ArSarcasm Shared Task: An Ensemble BERT Model for Sarcasm Detection in Arabic Tweets

Laila Bashmal
King Saud University
Riyadh, Saudi Arabia
lailabashmal@outlook.com

Daliyah H. Alzeer
Taif University
Taif, Saudi Arabia
dr.dalya.h@gmail.com

Abstract

Detecting Sarcasm has never been easy for machines to process. In this work, we present our submission of the sub-task1 of the shared task on the sarcasm and sentiment detection in Arabic organized by the 6th Workshop for Arabic Natural Language Processing (Abu Farha et al., 2021). In this work, we explored different approaches based on BERT models. First, we fine-tuned the AraBERTv02 model for the sarcasm detection task. Then, we used the Sentence-BERT model trained with contrastive learning to extract representative tweet embeddings. Finally, inspired by how the human brain comprehends the surface and the implicit meanings of sarcastic tweets, we combined the sentence embedding with the fine-tuned AraBERTv02 to further boost the performance of the model. Through the ensemble of the two models, our team ranked 5th out of 27 teams on the shared task of sarcasm detection in Arabic, with an F1-score of %59.89 on the official test data. The obtained result is %2.36 lower than the 1st place which confirms the capabilities of the employed combined model in detecting sarcasm.

1 Introduction

What is sarcasm? According to Cambridge Dictionary, it is "the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way" (Cambridge, 2021). Former research in psychology and neurology reports that the mind requires the use of two distinct regions to process sarcasm as it implies a contradicting meaning to the literal one (Shamay-Tsoory, 2005). According to Capelli and her colleagues, the ability to analyze sarcasm develops from childhood. In their early stages, children are not able to perceive sarcasm by the text alone and it has to be accompanied by intonation for clearer understanding. Unlike adults who identify sarcasm effortlessly with intonation or from the text alone (Capelli et al., 1990). Megan Dress and her colleagues noticed that there are regional differences in the use of sarcasm when other demographic factors were controlled (Dress et al., 2008). Moreover, Yoritaka Akimoto and his colleague concluded in their paper that there are individual differences in the intended use of sarcasm that affects the processing and interpretation of verbal irony (Ivanko et al., 2004). From this perspective, we can understand the challenging aspect of processing the different meanings of sarcasm with machines.

Automatic sarcasm detection is the process of identifying if the text has a sarcastic meaning or not. This process is an essential step in the sentiment analysis task which requires a clear understanding of the intended meaning of the text. Over the years, researchers have proposed several methods to detect sarcastic features in texts. Applying several algorithms like gradient boosting (Ahuja et al., 2018), using linguistic features (González-Ibáñez et al., 2011), using incongruity phenomenon for creating a sarcasm detection system (Joshi et al., 2015), A2Text-Net deep learning model (Liu et al., 2019), and using a model based on embeddings from a language model (Ilić et al., 2018).

In Gonzalez's paper and Joshi's paper, they both utilized linguistic features to assess sarcasm detection in written texts. Furthermore, Liyuan Liu and Suzana Ili separately investigated two different deep learning methods for the sake of better sarcastic detection. Liu's and his colleagues' model used a hypothesis layer and a feature processing layer to test the hypothesis and to select the supported features before training a neural network for detection. Moving to Suzana Ili´s paper, she and her colleagues found that the word-level representations model lacks the richness that embeddings from language models or ELMo have.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 323–328
Kyiv, Ukraine (Virtual), April 19, 2021.

Dialect	Non-Sarcastic	Sarcastic	Negative	Neutral	Positive	Total
Egyptian	1,745	930	1,376	793	506	2,675
Gulf	487	157	264	259	121	644
Levantine	486	138	285	197	142	624
Maghrebi	28	15	25	12	6	43
MSA	7,634	928	2,671	4,486	1405	8,562
Total	10,380	2,168	4,621	5,747	2,180	12,548

Table 1: ArSarcasm-v2 Training Dataset Statistics for Sarcasm and Sentiment over the Dialects (Abu Farha et al., 2021).

