Sarcasm and Sentiment Detection In Arabic Tweets Using BERT-based Models and Data Augmentation

Abeer Abuzayed and Hend Al-Khalifa
iWAN Research Group
College of Computer and Information Sciences
King Saud University
aabuzayed1@students.iugaza.edu.ps, hendk@ksu.edu.da

Abstract

In this paper, we describe our efforts on the shared task of sarcasm and sentiment detection in Arabic (Abu Farha et al., 2021). The shared task consists of two sub-tasks: Sarcasm Detection (Subtask 1) and Sentiment Analysis (Subtask 2). Our experiments were based on fine-tuning seven BERT-based models with data augmentation to solve the imbalanced data problem. For both tasks, the MARBERT BERT-based model with data augmentation outperformed other models with an increase of the F-score by 15% for both tasks which shows the effectiveness of our approach.

1 Introduction

Sarcasm is a form of figurative language, where the speaker expresses his/her thoughts in a sarcastic way. The process of sarcasm detection relies on understanding people's true sentiments and opinions. Application of sarcasm detection can be beneficial in several NLP applications, such as marketing research, opinion mining and information categorization.

In recent years, sarcasm detection has received considerable attention in the NLP community (Joshi, 2016). Different approaches were used for sarcasm detection; Early approaches for sarcasm detection used feature-based machine learning models (Ghosh et al., 2018). Recently, deep learning methods have been applied for this task (Ghosh et al., 2020). For a comprehensive survey on sarcasm and irony detection see (Joshi et al., 2017). However, work on Arabic sarcasm detection is still in its early stages with very few works. There are few efforts on Arabic sarcasm detection such as the works of (Karoui et al., 2017); (Ghanem et al., 2020) and the recent efforts to build standard datasets for sarcasm detection such as (Abbes et al., 2020) and (Abu Farha and Magdy, 2020).

The current small size of the shared task labeled data-set and its imbalance nature makes it extremely difficult to build effective detection systems. Also, the context of the current sarcasm tweets does not have enough information to decide on its state which indeed makes the tasks more challenging.

In this paper, we describe the system submitted for the shared task on sarcasm detection and sentiment analysis in Arabic. We approached this challenge first by experimenting with different classical machine learning classifiers such as Support Vector Machines (SVMs), XGBoost, Random Forest that are trained on tf-idf features. Then, we experimented with different Deep Neural Networks (DNNs) along with character and word-level features. Finally, we conducted experiments on several BERT-based models such as MARBERT and QARiB. We took the most promising BERT-based models results for subtask 1 and subtask 2 on the training set, which was MARBERT for both of them and tested it with a new augmented dataset.

The rest of the paper is organized as follows: Section 2, describes the dataset. In section 3, we describe our approach in tackling the problem. Section 4 provides and discusses the results of subtask 1 and subtask 2. And in section 5, we provide a conclusion of our work.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 312–317
Kyiv, Ukraine (Virtual), April 19, 2021.

2 Dataset

The ArSarcasm-v2 dataset (Abu Farha et al., 2021) released for both shared tasks by the competition organizers is the same containing 12,548 training tweets. In addition to 3000 tweets for testing. The dataset was annotated for sarcasm detection task (Subtask 1) with the label "TRUE" for sarcastic tweets and "FALSE" for not sarcastic tweets. For the second shared task (Subtask 2) on sentiment analysis the labels are (NEU, NEG or POS) for neutral, negative or positive, respectively. Table 1 illustrates the label distributions for both tasks. It can be seen that the training dataset is quite imbalanced having only about 17% of the tweets labeled as sarcastic (TRUE) and 17% of them labeled as positive (POS).

Task	Class	Training set
Sarcasm Detection	TRUE FALSE	**2168** 10380
Sentiment Analysis	NEU NEG POS	5747 4621 **2180**

Table 1: Label distributions for both tasks.

3 System

This section provides a description of the different data preprocessing steps, models we used in the experiments, and our data augmentation approach.

3.1 Preprocessing:

For the preprocessing we have done 4 major steps to prepare the dataset as follows:

1. **Cleaning:** we removed all of the diacritics such as (tashdid, fatha, damma, kasra, etc.), English and Arabic punctuations, English words and numbers, URLS and USER mention tokens.
2. **Elongation removal:** any repeated character for more than twice was removed. For example, the word "أمووووت" becomes "أموت" after the preprocessing.
3. **Letter normalisation:** letters which appeared in different forms were transformed into a single form. For example, {إأآ} was replaced with {ا}, {ى} with {ي} ,{ة} with {ه} and {گ} with {ك}.
4. **Extract #hashtag keywords:** we removed the hash sign"#" and replace the underscore "_" within a hashtag with a white space to extract understandable key words, For instance, "#باسم_طلع_حرامي" turns into " باسم طلع حرامي".

