Leveraging Offensive Language for Sarcasm and Sentiment Detection in Arabic

Fatemah Husain
Kuwait University / State of Kuwait
f.husain@ku.edu.kw

Ozlem Uzuner
George Mason University / USA
ouzuner@gmu.edu

Abstract

Sarcasm detection is one of the top challenging tasks in text classification, particularly for informal Arabic with high syntactic and semantic ambiguity. We propose two systems that harness knowledge from multiple tasks to improve the performance of the classifier. This paper presents the systems used in our participation to the two sub-tasks of the Sixth Arabic Natural Language Processing Workshop (WANLP); Sarcasm Detection and Sentiment Analysis. Our methodology is driven by the hypothesis that tweets with negative sentiment and tweets with sarcasm content are more likely to have offensive content, thus, fine-tuning the classification model using large corpus of offensive language, supports the learning process of the model to effectively detect sentiment and sarcasm contents. Results demonstrate the effectiveness of our approach for sarcasm detection sub-task over sentiment analysis sub-task.

1 Introduction

Current trend in sentiment analysis research is moving toward the special sub-field of sarcasm detection. It is becoming one of the most challenging and relevant task for the sentiment analysis community. The complexity of detecting sarcastic content is attributed to multiple factors including context understanding, cultural aspects, and personal traits (Oprea and Magdy, 2019).

This paper describes two systems that have been submitted to the shared sub-tasks of Sarcasm Detection and Sentiment Analysis at the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)(Abu Farha et al., 2021). The approach adopted in developing the systems is inspired by the findings from offensive language studies. Previous researchers highlight sarcastic content as one of the main causes of confusion for offensive language detection systems (Djandji et al., 2020; Keleg et al., 2020). Additionally, multiple studies applied sentiment features into the task of offensive language detection and report significant gain in performance (Haidar et al., 2017; Elmadany et al., 2020; Abu Farha and Magdy, 2020b). All of these findings demonstrate the relationship among sentiment analysis, sarcasm detection, and offensive language detection tasks.

Our main methodology is based on transfer learning that is performed by joint fine-tuning over the concatenated tasks; offensive language, sarcasm detection, sentiment analysis, and transfer corpora. Thus, the contextualized word embedding is learned based on the entire fine-tuning corpus. The main goal from applying this system pipeline is to examine the impact of offensive language linguistic features on sarcastic language and sentiment content.

2 Related Work

Abu Farha and Magdy (2020a) introduce the first version of ArSarcasm dataset, which includes labels for sentiment analysis, dialect identification, and sarcasm detection tasks with a total of 10,547 tweets, of which only 1,682 (16%) are sarcastic tweets. The ArSarcasm dataset is still a relatively new dataset and few researchers apply it in their studies. In (Abu Farha and Magdy, 2020a), the authors use Mazajak word embedding (Farha and Magdy, 2019) and develop a Bi-LSTM model. The results record very low performance; 62% precision, 38% recall, and F1-score of 0.46. Authors highlight the challenging in developing high performance system for detecting sarcasm because of the contextual and cultural aspects of sarcasm content. Abdul-Mageed et al. (2021) explore multiple BERT models for multiple text classification tasks in Arabic. Results for sentiment analysis task using

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 364–369
Kyiv, Ukraine (Virtual), April 19, 2021.

the ArSarcasm dataset record the highest F1 score of 71.50 when applied with MARBERT model (Abdul-Mageed et al., 2021). The same study also reports results for Sarcasm detection task using the ArSarcasm dataset. The highest achieved F1 score is 76.30 by MARBERT.

Multiple researchers from offensive language detection studies report findings from their experiments that highlight the relationship between sarcasm detection and offensive language detection. For example in (Djandji et al., 2020), the authors applied a multitask learning and multilabel classification approach with AraBERT model (Antoun et al., 2020), for hate speech detection and offensive language detection. The findings from their error analysis showed some mislabeled hate speech tweets that are mostly related to mockery, sarcasm, or mentioning other offensive and hateful statements within tweets. Keleg et al. (2020) also report similar findings. In (Keleg et al., 2020), multiple classification models for offensive language detection were explored, and the results from the error analysis highlight some issues that confused the classifiers including the use of sarcastic speech to quote scenes from popular movies.