Accordingly, they used the ELMo model to investigate the sarcastic features in textual data.

Even more, Arabic sarcasm detection is lagging behind in the field of sarcastic detection in NLP. Some efforts have been done such as Jihen Karoui and her colleagues who worked to detect irony in Arabic tweets (Karoui et al., 2017).

Our goal in this task is to classify Arabic tweets into sarcastic or not sarcastic by taking into account the direct and the overall meaning of the tweet. To this end, we explored the use of the BERT model in three different ways. First, we performed a knowledge-transfer from AraBERTv02 to a sarcasm detection task. Secondly, we employed Sentence-BERT trained model with the contrastive loss for extracting representative sentence-level embedding for sarcasm. Finally, mimicking the way the human brain processes the surface and intended meaning of a sentence in two different regions, we processed the data into two different models and the outputs of the two models were then concatenated into a one-tuned ensemble model by stacking a classifier on top of them.

2 Dataset and Methodology

2.1 Dataset Description

The shared task of sarcasm and sentiment detection in Arabic is based on the ArSarcasm-v2 dataset (Abu Farha et al., 2021). Table 1 shows some statistics of the released training set. The dataset is provided with dialects and sentiment labels. In total, it consists of 12,548 tweets with 2168 sarcastic tweets and 10380 non-sarcastic ones. The test set, which contains 3000 tweets, was released without labels for evaluation purposes.

2.2 Preprocessing

2.2.1 Cleaning

Hashtags and mentions: For the hashtags and mentions, we tried to extract hand-crafted features and tested if they would have any relation between the number of hashtags and mentions and the sarcasm in the tweet text. However, we did not notice any significant relation. Accordingly, we removed them from tweets for smoother analysis.

English letters, digits, punctuations, and Arabic diacritics: Similarly, we tried to extract hand-crafted features from the links, digits, and punctuations. Again, we did not notice any relevant connections between these features and the sarcasm in the tweet. Therefore, we removed the English letters and some punctuation. We also performed Arabic-letters normalization. This includes removing diacritics and normalizing different forms of Alef, Ya, and Ha each to a canonical form.

2.2.2 Emoji decoding

Unlike verbal sarcasm, written words do not have any intonations to help in detecting sarcasm (Kreuz and Caucci, 2007). Therefore, in textual contexts, emojis can help to provide us with emotional signals if these tweets have sarcastic features or not (Subramanian et al., 2019). Consequently, we decoded the emojis and their emotions by using the emojis' descriptions data (Emo). Since the emoji descriptions are in English, they were translated into Arabic using google translate API. We manually recheck the translated meaning to Arabic for more accuracy. For decoding, the emoji in each tweet is replaced with its equivalent textual description in Arabic.

2.3 Methodology

In this section, we provide details about the different BERT models we used in the experiments. These models are AraBERTv02, Sentence-BERT, and the Full-model that combines the two models.

2.3.1 AraBERTv02

Pre-trained bidirectional transformer-based models, known as BERT, represent the state-of-the-art

أنا راجع من الساحل وسأسعى للحضور بإذن الله ألف مبروك (not sarcasm)	بكم البرنامج السياحي من والى في فرنسا (not sarcasm)	similar
ويندوز 10 بيعمل تقريبا كل حاجه بمزاجه والموضوع مخيف (sarcasm)	أليس من المُضحك عندما يُحاضرك أحدهم عن الأخلاق والذوقيات وهو أكثر من يفتقر إليهما (sarcasm)	similar
ويندوز 10 بيعمل تقريبا كل حاجه بمزاجه والموضوع مخيف (sarcasm)	بكم البرنامج السياحي من والى في فرنسا (not sarcasm)	dissimilar

Figure 1: Dataset for Sentence-BERT Model

model in Natural Language Processing (NLP) and have proven to be successful on several NLP related tasks (Devlin et al., 2019).Unlike previous text representation models, BERT employs a self-attention mechanism method to capture long-range interactions among words in a text and enables the model to derive its contextual meanings. In order to employ this model in our task which has a relatively small dataset, we performed a transfer learning and fine-tuned a BERT model pre-trained on a very large corpus. We specifically used the AraBERTv02 base model (Antoun et al.). It is the second version of AraBERT that was pre-trained on 200 million Arabic sentences. The AraBERTv02 base model consists of 12 encoder layers with a hidden dimension of 768 and 12 attention heads. After performing a hyper-parameters search, we trained AraBERTv02 on a sarcasm detection task with the following hyper-parameters: the learning rate was set to 2e-05, the batch size is 16, the accumulation steps is 2, and fine-tuned the model with 2 iterations.

