Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022, pages 68–75
Marseille, 24 June 2022

Ara-Women-Hate: An annotated corpus dedicated to Hate speech detection against women in the Arabic community

Imane Guellil[1], Ahsan Adeel[3], Faical Azouaou[2],Mohamed Boubred[4],Yousra Houichi[5], Akram Abdelhaq Moumna[2]
[1]University of Edinburgh, Edinburgh, United Kingdom
[2]Laboratoire des Méthodes de Conception des Systèmes (LMCS), Ecole nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie,
[3]School of Mathematics and Computer Science, University of Wolverhampton,
[4]Capgemini, France
[5]Factory Digitale, Algeria,
imane.guellil@ed.ac.uk

Abstract

In this paper, an approach for hate speech detection against women in the Arabic community on social media (e.g. Youtube) is proposed. In the literature, similar works have been presented for other languages such as English. However, to the best of our knowledge, not much work has been conducted in the Arabic language. A new hate speech corpus (Arabic_fr_en) is developed using three different annotators. For corpus validation, three different machine learning algorithms are used, including deep Convolutional Neural Network (CNN), long short-term memory (LSTM) network and Bi-directional LSTM (Bi-LSTM) network. Simulation results demonstrate the best performance of CNN model which achieved an F1-score up to 86% for the unbalanced corpus as compared to LSTM and Bi-LSTM.

Keywords: Hate speech detection; Arabic language; Sexism detection; Deep learning

1. Introduction

With the online proliferation of hate speech, an important number of research studies have been presented in the last few years. The majority of these studies detect general hate speech (Burnap and Williams, 2014; Davidson et al., 2017; Wiegand et al., 2018) and focused on detecting sexism and racism on social media (Waseem and Hovy, 2016; Pitsilis et al., 2018; Kshirsagar et al., 2018). In contrast, only a few studies (Saha et al., 2018) focused on the detection of hate speech against women (only by distinguishing between hateful and non hateful comments). However, almost all studies are dedicated to English where other languages such as Arabic is also one of the four top used languages on the Internet (Guellil et al., 2018c; Guellil et al., 2018a; Guellil et al., 2021)). To bridge the gap, in this paper, we propose a novel approach to detect hate speech against women in Arabic community.

2. Background

2.1. Hate speech

Different definitions of hate speech are adopted by the research literature. However, the definition of (Nockleby, 2000) was recently largely used by many authors such as, (De Smedt et al., 2018; Schmidt and Wiegand, 2017; Zhang et al., 2018; Madisetty and Desarkar, 2018) and (Zhang and Luo, 2018). According to Nockleby, *"Hate speech is commonly defined as any communication that disparages or defames a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics"* (Nockleby, 2000). For illustrating how this hate can be presented in textual exchange, (Schmidt and Wiegand, 2017) provided some examples:

- Go fucking kill yourself and die already useless ugly pile of shit scumbag.
- The Jew Faggot Behind The Financial Collapse.
- Hope one of those bitches falls over and breaks her leg.

Based on the recent survey of (Schmidt and Wiegand, 2017), we decided to use the term *Hate speech* (which is the most commonly used) rather than other terms present in the literature for the same phenomenon such as: *abusive speech* (Andrusyak et al., 2018; Gorrell et al., 2018), *offensive language* (Risch et al., 2018; Pitsilis et al., 2018; Puiu and Brabete, 2019) or *cyberbullying* (Dadvar and Eckert, 2018; Van Hee et al., 2018). According to Chetty and Alathur (Chetty and Alathur, 2018), hate speech is categorized into four categories: gendered hate speech (including any form of misogyny, sexism, etc), religious hate speech (including any kind of religious discrimination, such as: Islamic sects, anti-Christian, anti-Hinduism, etc), racist hate speech (including any sort of racial offence or tribalism, xenophobia, etc) and disability (including any sort of offence to an individual suffering from health which limits to do some of the life activities) (Al-Hassan and Al-Dossari, 2019). However, this survey neglected a category which could influence important international outcomes which is political hate speech. Political hate speech can be referred to any abuse, offence, injuries regarding politicians.

