iCompass at SemEval-2020 Task 12: From a Syntax-ignorant N-gram Embeddings Model to a Deep Bidirectional Language Model

Abir Messaoudi
iCompass / Tunisia
abir@icompass.com

Hatem Haddad
iCompass / Tunisia
hatem@icompass.tn

Moez Ben Hajhmida
iCompass / Tunisia
moez@icompass.tn

Abstract

We describe our submitted system to the SemEval 2020. We tackled Task 12 entitled "Multilingual Offensive Language Identification in Social Media", specifically subtask 4A-Arabic. We propose three Arabic offensive language identification models: Tw-StAR, BERT and BERT+BiLSTM. Two Arabic abusive/hate datasets were added to the training dataset: L-HSAB and T-HSAB. The final submission was chosen based on the best performances which was achieved by the BERT+BiLSTM model.

1 Introduction

With the freedom of expression privilege granted after the Arab countries revolution, sensitive topics such as religion and politics have become popular and widely discussed across social media platforms. However, on the down side, offensive language spreads easily. Indeed, recent events; like Persian's gulf crisis, the parliamentary-presidential elections held in Tunisia or a football game between two Arabic clubs; caused intensive debates, most of them took place on social media networks leading to a high emergence of offensive speech. This evokes the need for tools to identify such online offensive language content.

Analyzing Arabic offensive language is significantly challenging due to the complex nature and morphology of the Arabic language. Furthermore, Arabic language in Social Media is mostly informal and written in Arabic dialects.

We describe our participation in SemEval 2020 Task 12 entitled "Multilingual Offensive Language Identification in Social Media" Zampieri et al. (2020), specifically Task 12 4A-Arabic: "Arabic Offensive Language Identification in Social Media" Mubarak et al. (2020) under the team name "iCompass". The task requires to distinguish between offensive (containing any form of non-acceptable language (profanity)) and non-offensive posts.

The remainder of the paper is organized as follows: in Section 2, we introduce L-HSAB and T-HSAB datasets. In Section 3, we describe the preprocessing step. Section 4 introduces the learning strategies and datasets used in the presented models. Results are reviewed and discussed in Section 5 while Section 6 concludes the study.

2 Arabic offensive language datasets

The proposed dataset for the SemEval task 12 subtask A-Arabic is composed mainly by Egyptian dialect comments and few other dialects like Libyan, Sudanese, Syrian, etc. Some comments, after the preprocessing phase, in different dialects are given in Table 1[1].

Consequently, in addition to the dataset proposed by SemEval Task 12 4A-Arabic, we used two other Arabic offensive language datasets : L-HSAB and T-HSAB.

L-HSAB (a Twitter dataset about abusive speech (AS) and hate speech (HS)) is introduced in Mulki et al. (2019a) as a benchmark dataset for automatic detection of online Levantine offensive language. The

[1]Comments numbers 700, 1088, 5 and 2439 of the training set.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1978–1982
Barcelona, Spain (Online), December 12, 2020.

Dialect	Comment	English translation
Egyptian	ايه يا شتا يا حبى مش خلاص كدة.	Hey my love, isn't enough like that..
North African	خليك غادي أنبح لعند بكرا صار تبي ...	Stay like that keep nagging till tomorrow..
Levantine	يا بكون بحياتك الأهم يا إما ما بدي أكون	I shall be at the highest priority in your life, or I don't wanna be in your life.
Golf arabic	يا حســـــايف قولتي لك يا حبيبي يوم قلبي جـــــــاهلٌ بك ما درى	Oh damage, I called you my beloved before I know the real you.

Table 1: Examples of comments in different dialects.

dataset is composed of 5.8K tweets, manually annotated as Normal, Abusive and Hate. The high obtained values of agreement without chance correction and inter-annotator agreement indicated the reliability of the dataset. The inter-rater agreement metric denoted by Krippendorff's alpha (α) was 76.5% and indicated the consistency of the annotations.