2.3.2 Sentence-BERT

Sentence-BERT is a siamese/triplet architecture that is based on BERT. It was mainly developed for information retrieval and paraphrasing tasks (Reimers and Gurevych, 2019). It accepts two (or more) sentences as inputs and projects each sentence to a vector space so that the semantically similar sentences are as close as possible to each other. One of the issues with word embeddings models is that they represent the semantic meaning of each word and lack the richness to represent the overall meaning of the piece of text. To overcome this obstacle, the sentence embeddings models solve the issue by showing the representation of the semantic meaning of an entire sentence. This representation is useful for applications where the implicit meaning of a sentence is different from the literal meaning of the sentence, especially in understanding the intended meaning of sarcastic sentences.

Figure 2: Sentence Embeddings Visualization via t-SNE: (a) Before Training the Sentence-Bert model;(b) After Training with 1 Iteration; (c) After Training with 5 Iterations.

Method	F1-score	Precision	Recall	Accuracy
BOW + LR classifier	45.88±2.07	39.78±1.68	55.28±3.10	78.02±1.50
TF-IDF + LR classifier	49.66±2.04	42.22±3.16	61.50±5.90	78.52±2.22
Sentence embedding + LR classifier	57.88±0.66	62.00±0.25	54.64±1.02	86.30±0.12
Fine-tuning BERT	61.06±0.99	64.40±1.14	58.20±2.49	86.78±0.23
Fine-tuning BERT and Sentence embedding	62.27±0.52	70.05±3.68	56.28±1.82	85.36±0.54

Table 2: Results on Validiation Set According to Different Methods.

For the implementation, we used the AraBERTv02 base model to adapt Sentence-BERT for sarcasm detection in the Arabic language (Antoun et al.). We fine-tuned the model to generate embeddings for the sarcasm detection task. For the sake of preparing the dataset for the siamese architecture, we rearranged the sentences in the dataset as pairs where similar sentences are given the label 1 and dissimilar sentences are given the label 0 as shown in the table (Figure 1).

We used a contrastive loss. Thus, if two tweets are expressing similar meanings, their embedding vectors are mapped to be closed to each other while maximizing the distance between the embedding of tweets with different meanings. Figure 2 shows how the sentence embeddings are mapped during the training phase. Finally, each tweet is represented by a fixed-size vector of its learned embedding.

2.3.3 Full-Model:

The architecture of the full model is shown in figure 3. The output of the fine-tuned BERT model is concatenated with the sentence embeddings obtained from the Sentence-BERT. This combined representation is then fed to an external classifier. The classifier consists of two linear layers: one with GELU activation and the second with a sigmoid activation to predict the sarcasm label. We fine-tuned the model using an SGD optimizer with a learning rate of 1e-04 for five iterations.

3 Experimental Results

The results of the five experiments of the method are shown in table 2 in the validation set. We split the dataset into two where %80 of the dataset is used for training and %20 is used for validation. In all experiments, we reported the results in terms of F1-score, recall, precision, and accuracy. F1-score is the official metric in this task, which combines the recall (TP/TP+FN) and the precision (TP/TP+FP) metrics. TP, FN, and FP refer respectively to the true positives, false negatives, and false positives. F1-score is given as 2 × (Precision × Recall)/(Precision + Recall) and it is more suitable than the accuracy for the binary classification task when the dataset is highly imbalanced.