2.2. Arabic in social media

Arabic is one of the six official languages of the United Nations[1] (Eisele and Chen, 2010; Ziemski et al., 2016; Guellil and Azouaou, 2016). It is the official language of 22 countries. It is spoken by more than 400 million speakers. Arabic is also recognized as the 4th most used language of the Internet (Al-Kabi et al., 2016; Boudad et al., 2017). All the works in the literature (Habash, 2010; Farghaly and Shaalan, 2009; Harrat et al., 2017; Guellil et al., 2019; **?**; Guellil et al., 2017a) classify Arabic in three main varieties: 1) Classical Arabic (CA) which is the form of Arabic language used in literary texts. The Quran [2] is considered to be the highest form of CA text (Sharaf and Atwell, 2012a). 2) Modern Standard Arabic (MSA) which is used for writing as well as formal conversations. 3) Dialectal Arabic which is used in daily life communication, informal exchanges, etc (Boudad et al., 2017). However, Arabic speakers on social media, discussion forums and Short Messaging Service (SMS) often use a non standard romanization called 'Arabizi' (Darwish, 2014; Bies et al., 2014). For example, the Arabic sentence: رَاني فرحَانة, which means I am happy, is written in Arabizi as 'rani fer7ana'. Hence, Arabizi is an Arabic text written using Latin characters, numerals and some punctuation (Darwish, 2014; Guellil et al., 2018a). Moreover, most of Arabic people are bilingual, where the Mashreq side (Egypt, Gulf, etc) often use English and the Maghreb side (Tunisia, Algeria, etc) often use French, as second language. This linguistic richness contributes to increase a well known phenomenon on social media which is *code switching*. Therefore, Arabic pages also contain messages such as: "رَاني super فرحَانة" or "رَاني very فرحَانة" meaning I am very happy. In addition, messages purely written in French or in English are also possible.

Many studies have been proposed, in order to deal with Arabic and Arabizi (Darwish, 2014; Guellil et al., 2017b). Extracting opinions, analysing sentiments and emotion represent an emerging research area for Arabic and its dialects (Guellil et al., 2017c; Guellil et al., 2018b; Imane et al., 2019). However, few studies are dedicated to analyze extreme negative sentiments such as hate speech. Arabic hate speech detection is relatively a new research area where we were able to to collect only few works. these approaches are described in more details in the following section.

[1] http://www.un.org/en/sections/about-un/official-languages/

[2] The Quran is a scripture which, according to Muslims, is the verbatim words of Allah containing over 77,000 words revealed through Archangel Gabriel to Prophet Muhammad over 23 years beginning in 610 CE. It is divided into 114 chapters of varying sizes, where each chapter is divided into verses, adding up to a total of 6,243 verses. The work of Sharaf et al. (Sharaf and Atwell, 2012b)

3. Related work

3.1. Hate speech detection

3.1.1. General hate speech detection

Burnap and Williams (Burnap and Williams, 2014) investigated the spread of hate speech after Lee Rigby murder in UK. The authors collected 450,000 tweets and randomly picked 2,000 tweets for the manual annotation conducted by CrowdFlower (CF) workers[3]. Each tweet was annotated by 4 annotators. The final dataset contains 1,901 annotated tweets. The authors used three classification algorithmsand the best achieved classification results were up to 0.77 (for F1-score) using the Binary Logistic Regression (BLR). Davidson et al. (Davidson et al., 2017) distinguished between hateful and offensive speech by applying the Logistic Regression (LR) classifier The authors automatically extracted a set of tweets and manually annotated 24,802, randomly selected by CF workers. Their model achieved an F1 score of 0.90 but suffered poor generalization capability with up to 40% misclassification. Weigand et al. (Wiegand et al., 2018) also focused on the detection of abusive language. The authors used several features and lexical resources to build an abusive lexicon. Afterwards, constructed lexicon in an SVM classification was used. In this work, publicly available datasets were used (Razavi et al., 2010; Warner and Hirschberg, 2012; Waseem and Hovy, 2016).

