic social media content. In *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)*, pages 37–49, Kuala Lumpur, Malaysia. Citeseer.

- [Rahab et al., 2017] Rahab, H., Zitouni, A., and Djoudi, M. (2017). Siaac: Sentiment polarity identification on arabic algerian newspaper comments. In *Proceedings of the Computational Methods in Systems and Software*, pages 139–149. Springer.
- [Ranasinghe et al., 2019] Ranasinghe, T., Saadany, H., Plum, A., Mandhari, S., Mohamed, E., Orasan, C., and Mitkov, R. (2019). RGCL at IDAT: deep learning models for irony detection in arabic language. In Mehta, P., Rosso, P., Majumder, P., and Mitra, M., editors, *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15*, volume 2517 of *CEUR Workshop Proceedings*, pages 416–425. CEUR-WS.org.
- [Rangel et al., 2014] Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., and Daelemans, W. (2014). Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–30.
- [Ravi and Ravi, 2015] Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.
- [Recupero et al., 2015] Recupero, D. R., Presutti, V., Consoli, S., Gangemi, A., and Nuzozese, A. G. (2015). Sentilo: frame-based sentiment analysis. *Cognitive Computation*, 7(2):211–225.
- [Refaee and Rieser, 2014] Refaee, E. and Rieser, V. (2014). An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.
- [Rei et al., 2016] Rei, M., Crichton, G., and Pyysalo, S. (2016). Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, Osaka, Japan. The COLING 2016 Organizing Committee.
- [Rosas et al., 2013] Rosas, V. P., Mihalcea, R., and Morency, L.-P. (2013). Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45.
- [Rosenthal et al., 2017] Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- [Rosso et al., 2018] Rosso, P., Rangel, F., Farías, I. H., Cagnina, L., Zaghouani, W., and Charfi, A. (2018). A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass*, 12(4).
- [Ruder et al., 2019] Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.
- [Rula et al., 2016] Rula, A., Maurino, A., and Batini, C. (2016). Data quality issues in linked open data. In *Data and information quality*, pages 87–112. Springer.
- [Rushdi-Saleh et al., 2011] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011). Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

- [Saad and Ashour, 2010] Saad, M. K. and Ashour, W. M. (2010). Osac: Open source arabic corpora. *Osac: Open source arabic corpora*, 10.
- [Saadane and Habash, 2015] Saadane, H. and Habash, N. (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79.
- [Salameh et al., 2015] Salameh, M., Mohammad, S., and Kiritchenko, S. (2015). Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 767–777.
- [Salem, 2017] Salem, F. (Feb 5, 2017). Social media and the internet of things towards data-driven policymaking in the arab world: Potential, limits and concerns. *The Arab Social Media Report, Dubai: MBR School of Government, Vol. 7, 2017. Available at SSRN: <https://ssrn.com/abstract=2911832>*.
- [Sánchez-Rada and Iglesias, 2016] Sánchez-Rada, J. F. and Iglesias, C. A. (2016). Onyx: A linked data approach to emotion representation. *Information Processing & Management*, 52(1):99–114.
- [Sánchez-Rada et al., 2014] Sánchez-Rada, J. F., Vulcu, G., Iglesias, C. A., and Buitelaar, P. (2014). Eurosentiment: linked data sentiment analysis. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pages 145–148. CEUR-WS. org.
- [Sankoff and Poplack, 1981] Sankoff, D. and Poplack, S. (1981). A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45.
- [Santos and Zadrozny, 2014] Santos, C. D. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st international conference on machine learning (ICML-14)*, pages 1818–1826, Beijing, China.
- [Saroufim et al., 2018] Saroufim, C., Almatarky, A., and Abdel Hady, M. (2018). Language independent sentiment analysis with sentiment-specific word embeddings. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 14–23, Brussels, Belgium. Association for Computational Linguistics.
- [Sayadi et al., 2016] Sayadi, K., Liwicki, M., Ingold, R., and Bui, M. (2016). Tunisian dialect and modern standard arabic dataset for sentiment analysis: Tunisian election context. In *Second International Conference on Arabic Computational Linguistics, ACLING*, pages 35–53, Konya, Turkey.
- [Schröder et al., 2011] Schröder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C., and Zovato, E. (2011). Emotionml—an upcoming standard for representing emotions and related states. *Affective computing and intelligent interaction*, pages 316–325.
- [Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