3.2 Models

The past few years have witnessed a huge revolution in building various bidirectional transformer-based models, particularly for Arabic. Where they perform as powerful transfer learning tools that help in improving a wide range of natural language processing (NLP) tasks such as text classification, question answering, named entity recognition (NER), etc. While fine-tuning BERT-based models achieved state-of-the-art results on various downstream NLP tasks we will experiment the following BERT- based models:

- **MARBERT and ArBERT**: released by (Abdul-Mageed at el., 2020). Both are built based on the BERT-based model except for MARBERT which does not use the next sentence prediction (NSP) objective as it is trained on tweets which are basically short. ArBERT was trained on a collection of Arabic datasets which are mostly books and articles written in Modern Standard Arabic (MSA) with 6.5B tokens. While MARBERT trained on both Dialectal (DA) and MSA tweets and has 15.6B tokens. Additionally, MARBERT and ArBERT were experimented on ArSarcasm dataset (Abu Farha and Magdy, 2020).
- **QARiB** (QCRI Arabic and Dialectal BERT): was trained on a collection of Arabic tweets and sentences of text written on MSA with a total token number of 14B (Abdelali et al., 2020)[1].
- **AraBERTv02:** It was trained on Arabic corpora consisting of internet text and news articles of (8.6B tokens) (Antoun et al., 2020).
- **GigaBERTv3:** is a bilingual BERT for English and Arabic. It was pre-trained in a corpus (Gigaword, Oscar and Wikipedia) with ~10B tokens (Lan et al., 2020).
- **Arabic BERT:** was trained on about 8.2B words of MSA and dialectical Arabic too. (Safaya et al., 2020).

[1] https://github.com/qcri/QARiB

- **mBERT** (BERT multilingual base model (cased)): pretrained model supports 104 languages including Arabic was pre-trained on the entire Wikipedia for each language. (Devlin et al., 2018).

Our architecture as shown in Figure 1 is mainly based on fine-tuning the BERT- based models mentioned above. In our initial experiment MARBERT outperforms the other BERT-based models followed by QARiB. Thus, we decided to improve the performance of these two best models by adopting data augmentation techniques (described below) to solve the imbalanced data issue. In our experiments, all of the BERT- based models were fine-tuned with the same settings. The training set was splitted into 80% for training our models and 20% for validation. Where each model was trained for 5 epochs with a learning rate of 2e-06, maximum sequence length equals the maximum length seen in the training set and a batch size of 32. Google Colab free GPU and Huggingface pytorch versions of the previous mentioned models were used.

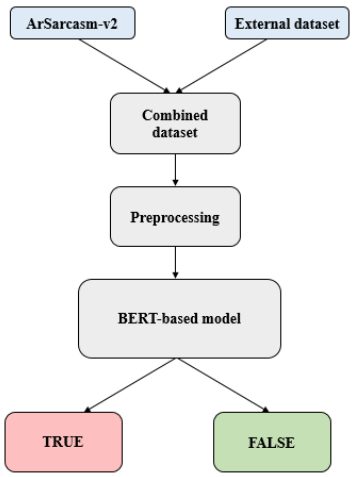

Figure 1: BERT- based model with data augmentation for sarcasm detection.

3.3 Data Augmentation

It is indicated that the sarcastic utterance usually has a negative implicit sentiment (Abu Farha at el., 2020). Moreover, from the provided dataset we found that 1939 (about 89%) of the tweets labeled as sarcastic and negative at the same time. Consequently, we hypothesis that every negative tweet can be sarcastic too. To investigate our hypothesis we used ASAD sentiment dataset (Alharbi at el., 2020) annotated with three sentiment labels (positive, negative and neutral). We could successfully retrieve 29924 tweets using the public tweet ids shared by the authors.

For sarcasm detection shared task, we replaced the labels annotated as “negative” with the label “TRUE” and replaced “positive” labels with “FALSE”. This produced an extra 4930 "FALSE” tweets and 4739 "TRUE" added to the original dataset which made the total number of the training set 22,217 tweets.

For the sentiment analysis task, we used the same dataset (which is basically annotated for sentiment) to provide the original training dataset with more positive and negative examples. A dataset of 4930 positive and 4739 negative examples were combined with the original training dataset and then tested on the best performing BERT-based model achieved by training them on the original training set.

4 Results and Discussion

In this section, we present and analyse the results of our experiments for subtask 1 and 2.

4.1 Results

The evaluation metrics used to test our system are F-score of the sarcastic class for subtask1 and macro-average F-score of the positive and negative classes (F-PN) for subtask 2. Both metrics were specified by the competition organizers.