Furthermore, various offensive language detection studies apply sentiment features in their models and report positive effects in system performance. For instance, Haidar et al. (2017) deploy a cyberbullying detection system that consists of SentiStrength[1] features related to sentiment polarity of Twitter users to train a classifier. Results report higher performance when using the system with SentiStrength features from the model that does not apply sentiment features. Elmadany et al. (2020) develop an offensive language detection system based on assuming a correlation between negative sentiment and offensive language. They use AraNet(Abdul-Mageed et al., 2020) to augment the imbalanced dataset by adding negative tweets and develop M-BERT[2]-based classifiers. Results show higher performance of the model that applies negative sentiment augmented dataset over the others that do not consider the sentiment augmentation approach. Additionally, Abu Farha and Magdy (2020b) explored various classifiers using different multitask learning settings across offensive language, hate speech, and sentiment analysis. Result demonstrates higher performance for the

[1] http://sentistrength.wlv.ac.uk/

[2] https://github.com/google-research/bert/blob/master/multilingual.md

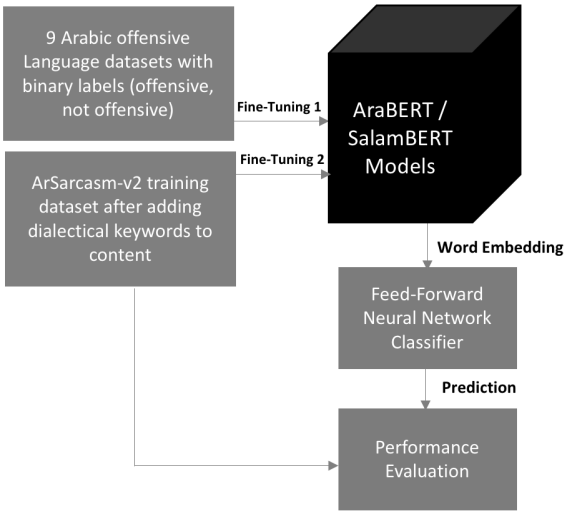

Figure 1: System Pipeline

model that is trained on offensive language, hate speech, and sentiment, assuming all offensive and hate speech tweets are negative sentiment and others are positive sentiment.

3 Methodology

Transfer learning is performed by joint training over the concatenated task and transfer corpora, and subwords are learned over the concatenation of both corpora; offensive language and sarcasm detection or sentiment analysis. The overall proposed system architecture is described in Fig.1. We submitted two separate submissions that were independently trained and tested for each sub-task.

3.1 Tasks and Datasets

The main dataset of our study is the ArSarcasm-v2 dataset (Abu Farha et al., 2021), which includes 15,548 tweets with three labels assigned to each tweet; sentiment, sarcasm, and dialect, and consists of two parts; 12,548 tweets in training dataset and 5,000 tweets in testing dataset. During our experimental studies, the ArSarcasm-v2 dataset (Abu Farha et al., 2021) is randomly classified into 80% training dataset and 20% development dataset to evaluate our model. The training part of the dataset consists of the following sentiment tweets: 4,623 neutral, 3,672 negative, and 1,743 positive, and the following sarcasm tweets: 1,749 sarcasm and 8,289 not sarcasm. While the development part of the dataset consists of 1,124 neutral, 949 negative, and 437 positive sentiment tweets, and 419 sarcasm and 2,091 not sarcasm tweets. Labels for the testing dataset are not available during the

Model	Class	Precision	Recall	Macro-F1
Baseline	Not Sarcasm	0.91	0.94	0.92
AraBERT	Sarcasm	0.63	0.52	0.57
	Average	**0.77**	0.73	**0.75**
Main	Not Sarcasm	0.90	0.92	0.91
AraBERT	Sarcasm	0.56	0.54	0.54
	Average	0.73	0.72	0.73
Baseline	Not Sarcasm	0.91	0.93	0.92
SalamBERT	Sarcasm	0.59	0.52	0.56
	Average	0.75	0.72	0.74
Main	Not Sarcasm	0.91	0.92	0.92
SalamBERT	Sarcasm	0.59	0.57	0.58
	Average	0.75	**0.74**	**0.75**

Table 1: Performance Evaluation Results of the Development Dataset for the Sarcasm Detection Sub-Task.

time of writing this paper, however, results are announced by the shared task organizers and included in the results section of this paper. The full details of the dataset are available in the task guidelines (Abu Farha et al., 2021).