It is to be noted that all the aforementioned studies have been conducted with English language. However, a few other studies in some other languages are also conducted recently such as Italian (Del Vigna12 et al., 2017), German (Köffer et al., 2018), Indonesian (Alfina et al., 2017), Russian (Andrusyak et al., 2018). However, only a limited number of researches have focused on hate speech detection in Arabic language. Abozinadah et al. (Abozinadah et al., 2015) evaluated different machine learning algorithms to detect abusive Arabic tweets. The authors manually selected and annotated 500 accounts associated to the abusive extracted tweets and used three classification algorithms The best results were obtained with the Naîve Bayes (NB) classifier with F1-score up to 0.90. Mubarek et al. (Mubarak et al., 2017) focused on the detection and classification of the obscene and offensive Arabic tweets. The authors used the Log Odds Ration (LOR) For evaluation, the authors manually annotated 100 tweets and obtained a F1-score up to 0.60. Haidar et al. (Haidar et al., 2017) proposed a system to detect and stop cyberbullying on social media. The authors manually annotated a dataset of 35,273 tweets from Middle East Region (especially from Lebanon, Syria, Gulf Area and Egypt). For classification, the authors used SVM and NB and obtained the best results with SVM achieving F1-score up to 0.93. More recently, Alakrot et al. (Alakrot et al., 2018) described a step by step

[3] https://www.figure-eight.com/

construction of an offensive dataset of Youtube Arabic comments. The authors extracted 167,549 Youtube comments from 150 Youtube video. For annotation, 16,000 comments were randomly picked (annotated by 3 annotators). Finally, Albadi et al. (Albadi et al., 2018) addressed the detection of Religious Arabic hate speech. The authors manually annotated 6,136 tweets (where 5,569 were used for training and 567 for testing). For feature extraction, AraVec (Soliman et al., 2017) was used.

3.1.2. Sexism detection (Hate speech against women)

Waseem et al. (Waseem and Hovy, 2016) used LR classification algorithm to detect sexism and racism on social media. The authors manually annotated a dataset containing 16,914 tweets where 3,383 tweets are for sexist content, 1,972 for racist content, and 11,559 for neither sexist or racism. For dataset generation, the authors used Twitter API for extracting tweets containing some keywords related to women. The authors achieved F1-score up to 0.73. The work of Waseem et al. (Waseem and Hovy, 2016) is considered as a benchmark by many researchers (Al-Hassan and Al-Dossari, 2019; Pitsilis et al., 2018; Kshirsagar et al., 2018). The idea of Pitsilis et al. (Pitsilis et al., 2018) is to employ a neural network solution composed of multiple Long-Short-Term-Memory (LSTM) based classifiers in order to detect sexism and racism in social media. The authors carried out many experiments achieving the best F1-score of 0.93. Kshirsagar et al. (Kshirsagar et al., 2018) also focused on racism and sexism detection and their approach is also based on neural networks. However, in this work, the author also used word embedding for extracting feature combining with a Muli-Layer Perception (MLP) based classifier. The best achieved F1-score was up to 0.71. Saha et al. (Saha et al., 2018) presented a model to detect hate speech against women. The authors used several algorithms to extract features such as bag-of-words (BOW), TF-IDF and sentence embeddings with different classification algorithms such as LR, XGBoost and CatBoost. The best achieved F1-score was 0.70 using LR classifier. Zhang et al. (Zhang et al., 2018) proposed a hybrid model combining CNN and LSTM to detect hate speech. The authors applied their model to 7 datasets where 5 are publicly available (Waseem and Hovy, 2016; Waseem, 2016; Gambäck and Sikdar, 2017; Park and Fung, 2017; Davidson et al., 2017).