4.1.1 Subtask 1:

Table 2 shows the results on the validation set in addition to the time the models took to train. MARBERT outperforms the other models with 0.647 F1-score on sarcastic class followed by QARIB with 0.597 F1-score. While mBERT gives the lowest F1- score of 0.411. The results of using MARBERT and QARiB with data augmentation are shown in Table 3. Data augmentation improves results by about 15% for both MARBERT and QARiB. This shows the effectiveness of our hypothesis to augment the dataset.

Model	F1-sarcastic	Training time (min: sec)
MARBERT	**0.65**	06:01
ArBERT	0.57	10:48
QARIB	0.60	10:25
AraBERTv02	0.56	10:07
GigaBERT	0.51	11:18
Arabic BERT	0.53	10:40
mBERT	0.41	06:30

Table 2: Results on original dataset for subtask 1.

Model	F1- sarcastic (valid set)	Training time (min: sec)
MARBERT	**0.80**	18:27
QARIB	0.75	18:00

Table 3: Results with data augmentation on subtask 1.

4.1.2 Subtask 2:

Similarly, the results of subtask 2 are shown in Table 4, QARiB achieved slightly higher F-PN score than MARBERT however, it has higher overfitting than MARBERT. Thus, we decided to try MARBERT with data augmentation. Expanding dataset size improves the performance of MARBERT by 15% as shown in Table 5 which also shows the effectiveness of our data augmentation approach.

Model	F-PN	Training time (min: sec)
MARBERT	0.71	05:59
QARIB	0.73	09:58

Table 4: Results on original dataset for subtask 2.

Model	F-PN (valid set)	Training time (min: sec)
MARBERT	**0.86**	19:28

Table 5: Results with data augmentation on subtask 2.

4.2 Official Results:

Based on the results above for both tasks we submitted the results of MARBERT on the test set. Table 6 presents the results of the MARBERT on the test set as reported by the competition organizers, compared to the results on the validation set. Obviously, there is a significant decrease in the model performance on the test set for both tasks, this is likely because of the overfitting issue.

Subtask 1		Subtask 2	
F1- sarcastic (valid set)	**F1- sarcastic (test set)**	**F-PN (valid set)**	**F-PN (test set)**
0.80	0.52	0.86	0.71

Table 6: Results of submitted model (MARBERT) for both tasks.

4.3 Error Analysis:

For further analysis for our proposed model results, extra error analysis is conducted to check where the model failed to correctly classify the tweets and try to find the reasons behind this mis-classifitcion. We randomly check a sample of 50 mis-classified examples. Table 7 lists some mis-classified tweets by our best performing model on sarcasm detection task. We found that there are several reasons for classifying sarcastic tweets as not sarcastic and vice versa. We summarise these reasons as follows:

- **Human annotation** is not 100% correct because annotators' cultures and backgrounds diversity might not be considered in the annotation process. For example, we believe that tweets 2 and 5 should be annotated as FALSE/ not sarcastic and TRUE respectively.
- **The absence of context:** in some tweets the context is missed and it was not possible for our model to understand the context and predict the label correctly. In examples 4 and 6 "Justin Bieber" is mentioned and it is expected that there is some special event happened, however, it is not clear using only one tweet. Thus, our model failed to classify both examples correctly. In addition, some tweets have media content and URLs which definitely clarify the context more which is not considered in this dataset.

	Tweet	True label	Predicted label
1	" بريطانيا السعودية واخد بالك أنت😂😂😂"	TRUE	FALSE
2	"#شبكة_أخبار_المعارك أطفال اصيبوا نتيجة الغارات الروسية على مدينة #حلب صباح اليوم حسبنا الله ونعم والوكيل https://t.co/JHhjktp7qu"	TRUE	FALSE
3	" اجتماع لوزراء الخارجيه ف الجامعة العربيه اللي مقرها القاهرة. كل دول مش جايين لمصر دول جايين لجامعة الدول العربية. بسيطة😉"	TRUE	FALSE
4	" الصوره متلاعبين فيها دامجين ملامح زين مالك و جستن بيبر"	TRUE	FALSE
5	" يعني على اساس لو ما حلف الناتو والسعودية وتركيا وقطر كنتوا اساسن حسنتوا تسيطروا على مخفر"	FALSE	TRUE
6	" جستن بيبر فرع اليمن"	FALSE	TRUE
7	"حلقة فيها لفتة للشهيد "جبران توييني" ليك مين بين ضيوفها !! مسخرة !"	FALSE	TRUE
8	"اقسم بالله قمة المسخره ماهم عرفين يخترعو عذر لكن الرجال حيردو عليهم في الملعب"	FALSE	TRUE

Table 7: Examples of mis-clasffied tweets for sarcasm class.