We also use 9 Arabic offensive language datasets in developing the proposed classification system. The datasets consist of Aljazeera.net Deleted Comments (Mubarak et al., 2017), Egyptian Tweets (Mubarak et al., 2017), YouTube Comments (Alakrot et al., 2018a), Religious Hate Speech (Albadi et al., 2018), L-HSAB (Mulki et al., 2019), T-HSAB (Haddad et al., 2019), MPOLD (Chowdhury et al., 2020), OSACT4 (Mubarak et al., 2020), and the Multi-Platform Hate Speech Dataset (Omar et al., 2020). All datasets were used without any changes in content, no filtering or preprocessing were performed. However, to maintain consistency among all datasets, the labels were changed for some datasets. Only binary classes are applied; offensive or not offensive. Thus, we convert different types of offensive languages to offensive class. For example, the L-HSAB and T-HSAB datasets differentiate between hate and abusive languages classes; which were both converted to offensive class.

3.2 Preprocessing

Only one preprocessing procedure is conducted over ArSarcasm-v2 dataset (Abu Farha et al., 2021), which consists of adding a keyword token to the end of the tweet to refer to the dialect of the tweet. Thus, if the record in ArSarcasm-v2 Dataset (Abu Farha et al., 2021) shows a label for the dialect as "gulf", then the keyword ‘خليجي/ Gulfian’ is added as the last token in the tweet corresponding to that record. Similarly to the other dialects; ‘msa’, ‘egypt’, ‘levant’, and ‘magreb’, which were assigned the following keyword tokens respectively, ‘عربي/ Arabic’, ‘مصري/ Egyptian’, ‘شامي/ Levantine’, and ‘مغربي/ Moroccan’.This preprocessing procedures is based on the assumptions that adding a word of similar meaning to the dialect to create a semantic relationship with the dialect of the tweet can enrich the process of contextual understanding for the classification model.

3.3 Classification Model

The experiment depends mainly on AraBERT model (aubmindlab/bert-base-arabert)(Antoun et al., 2020). The architecture of AraBERT model is adopted from HuggingFace[3] library with AutoModelForSequenceClassification module. The pooled output from AraBERT encoder is used with a simple Feed Forward Neural Network layer to build the classification model. Experimental settings include maximum sample length of 256, patch size of 8, 4 epochs, 1e-8 epsilon, and 3e-5 learning rate. All experiments are created in Python using PyTorch-Transformers library, and evaluation metrics were developed using Scikit-Learn Python library. The implementation environment is based on Google Colab Pro for all experiments. We develop two classifications models using exactly the same system pipeline. Firstly, AraBERT model is used with fine-tuning on the 9 offensive language datasets as mentioned earlier, then, the same model is further fine-tuned using the training dataset from ArSarcasm-v2

[3] https://huggingface.co/

Model	Class	Precision	Recall	Macro-F1
	Positive	0.66	0.58	0.62
Baseline	Neutral	0.74	0.78	0.76
AraBERT	Negative	0.76	0.75	0.76
	Average	**0.72**	**0.71**	**0.71**
	Positive	0.62	0.52	0.56
Main	Neutral	0.73	0.79	0.76
AraBERT	Negative	0.74	0.74	0.74
	Average	0.70	0.68	0.69
	Positive	0.54	0.66	0.60
Baseline	Neutral	0.73	0.79	0.76
SalamBERT	Negative	0.76	0.73	0.74
	Average	0.71	0.69	0.70
	Positive	0.67	0.56	0.61
Main	Neutral	0.73	0.80	0.77
SalamBERT	Negative	0.75	0.73	0.74
	Average	**0.72**	0.70	0.70

Table 2: Performance Evaluation Results of the Development Dataset for the Sentiment Analysis Sub-Task.

Dataset for the target task. The second system deploy a customized version of AraBERT, called SalamBERT[4], which adds more tokens to the vocabulary of AraBERT and continue pre-training the model using the MADAR corpus (Salameh et al., 2018; Bouamor et al., 2019), which consists of multiple Arabic dialects to ensure the coverage of dialectical Arabic in word embeddings.

3.4 Results

Training and Development datasets experiments were evaluated with 5-fold cross validation for all performance metrics. Baseline models do not consider transfer learning across tasks, and are used as benchmarks for the evaluation process. Thus, baseline models in each sub-task are trained using the training dataset and evaluated using the development dataset. Results for the sarcasm detection sub-task are shown in Table 1 and for the sentiment analysis sub-task are shown in Table 2 from the development dataset. Main models refer to the model that consider transfer learning from offensive language detection task to the targeted task. As can be noticed from the tables, the variation among the performance of all models is insignificant for the development dataset.