T-HSAB was constructed out of Tunisian comments harvested from different social media platforms and introduced in Haddad et al. (2019) as a benchmark dataset for automatic detection of online Tunisian offensive language. The dataset is composed of 6K tweets, manually annotated as Normal, Abusive and Hate. It has high values of agreement without chance correction and inter-annotator agreement and Krippendorff's alpha (α) value of 75%.

3 Data Preprocessing

Data was preprocessed by cleaning the tweets from the social media-inherited symbols such as (Rt,<LF> and @), URLs, Usernames, dates, retweets, symbols, punctuations, emojis and non-Arabic characters, in order to remove noise and to get the Arabic text only.
Table 2 shows an example of the comment number 2372 of the training set, before and after preprocessing.

Before	RT @USER: حبيباتِ العرقْ ... أنت يا شعبي \ أنت يا صديقي \ أما أنت <LF> ::...
After	أما أنت أنت يا صديقي أنت يا شعبي حبيبات العرق

Table 2: Example of a comment before and after preprocessing phase.

4 Used models and Learning Strategies

In this section, we describe the learning strategies and the different architectures used. The mechanism of each strategy is briefly reviewed. To accomplish this mission, we have used three classification approaches.

4.1 Tw-StAR

Tw-StAR is a syntax-ignorant n-gram embeddings model used in sentiment analysis of several Arabic dialects Mulki et al. (2019b). Tw-StAR's embeddings are composed and learned using an unordered composition function and a shallow neural model. The model performed state of art results at Semeval-2017 Task 4: "Sentiment Analysis in Twitter" Mulki et al. (2017) including Arabic language and SemEval-2018 Task 1: "Affect in Tweets", subtask Ec "Detecting Emotions (multi-label classification)" Mulki et al. (2018).

In addition, we have used an offensive lexicon extracted from L-HSAB and T-HSAB datasets. The offensive lexicon is composed by offensive words in Arabic dialects such as **خائن، جزمة، عاهرة، ساقط، وسخة** that mean Thief, traitor... We have combined the predictions of our classifier with the results of the lexicon-based method to have a hybrid solution. This mixed solution have decreased the results since it can neither identify negations nor detect non-offensive comments that contain offensive words, example in table 3.

Arabic	... ده كان انسان طيب ومتقي مش حرامي وخاين زي ماكانو يقولو
English	... this was a good and fearful person, not a thief and a traitor as they said

Table 3: Example of a non offensive comment that contains offensive words.

Tw-StAR+Lexicon approach has classified this comment as offensive because it contains حرامي and خاين, which means Thief and Traitor, that appear in the lexicon of offensive words and confuse the prediction results of this type of sentences. This approach decreased the performance results as given in Table 5.

4.2 BERT

Context-free models such as word2vec or GloVe generate a single word embedding representation for each word in the vocabulary. Differently, contextual models generate a representation of each word based on the other words in the sentence. For this reason, we selected the Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. (2019) as a contextual language model in its light multilingual version as an embedding technique.

BERT is a deep bidirectional language model, pretrained on large corpora (BooksCorpus: 800M words, and Wikipedia: 2,500M words), that can be fine-tuned to solve many NLP tasks such as named entity recognition (NER), question answering (QA) and text classification Devlin et al. (2019).

We used the BERT base multilingual cased model, with 12 transformer layers, 12 attention heads and 110M parameters. We used the already pretrained model and trained the classifier to predict the probabilities of the labels (OFF or NOT) tuning different values of hyper parameters in order to have the best performances. Values for fine-tuning are stated in Table 4.

4.3 BERT + BiLSTM

The dataset was tokenized using the BERT tokenizer mapping Arabic words to their indexes. BERT embedding matrix was used at the embedding layer level. Then, BILSTM model was used as classifier in order to predict label's probabilities.

	Tw-StAR	**BERT**	**BERT+BILSTM**
Batch size	128	8	16
Sequence length	-	128	128
Epochs	6	3	6
N-grams (N)	8	-	-

Table 4: Hyper parameters used during training phase.