3.2. Motivation and contribution

The hate speech detection on social media is relatively new but an important topic. There are very few publicly available corpora mostly dedicated to English. Even for English, less than 10 resources are publicly available. More recently, researchers have presented work in other languages including German, Italian, Arabic. However, most of the work focuses on detecting a general hate speech not against a specific community. In Arabic, only 5 research studies are presented in the literature which are mainly focused on Twitter. This paper focuses on Youtube which is the second biggest social media platform, after Facebook, with 1.8 billion users (Kallas, 2017; Alakrot et al., 2018). The major contributions of this study are: Development of a novel hate speech corpus against women containing MSA and Algerian dialect, written in Arabic, Arabizi, French, and English. The corpus constitutes 5,000 manually annotated comments. For corpus validation, three deep learning algorithms (CNN, LSTM, and bi-LSTM) are used for hate speech classification. For feature extraction, algorithms such as word2vec, FasText, etc., are used.

4. Methodology

4.1. Dataset creation

4.1.1. Data collection

Youtube comments related to videos about women are used. Feminine adjective such as: جميلة meaning beautiful, جَايحة meaning stupid or كلبة meaning a dog are targeted. A video on Youtube is recognised by a unique identifier (*video_id*). For example the video having an id equal to *"TJ2WfhfbvZA"* handling a radio emission about unfaithful women and the video having an id equal to *"_VimCUVXwaQ"* gives advices to women for becoming beautiful. Three annotators, manually review the obtained video from the keyword and manually selected 335 *video_id*. We used Youtube Data API[4] and a python script to automatically extract comments of each *video_id* and their replies. At the end, we were able to collect *373,984* comments extracted for the period between February and March 2019, we call this corpus *Corpus_Youtube_women*.

4.1.2. Data annotation

For the annotation, we randomly select 5,000 comments. The annotation was done by three annotators, natives speaker of Arabic and its dialects. The annotators were separated and they had one week for manually annotated the selected comments using two labels, 1 (for hate) and 0 (for non-hate). The following points illustrate the main aspects figuring in the annotators guideline:

- The annotators should classify each comments containing injuries, hate, abusive or vulgar or offensive language against women as a comment containing hate.
- The annotators should be as objective as they can. Even if they approve the comment, they should consider it as containing hate speech is it is offensive against women.
- For having a system dealing with all type of comments, the annotators were asked to annotate all

[4]https://developers.google.com/youtube/v3/

the 5,000 comments, even if the comment speak about football or something not related to women at all. However they asked to annotate this comment with 0 and to add the label w (meaning without interest).

- When the annotators are facing a situation where they really doubt about the right label, they were asked to put the label p (for problem) rather than putting a label with which they are not convinced.

At the beginning of the annotation process, we received lots questions such as: 1) Have the hate have to be addressed to women, how to classify a message containing hate regarding men? 2) Have the hate comments absolutely contains terms indicating hate or have the annotators to handle irony?, etc. For the first question, we precise that the comments have to be addressed to women. Any others comment have to be labelled with 0 For the second question, we asked the annotators to also consider the irony and sarcasm.

After completion of the annotation process, we concentrate on the comments obtaining the same labels from all annotators. Then, we constructed two dataset. The first one (*Corpus_1*) contains 3,798 comments which are annotated with the same labels (0 or 1) from the three annotators. Among this comments 792 (which represent 20.85%) are annotated as hateful and 3006 as non-hateful. Hence, this corpus is very unbalanced. The second one (*Corpus_2*) represents the balanced version of (*Corpus_1*). For constructing this corpus, we randomly picked up 1,006 comments labelled as non-hateful and we picked up all the comments annotated as hateful. Then, we constructed a balanced corpus containing 1,798 comments.