- **Usage of sarcastic keywords:** tweets 7 and 8 use the keywords "مسخرة" which means "sarcasm", but the tweet itself is not sarcastic. It is likely our system picked up the sarcastic word but failed to take into account the context in which the word was used.
- **Emojis are not processed**: as we left the emojis in tweets without any kind of preprocessing, we noticed that some emojis impact the classification process. For instance, in examples 1 and 3 it is probably that the emojis (😁 and ☺) give positive indication for our model that the tweets are not sarcastic, thus our model failed to classify them as sarcastic.

	Tweet	True label	Predicted label
1	"بيونسيه في أناقة مكياجها العيون – مجلة رنود https://t.co/UbnwcotH5v"	POS	NEU
2	"غوتشي ❤ https://t.co/vR5sY36zdA"	POS	NEU
3	"تهنئة لكل طالب وطالبة قضوا ما يزيد عن 150 يوماً معتصمين وواجهوا كل انواع التنكيل #جامعة_النيل"	POS	NEG
4	"وغاب عني وفراقه حبيبي شيء تآني ..عذاب مش بآلكلام احكيه 💔#اليسا https://t.co/LNjehLMxsQ "	POS	NEG
5	"# قروس يقولكم هاتوا ميسي وسواريز ونيمار 😎بيجلدهم مثل ماجلد الهلال في ثلاث ايام رايح جاي 😎 #الاهلي_برشلونه https://t.co…"	NEG	POS
6	" انشالله ما الاقي نفسي بتيك توك بعد جم يوم 😂😂"	NEG	POS

Table 8: Examples of mis-clasffied tweets for sentiment class.

Similarly, we examined a random sample of mis-classified tweets for subtask 2 and investigated them as shown in Table 8. We found some reasons similar to the previous findings for sarcasm detection task. We argue that there are some errors in the annotations process such as examples number 1 and 2 should be NEU (neutral) instead of positive. Also, the usage of negative terms in positive context impacts the model performance. Negative terms such as "تنكيل" in example 3 and "عذاب" in example 4 were recognized by our system but without any further understanding of the context where they were mentioned. Finally, positive emojis such as (😎 and 😂) in examples 5 and 6 respectively are likely to skew the polarity of the sentence and reverse its detected sentiment.

5 Conclusion

In this paper we experimented with seven BERT-Based models and we augmented the shared task data set to identify the sentiment of a tweet or detect if a tweet is sarcasm. We achieved promising results for sarcasm detection and sentiment identification by MARBERT model with data augmentation. Our error analysis indicates certain types of errors in the dataset and in the annotation that can be addressed in the future.

Last but not least, if we had more time to work on this shared task, we will build a lexicon for sarcasm, try other approaches for data augmentation and also revise the annotation in the dataset, since we found several tweets were mis-annotated.

References

Abbes, I., Zaghouani, W., El-Hardlo, O. & Ashour, 2020. DAICT: A Dialectal Arabic Irony Corpus Extracted from Twitter. LREC 2020.

Abdul-Mageed, M., Elmadany, A. and Nagoudi, E.M.B., 2020. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. arXiv preprint arXiv:2101.01785.

Abu Farha, I. and Magdy, W., 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. OSACT4 - LREC 2020.

Abu Farha, I., Zaghouani, W. and Magdy, W., 2021. Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic. *Proceedings of the Sixth Arabic Natural Language Processing Workshop.*

Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I.I. and Zhang, X., 2020. ASAD: A Twitter-based Benchmark Arabic Sentiment Analysis Dataset. arXiv preprint arXiv:2011.00578.

Antoun, W., Baly, F., and Hajj, H., 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. *LREC 2020 Workshop Language Resources and Evaluation Conference.*

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ghanem, B., Karoui, J., Benamara, F., Rosso, P. and Moriceau, V., 2020, April. Irony detection in a multilingual context. In *European Conference on Information Retrieval* (pp. 141-149). Springer, Cham.

Ghosh, D., Fabbri, A.R. and Muresan, S., 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, *44*(4), pp.755-792.

Ghosh, D., Vajpayee, A. and Muresan, S., 2020. A report on the 2020 sarcasm detection shared task. *arXiv preprint arXiv:2005.05814*.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. ACM Computing Surveys (CSUR), 50(5), 73. American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Karoui, J., Zitoune, F.B. and Moriceau, V., 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, *117*, pp.161-168.

Lan, W., Chen, Y., Xu, W. and Ritter, A., 2020, November. An Empirical Study of Pre-trained Transformers for Arabic Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4727-4734).

Safaya, A., Abdullatif, M. and Yuret, D., 2020, December. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. *In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 2054-2059).*