5 Results and Discussion

Three datasets are provided by SemEval 2020 Task 12 subtask 4A-Arabic: TRAIN (7000 comments) for training models, DEV (1000 comments) for tuning models, and TEST (2000 comments) for the official evaluation.

Data was preprocessed using regular expressions recognition and regular expressions substitution provided by the re Python module[2]. Having the data preprocessed and the features extracted, training dataset was splitted into 80% for training and 10% for cross-validation. We have trained our Tw-StAR, BERT, and BiLSTM models on the DEV set.

Table 5 lists the results of the four classification architectures: Tw-StAR, Tw-StAR+Lexicon, BERT, and BERT+BiLSTM. As a result, a slight improvement was achieved by BERT when compared to the Tw-StAR baseline and BiLSTM by achieving the best performances with a macro F-score (Macro F1) of 0.825 and an accuracy of 0.898.

[2] https://docs.python.org/3/library/re.html

Model	Accuracy	Micro F1	Macro F1
Tw-StAR	0.813	0.813	0.331
Tw-StAR + Lexicon	0.818	0.818	0.332
BERT	0.821	0.821	0.791
BERT+BiLSTM	**0.898**	**0.898**	**0.825**

Table 5: Results on the SemEval Devset.

In order to improve the performances, we used T-HSAB and L-HSAB datasets for training in addition to the SemEval 2020 Task 12 subtask 4A-Arabic provided training dataset. Characteristics of each dataset are given in Table 6.

Dataset	Dialect	NOT	OFF	Total
L-HSAB	Levantine	3650	2196	5846
T-HSAB	Tunisian	3835	2204	6039
SemEval 2020	mostly Egyptian	5666	1334	7000
Total	–	13151	5734	18885

Table 6: Datasets used during the training phase.

Considering the results stated in Table 5 and Table 7, the supervised learning-based model with BiLSTM algorithm achieved the best average F-score (micro and marco) performances compared to the BERT base and the Tw-StAR model. Hence, BERT+BiLSTM was used to provide the TEST set classification results for the final submission.

Training dataset	Accuracy	Micro F1	Macro F1
SemEval	0.885	0.885	0.784
SemEval+L-HSAB	**0.900**	**0.900**	0.794
SemEval+L-HSAB + T-HSAB	0.898	0.898	**0.825**

Table 7: Results on SemEval Devset according to the training dataset.

Table 8 reviews the official results of iCompass system against the top three ranked systems.

6 Conclusion and Future work

Three classification architectures were used to identify offensive language Arabic tweets. The best F-score macro results were obtained by the BiLSTM classification architecture using BERT base multilingual word embedding which was selected for the final submission.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-HSAB: A tunisian hate speech and abusive dataset. In Kamel Smaïli, editor, *Arabic Language Processing: From Theory to Practice - 7th International Conference, ICALP 2019, Nancy, France, October 16-17, 2019, Proceedings*, volume 1108 of *Communications in Computer and Information Science*, pages 251–263. Springer.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Team	Rank	macro F1-score
ALAMI Hamza	1	0,9017
alt	2	0,9015
Galileo	3	0,8988
iCompass	14	0,8519

Table 8: Ranking and results on SemEval Test set.

Hala Mulki, Hatem Haddad, Mourad Gridach, and Ismail Babaoğlu. 2017. Tw-star at semeval-2017 task 4: Sentiment classification of arabic tweets. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 664–669, Vancouver, Canada, August.

Hala Mulki, Chedi Bechikh Ali, Hatem Haddad, and Ismail Babaoğlu. 2018. Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 167–171, New Orleans, Louisiana, June.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019a. L-HSAB: A Levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy, August. Association for Computational Linguistics.

Hala Mulki, Hatem Haddad, Mourad Gridach, and Ismail Babaoğlu. 2019b. Syntax-ignorant n-gram embeddings for sentiment analysis of Arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 30–39, Florence, Italy, August.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.