4.2. Hate speech detection

4.2.1. Features extraction

We use two different algorithm for features extraction which are, Word2vec (Mikolov et al., 2013) and FasText (Joulin et al., 2016). We use Word2vec with classic methods and we use FasText with Deep learning methods. Word2vec describes two architectures for computing continuous vectors representations, the Skip-Gram (SG) and Continuous Bag-Of-Words (CBOW). The former predicts the context-words from a given source word, while the latter does the inverse and predicts a word given its context window (Mikolov et al., 2013). As for Word2vec, Fastext models is also based on either the skip-gram (SG) or the continuous bag-of-words (CBOW) architectures. The key difference between FastText and Word2Vec is the use of n-grams. Word2Vec learns vectors only for complete words found in the training corpus. FastText learns vectors for the n-grams that are found within each word, as well as each complete word (Joulin et al., 2016). In this work we rely on both representations of word2vec and fasText (i.e SG and CBOW).

For Word2vec model, we used the Gensim toolkit[5]. For fasText, we use the fasText library proposed by Facebook on Github[6]. For both Word2vec/fasText, we use a context of 10 words to produce representations for both CBOW and SG of length 300. We trained the Word2vec/fasText models on the corpus *Corpus_Youtube_women*

4.2.2. Classification

For comparing the results, we use both classification methods, classic and deep learning based. For classic method, we use five classification Algorithms such as: GaussianNB (GNB), LogisticRegression (LR), RandomForeset (RF), SGDClassifier (SGD, with loss='log' and penalty='l1') and LinearSVC (LSVC with C='1e1'). For their implementation phase, we were inspired by the classification algorithm proposed by Altowayan et al. (Altowayan and Tao, 2016). For the deep learning classification we use three models CNN, LSTM and Bi-LSTM. For each model, we use six layers. The first layer is a randomly-initialized word embedding layer that turns words in sentences into a feature map. The weights of embedding_matrix are calculated using fasText (with both SG and CBOW implementation). This layer is followed by a CNN/ LSTM/BiLSTM layer that scans the feature map (depending on the model that we defined). These layers are used with 300 filters and a width of 7, which means that each filter is trained to detect a certain pattern in a 7-gram window of words. Global maxpooling is applied to the output generated by CNN/LSTM/BiLSTM layer to take the maximum score of each pattern. The main function of the pooling layer is to reduce the dimensionality of the CNN/LSTM/BiLSTM representations by down-sampling the output and keeping the maximum value. For reducing over-fitting by preventing complex co-adaptations on training data, a Dropout layer with a probability equal to 0.5 is added. The obtained scores are then feeded to a single feed-forward (fully-connected) layer with Relu activation. Finally, the output of that layer goes through a sigmoid layer that predicts the output classes. For all the models we used Adam optimizers with epoch 100 and an early_stopping parameter for stopping the iteration in the absence of improvements.

5. Experimentation and Results

5.1. Experimental results

Table 1 presents the results obtained on *Corpus_1* and *Corpus_2*. For showing the impact of balanced/unbalanced corpus, we present the different results related to the detection of Hateful/non hateful detection separately. It can be seen from Table 1 that the F1-score obtained on unbalanced corpus (*Corpus_1*, up to 86%) are slightly better than those obtained on the balanced corpus (*Corpus_1*, up to 85%). However only

[5]https://radimrehurek.com/gensim/models/word2vec.html
[6]https://github.com/facebookresearch/fastText