- [Severyn and Moschitti, 2015] Severyn, A. and Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962, Santiago, Chile.
- [Shapiro and Duh, 2018] Shapiro, P. and Duh, K. (2018). Morphological word embeddings for arabic neural machine translation in low-resource settings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 1–11.
- [Shi et al., 2015] Shi, X., Zhai, J., Yang, X., Xie, Z., and Liu, C. (2015). Radical embedding: Delving deeper to chinese radicals. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 594–598.
- [Shoukry and Rafea, 2012a] Shoukry, A. and Rafea, A. (2012a). Preprocessing egyptian dialect tweets for sentiment mining. In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, pages 47–56.
- [Shoukry and Rafea, 2012b] Shoukry, A. and Rafea, A. (2012b). Sentence-level arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 546–550. IEEE.
- [Siddiqui et al., 2015] Siddiqui, M. A., Dahab, M. Y., and Batarfi, O. A. (2015). Building a sentiment analysis corpus with multifaceted hierarchical annotation. *Int. J. Comput. Ling*, 6(2):11–25.
- [Sigar and Taha, 2012] Sigar, A. H. and Taha, Z. S. (2012). A contrastive study of ironic expressions in english and arabic. *College Of Basic Education Researches Journal*, 12(2):795–815.
- [Sindhwani and Melville, 2008] Sindhwani, V. and Melville, P. (2008). Document-word co-regularization for semi-supervised sentiment analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1025–1030. IEEE.
- [Škrjanec and Pollak, 2016] Škrjanec, I. and Pollak, S. (2016). Topic ontologies of the slovene blogosphere: a gender perspective. In *V: Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, pages 27–28.
- [Smith, 2019] Smith, N. A. (2019). Contextual word representations: A contextual introduction. *CoRR*, abs/1902.06006.
- [Sobkowicz et al., 2012] Sobkowicz, P., Kaschesky, M., and Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470–479.
- [Soliman et al., 2017] Soliman, A. B., Eissa, K., and El-Beltagy, S. R. (2017). Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.
- [Speer and Chin, 2016] Speer, R. and Chin, J. (2016). An ensemble method to produce high-quality word embeddings. *arXiv preprint arXiv:1604.01692*.
- [Sperr et al., 2013] Sperr, H., Niehues, J., and Waibel, A. (2013). Letter n-gram-based input encoding for continuous space language models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 30–39.

- [Strapparava et al., 2004] Strapparava, C., Valitutti, A., et al. (2004). Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, pages 1083–1086. Citeseer.
- [Sykora et al., 2013] Sykora, M. D., Jackson, T., O’Brien, A., and Elayan, S. (2013). Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *IADIS Intelligent Systems and Agents Conference, Prague (Czech Republic)*.
- [Sánchez-Rada and al, 2016] Sánchez-Rada, J. F. and al (2016). Towards a common linked data model for sentiment and emotion analysis. www.sewaproject.eu/files/fbbe2939-91c0-42bd-c87b-d2e98d45cbac.pdf.
- [Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.
- [Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- [Uryupina et al., 2014] Uryupina, O., Plank, B., Severyn, A., Rotondi, A., and Moschitti, A. (2014). Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*, pages 4244–4249.
- [Üstün et al., 2018] Üstün, A., Kurfalı, M., and Can, B. (2018). Characters or morphemes: How to represent words? In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 144–153, Melbourne, Australia. Association for Computational Linguistics.
- [Vania and Lopez, 2017] Vania, C. and Lopez, A. (2017). From characters to words to in between: Do we capture morphology? *CoRR*, abs/1704.08352.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- [Volkova et al., 2013] Volkova, S., Wilson, T., and Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.
- [Wallace, 2015] Wallace, B. C. (2015). Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, 43(4):467–483.
- [Wan, 2008] Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561, Honolulu, Hawaii. Association for Computational Linguistics.
- [Wang and Xu, 2017] Wang, C. and Xu, B. (2017). Convolutional neural network with word embeddings for Chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 1, pages 163–172, Taipei, Taiwan. Asian Federation of Natural Language Processing.