The official results from the shared task organizers from the testing dataset are presented in Table 3 for the sarcasm detection sub-task and Table 4 for the sentiment analysis sub-task. Overall, SalamBERT-based system reports higher performance in sarcasm detection sub-task, while AraBERT-based system demonstrates higher performance in sentiment analysis sub-task. Among the 27 teams who participated in sarcasm detection sub-task, SalamBERT-based system ranked the 12th (0.5348 F1-sarcastic) and AraBERT-based system the 18th (0.5041 F1-sarcastic) based on F1-sarcastic metric. For the sentiment analysis sub-task, a total of 22 participants were included in the competition from which the AraBERT-based system ranked the 12th (0.6877 F-PN) and the

Metric	AraBERT	SalamBERT
F1-Sarcasm	0.5041	**0.5348**
Precision	0.6950	**0.7128**
Recall	0.6622	**0.6807**
Macro-F1	0.6732	**0.6922**
Accuracy	0.7607	**0.7727**

Table 3: Official Shared Task Results for the Testing Dataset from Sarcasm Detection Sub-Task.

Metric	AraBERT	SalamBERT
F-PN	**0.6877**	0.6259
Precision	**0.6136**	0.5580
Recall	**0.6318**	0.5813
Macro-F1	**0.6210**	0.5635
Accuracy	**0.6630**	0.6073

Table 4: Official Shared Task Results for the Testing Dataset from Sentiment Analysis Sub-Task.

[4] https://huggingface.co/Fatemah

SalamBERT-based system the 16th (0.6259 F-PN) according to the F-PN metric.

4 Error Analysis and Discussion

We manually investigate samples of the misclassified samples for each sub-task from both models. Among the common errors for sarcasm detection sub-task is offensive tweets that were classified as sarcasm, while it is not sarcasm. For example the following tweet is misclassified as sarcasm by all experiments using both models: 'ميسي قليل أدب المدربين يحطون خطين دفاع وهو يضربهم بتمريره وحده/ Messi is impolite, the coaches draw two lines of defense and he hits them alone'. For the sentiment analysis sub-task, most of the common errors are from the neutral class, such as 'اخبار الدوري الاسباني ـ الإصابة تضرب حارس مرمى برشلونةالإسباني/ La Liga news -"The injury hits Barcelona's Spanish goalkeeper', which has some offensive terms 'تضرب/ hit' but it is not negative, while all models classify it as a negative sentiment tweet. Further analysis is required to examine the extent of the overlap among the three tasks and the forms of content that are shared across them.

5 Conclusions

This paper represents the system used in submissions for Sarcasm Detection and Sentiment Analysis sub-tasks at the Sixth Arabic Natural Language Processing Workshop (WANLP)(Abu Farha et al., 2021). Our approach examines transfer learning across offensive language, saracsm detection, and sentiment analysis. The results highlight valuable impact of our approach for sarcasm detection task over sentiment analysis task.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. *ArXiv*, abs/2101.01785.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, and El Moatez Billah Nagoudi. 2020. AraNet: A deep learning toolkit for Arabic social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 16–23, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2020a. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2020b. Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Azalden Alakrot, Liam Murray, and Nikola S. Nikolov. 2018a. Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem M. Hajj. 2020. Multi-task learning using arabert for offensive language detection. In *OSACT*.

AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020. Leveraging affective bidirectional transformers for offensive language detection.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *Arabic Language Processing: From Theory to Practice*, pages 251–263, Cham. Springer International Publishing.

Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. 2017. Multilingual cyberbullying detection system: Detecting cyberbullying in arabic content. In *2017 1st Cyber Security in Networking Conference (CSNet)*, pages 1–8.

Amr Keleg, Samhaa R. El-Beltagy, and Mahmoud Khalil. 2020. ASU_OPTO at OSACT4 - offensive language detection for Arabic text. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 66–70, Marseille, France. European Language Resource Association.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of osact4 arabic offensive language detection shared task. 4.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. 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. Association for Computational Linguistics.

Ahmed Omar, Tarek M. Mahmoud, and Tarek Abd-El-Hafeez. 2020. Comparative performance of machine learning and deep learning algorithms for arabic hate speech detection in osns. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, pages 247–257, Cham. Springer International Publishing.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.