Table 1: Classification results on Corpus_1

Corpus	Models	Type	ML Alg	Hateful			Non-hateful			Average		
				P	R	F1	P	R	F1	P	R	F1
			GNB	0.32	0.80	0.46	0.91	0.56	0.69	0.79	0.61	0.64
			LR	0.70	0.19	0.30	0.82	0.98	0.89	0.80	0.81	0.77
		SG	RF	0.69	0.33	0.44	0.84	0.96	0.90	0.81	0.83	0.80
			SGD	0.81	0.13	0.23	0.81	0.99	0.89	0.81	0.81	0.75
			LSVC	0.70	0.41	0.52	0.86	0.95	0.90	0.83	0.83	0.82
	Word2vec											
			GNB	0.30	0.82	0.44	0.91	0.48	0.63	0.78	0.55	0.59
			LR	0.75	0.04	0.07	0.79	1.00	0.88	0.79	0.79	0.71
		CBOW	RF	0.50	0.17	0.26	0.81	0.95	0.88	0.75	0.79	0.75
			SGD	0.67	0.04	0.07	0.79	0.99	0.88	0.77	0.79	0.71
Corpus_1			LSVC	0.57	0.15	0.24	0.81	0.97	0.88	0.76	0.80	0.75
			CNN	0.77	0.56	**0.65**	0.89	0.96	**0.92**	0.87	0.87	**0.86**
		SG	LSTM	0.82	0.45	0.58	0.87	0.97	0.92	0.86	0.86	0.85
			Bi-LSTM	0.89	0.36	0.51	0.85	0.99	0.91	0.86	0.86	0.83
	FasText											
			CNN	0.71	0.46	0.56	0.87	0.95	0.91	0.84	0.85	0.83
		CBOW	LSTM	0.67	0.53	0.59	0.88	0.93	0.90	0.83	0.84	0.84
			Bi-LSTM	0.56	0.61	0.59	0.89	0.87	0.88	0.82	0.82	0.82
			GNB	0.63	0.82	0.71	0.83	0.63	0.71	0.74	0.71	0.71
			LR	0.79	0.75	0.77	0.82	0.85	0.84	0.81	0.81	0.81
		SG	RF	0.81	0.62	0.71	0.76	0.89	0.82	0.78	0.78	0.77
			SGD	0.72	0.85	0.78	0.87	0.75	0.81	0.81	0.79	0.79
			LSVC	0.79	0.74	0.76	0.81	0.85	0.83	0.80	0.80	0.80
	Word2vec											
			GNB	0.54	0.85	0.66	0.80	0.45	0.58	0.69	0.62	0.61
			LR	0.72	0.58	0.65	0.73	0.83	0.77	0.72	0.72	0.72
		CBOW	RF	0.73	0.63	0.68	0.75	0.82	0.78	0.74	0.74	0.74
			SGD	0.77	0.57	0.65	0.73	0.87	0.79	0.74	0.74	0.73
Corpus_2			LSVC	0.75	0.70	0.72	0.79	0.82	0.80	0.77	0.77	0.77
			CNN	0.86	0.69	0.77	0.80	0.92	0.85	0.83	0.82	0.82
		SG	LSTM	0.93	0.60	0.73	0.76	0.97	0.85	0.83	0.81	0.80
			Bi-LSTM	0.85	0.81	**0.83**	0.86	0.89	**0.88**	0.85	0.86	**0.85**
	FasText											
			CNN	0.81	0.62	0.70	0.76	0.89	0.82	0.78	0.77	0.77
		SG	LSTM	0.94	0.57	0.71	0.75	0.97	0.85	0.83	0.80	0.79
			Bi-LSTM	0.73	0.82	0.77	0.85	0.77	0.81	0.80	0.79	0.79

65% of hateful comment were correctly classified using *Corpus_1*, where 83% are correctly classified using *Corpus_2*. Deep learning classifiers (CNN, Bi-LSTM) associated to SG model of fasText outperformed other classifiers for both corpus (1 and 2). In addition SG model outperformed CBOW model for both corpus and and for all the used classifiers. It also can be observed that deep learning classifiers are more appropriate with unbalanced data (F1-score up to 65%) where the classic classifiers (GNB, LR, ect) are able to correctly classify only 52%.