- [Wang et al., 2016] Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.-L., and Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 174:806–814.
- [Watson et al., 2018] Watson, D., Zalmout, N., and Habash, N. (2018). Utilizing character and word embeddings for text normalization with sequence-to-sequence models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 837–843, Brussels, Belgium. Association for Computational Linguistics.
- [Wieting et al., 2016] Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Chara-gram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515, Austin, Texas. Association for Computational Linguistics.
- [Wilson et al., 2004] Wilson, T., Wiebe, J., and Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *aaai*, volume 4, pages 761–769.
- [Wu et al., 2013] Wu, H.-H., Tsai, A. C.-R., Tsai, R. T.-H., and Hsu, J. Y.-J. (2013). Building a graded chinese sentiment dictionary based on commonsense knowledge for sentiment analysis of song lyrics. *J. Inf. Sci. Eng.*, 29(4):647–662.
- [Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.
- [Xiao and Cho, 2016] Xiao, Y. and Cho, K. (2016). Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*.
- [Xie et al., 2017] Xie, X., Ge, S., Hu, F., Xie, M., and Jiang, N. (2017). An improved algorithm for sentiment analysis based on maximum entropy. *Soft Computing*, pages 1–13.
- [Xu et al., 2019] Xu, D., Laparra, E., and Bethard, S. (2019). Pre-trained contextualized character embeddings lead to major improvements in time normalization: a detailed analysis. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 68–74, Minneapolis, Minnesota. Association for Computational Linguistics.
- [Yaghan, 2008] Yaghan, M. A. (2008). Arabizi: A contemporary style of arabic slang. *Design issues*, 24(2):39–52.
- [Yamada et al., 2018] Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2018). Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. *arXiv e-prints*, page arXiv:1812.06280.
- [Ye et al., 2009] Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, 36(3):6527–6535.
- [Yerva et al., 2010] Yerva, S. R., Miklós, Z., and Aberer, K. (2010). It was easy, when apples and blackberries were only fruits. Technical report, WePS-3, colocated with CLEF.

- [You et al., 2016] You, Q., Luo, J., Jin, H., and Yang, J. (2016). Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 13–22. ACM.
- [Zahran et al., 2015] Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., and Atyia, A. (2015). Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–443. Springer.
- [Zaidan, 2012] Zaidan, O. F. (2012). *Crowdsourcing Annotation for Machine Learning in Natural Language Processing Tasks*. Johns Hopkins University.
- [Zaidan and Callison-Burch, 2014] Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.
- [Zbib et al., 2012] Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.
- [Zhai et al., 2018] Zhai, Z., Nguyen, D. Q., and Verspoor, K. (2018). Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 38–43, Brussels, Belgium. Association for Computational Linguistics.
- [Zhang and Abdul-Mageed, 2019] Zhang, C. and Abdul-Mageed, M. (2019). Multi-task bidirectional transformer representations for irony detection. In Mehta, P., Rosso, P., Majumder, P., and Mitra, M., editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 391–400. CEUR-WS.org.
- [Zhang and Abdul-Mageed, 2019] Zhang, C. and Abdul-Mageed, M. (2019). No army, no navy: Bert semi-supervised learning of arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284.
- [Zhang et al., 2018] Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.
- [Zhang et al., 2019] Zhang, S., Zhang, X., Chan, J., and Rosso, P. (2019). Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633 – 1644.
- [Zhang et al., 2015] Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.
- [Zhang et al., 2016] Zhang, Y., Roller, S., and Wallace, B. C. (2016). MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1522–1527, San Diego, California. Association for Computational Linguistics.

- [Zhao et al., 2016] Zhao, J., Liu, K., and Xu, L. (2016). Sentiment analysis: mining opinions, sentiments, and emotions.
- [Zhao et al., 2018] Zhao, J., Mudgal, S., and Liang, Y. (2018). Generalizing word embeddings using bag of subwords. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 601–606, Brussels, Belgium. Association for Computational Linguistics.
- [Zheng et al., 2018] Zheng, L., Wang, H., and Gao, S. (2018). Sentimental feature selection for sentiment analysis of chinese online reviews. *International journal of machine learning and cybernetics*, 9(1):75–84.
- [Zhu et al., 2019] Zhu, Y., Vulić, I., and Korhonen, A. (2019). A systematic study of leveraging subword information for learning word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 912–932, Minneapolis, Minnesota. Association for Computational Linguistics.
- [Zimbra and Chen, 2012] Zimbra, D. and Chen, H. (2012). Scalable sentiment classification across multiple dark web forums. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*, pages 78–83. IEEE.
- [Zou et al., 2013] Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.
- [Zribi et al., 2014] Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. H., and Habash, N. (2014). A conventional orthography for tunisian arabic. In *LREC*, pages 2355–2361.