5.2. Discussion and Analysis

The presented results are pretty good but they could be improved by integrating some pre-treatments. The first one is related to Arabizi transliteration. As Arabic people used both scripts Arabic and Arabizi. Handling them together or classifying Arabizi without calling the transliteration step could give wrong results. We previously showed that the transliteration consequently improved the results of sentiment analysis (Guellil et al., 2018a). We previously present a transliteration based on rules-based approach (Guellil et al., 2018a; Guellil et al., 2018c) but we conclude that a corpus based approach would certainly improve the results. Hence, we

plan to propose a corpus-based approach for transliteration and apply this approach on the annotated corpus for having one script used for Arabic language. In addition to scripts, Arabic people also use other languages to express their opinions in social media, such as French or English. However, the proportion of these languages is not really important comparing to the proportion of Arabic and Arabizi. In the context of this study, we handle all the languages in the same corpus. However, a language identification step would consequently improve the results. Hence, as an improvement to this work, we plan to propose an identification approach between Arabizi, French and English (because they share the same script).

6. Conclusion

Hate speech detection is a research area attracting the research community interest more and more. Different studies have been proposed and most of them are quietly recent (during 2016 and 2019). The purpose of this studies is mitigated between the detection of hate speech in general and hate speech targeting a special community or a special group. In this context, the principal aim of this paper is to detect hate speech against women in Arabic community on social media. We automatically collected data related to women from Youtube. Afterwards, we randomly select 5,000 comments and give them to three annotators in order to labelled them as hateful or non-hateful. However, for increasing the precision, we concentrate on the portion of the corpus were all the annotators were agree. It allows us to construct a corpus containing 3,798 comments (where 3.006 are non-hateful and 792 are hateful). We also constructed a balanced corpus containing 1,798 comment randomly picked up from the aforementioned one. For validating the constructed corpus, we used different machine learning algorithm such as LSVC, GNB, SGD, etc and deep learning one such as CNN? LSTM, etc. However, The exeperimental results showed that the deep learning classifiers (especially CNN, Bi-LSTM) outperform the other classifiers by respectively achieving an F1-score up to 86%.

For improving this work we plan to integrate a transliteration system for transforming Arabizi to Arabic. We also plan to identify the different language before proceeding to the classification. Finally, we also plan to automatically increase the training corpus.

7. Acknowledgements

We would like to thank the Edinburgh Futures Institute (EFI)[7] for funding the fees related to the presentation of this paper. The purpose of the Edinburgh Futures Institute (EFI) is to pursue knowledge and understanding that supports the navigation of complex futures.

[7]https://efi.ed.ac.uk/

8. Bibliographical References

Abozinadah, E. A., Mbaziira, A. V., and Jones, J. (2015). Detection of abusive accounts with arabic tweets. *International Journal of Knowledge Engineering*, 1(2):113–119.

Al-Hassan, A. and Al-Dossari, H. (2019). Detection of hate speech in social networks: A survey on multilingual corpus. *Computer Science & Information Technology (CS & IT)*, 9(2):83.

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.

Alakrot, A., Murray, L., and Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Albadi, N., Kurdi, M., and Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Alfina, I., Mulia, R., Fanany, M. I., and Ekanata, Y. (2017). Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Altowayan, A. A. and Tao, L. (2016). Word embeddings for arabic sentiment analysis. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 3820–3825. IEEE.

Andrusyak, B., Rimel, M., and Kern, R. (2018). Detection of abusive speech for mixed sociolects of russian and ukrainian languages. *RASLAN 2018 Recent Advances in Slavonic Natural Language Processing*, page 77.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 93–103.

Boudad, N., Faizi, R., Thami, R. O. H., and Chiheb, R. (2017). Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*.

Burnap, P. and Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. pages 1–18.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior*.

Dadvar, M. and Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *arXiv preprint arXiv:1812.08046*.

Darwish, K. (2014). Arabizi detection and conversion to arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

De Smedt, T., De Pauw, G., and Van Ostaeyen, P. (2018). Automatic detection of online jihadist hate speech. *arXiv preprint arXiv:1803.04596*.

Del Vigna12, F., Cimino23, A., Dell Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.

Eisele, A. and Chen, Y. (2010). Multiun: A multilingual corpus from united nation documents. In *LREC*.

Farghaly, A. and Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):14.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Gorrell, G., Greenwood, M. A., Roberts, I., Maynard, D., and Bontcheva, K. (2018). Twits, twats and twaddle: Trends in online abuse towards uk politicians. In *Twelfth International AAAI Conference on Web and Social Media*.

Guellil, I. and Azouaou, F. (2016). Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect. In *Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on*, pages 724–731. IEEE.

Guellil, I., Azouaou, F., and Abbas, M. (2017a). Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31 (2017)*.

Guellil, I., Azouaou, F., Abbas, M., and Fatiha, S. (2017b). Arabizi transliteration of algerian arabic dialect into modern standard arabic. In *Social MT 2017: First workshop on Social Media and User Generated Content Machine Translation (co-located with EAMT 2017)*.

Guellil, I., Azouaou, F., Saâdane, H., and Semmar, N. (2017c). Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien.

Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A.-e., and Hussain, A. (2018a). Arabizi sentiment analysis based on transliteration and automatic corpus annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 335–341.

Guellil, I., Adeel, A., Azouaou, F., and Hussain, A. (2018b). Sentialg: Automated corpus annotation for algerian sentiment analysis. In *9th International Conference on Brain Inspired Cognitive Systems(BICS 2018)*.

Guellil, I., Azouaou, F., Benali, F., Hachani, a.-e., and Saadane, H. (2018c). Approche hybride pour la translitération de l'arabizi algérien : une étude préliminaire. In *Conference: 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), May 2018, Rennes, FranceAt: Rennes, France*. https://www.researchgate.net/publication ....

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., and Nouvel, D. (2019). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*.

Guellil, I., Azouaou, F., Benali, F., and Ala-Eddine, H. (2021). One: Toward one model, one algorithm, one corpus dedicated to sentiment analysis of arabic/arabizi and its dialects. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 236–249.

Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Haidar, B., Chamoun, M., and Serhrouchni, A. (2017). A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Advances in Science, Technology and Engineering Systems Journal*, 2(6):275–284.

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*.

Imane, G., Kareem, D., and Faical, A. (2019). A set of parameters for automatically annotating a sentiment arabic corpus. *International Journal of Web Information Systems*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kallas, P. (2017). Top 15 most popular social networking sites and apps. *Consultado em Setembro*, 20:2017.

Köffer, S., Riehle, D. M., Höhenberger, S., and Becker, J. (2018). Discussing the value of automatic hate speech detection in online debates. *Multikonferenz*

Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana, Germany.

Kshirsagar, R., Cukuvac, T., McKeown, K., and McGregor, S. (2018). Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644*.

Madisetty, S. and Desarkar, M. S. (2018). Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018). Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Puiu, A.-B. and Brabete, A.-O. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media. *arXiv preprint arXiv:1903.00665*.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Risch, J., Krebs, E., Löser, A., Riese, A., and Krestel, R. (2018). Fine-grained classification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*.

Saha, P., Mathew, B., Goyal, P., and Mukherjee, A. (2018). Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Sharaf, A.-B. M. and Atwell, E. (2012a). Qurana: Corpus of the quran annotated with pronominal anaphora. In *LREC*. Citeseer.

Sharaf, A.-B. M. and Atwell, E. (2012b). Qursim: A corpus for evaluation of relatedness in short texts. In *LREC*, pages 2295–2302.

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.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). Inducing a lexicon of abusive words–a feature-based approach. pages 1046–1056.

Zhang, Z. and Luo, L. (2018). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, (Preprint):1–21.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In *LREC*.