3.1 Results of Baseline Methods

As a first attempt, we implemented a few baseline models, specifically the logistic regression(LR) classifier with TF-IDF and BOW word encodings. As shown in Table 2, BOW yields an F1-score of %45.88 compared to %49.66 to TF-IDF. These results indicate that the traditional word encodings methods are insufficient to model the complex meanings in sarcastic tweets.

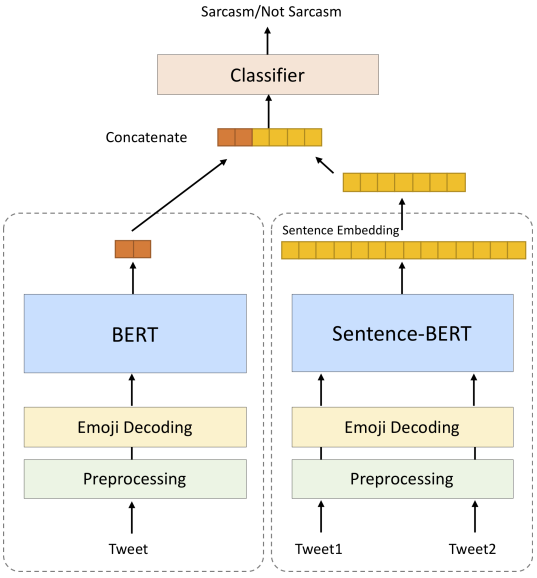

Figure 3: The Chart of the Full-Model that Combines the BERT Model and the Sentence-BERT Model. The Outputs of the Two Models are Concatenated and Fed into an External Classifier.

3.2 Results on the validation set

Next, we compared the results of the different BERT-based models on the validation set. As shown in Table 2, the results indicate that using BERT-based models can remarkably increase the

	F1-sarcastic	Accuracy	Macro-F1	Precision	Recall
Full-model	59.89	78.30	72.51	72.68	72.35

Table 3: Official Results Achieved by the Full-Model on the Test Set for the Sarcasm Detection task.

F1 score and the accuracy compared to traditional word encodings methods. Specifically, using sentence embeddings extracted from Sentence-BERT has improved the F1-score from %49.66 with TF-IDF to %57.88 using the same classifier. Moreover, fine-tuning AraBERTv02 with few iterations shows also a significant improvement with %61.06 F1-score. Finally, by simply combining the outputs of the fine-tuned AraBERTv02 and Sentence-BERT embeddings, the F1 score has increased to %62.27 and with a %1.42 decrease in the accuracy. This indicates that concatenating learned Sentence-BERT embeddings with BERT models has promising capabilities in detecting sarcasm.

3.3 Official Results on the Test set

Finally, we reported our submission results of the shared task on sarcasm detection in Arabic subtask-1 (Abu Farha et al., 2021). We submitted the full-model outputs as our primary submission, which achieved a %59.89 F1-sarcastic score (ranked 5th out of 27). The detailed results on the test set are shown in Table 3.

4 Conclusion

We presented our submission on the shared task on sarcasm detection and sentiment analysis in Arabic to tackle the problem of detecting sarcasm in Arabic. We explored different models based on BERT. We utilized a fine-tuning BERT and a Sentence-BERT to generate sentence embeddings that can be effective in detecting sarcasm. The final submission model combined the AraBERTv02 model and a representation obtained from a Sentence-BERT model to determine the sarcastic meaning of tweets. In the future, we will work on improving the performance of the model by exploring more sentence representation techniques.

References

Full Emoji List, v13.1, https://unicode.org/emoji/charts/full-emoji-list.html, note = Accessed: 2021-02-01.

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*.

Ravinder Ahuja, Shantanu Bansal, Shuvam Prakash, Karthik Venkataraman, and Alisha Banga. 2018. Comparative study of different sarcasm detection algorithms based on behavioral approach. *Procedia Computer Science*, 143:411–418. 8th International Conference on Advances in Computing Communications (ICACC-2018).

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

University Press Cambridge. 2021. *Cambridge Dictionary*. Cambridge University Press.

Carol A. Capelli, Noreen Nakagawa, and Cary M. Madden. 1990. How children understand sarcasm: The role of context and intonation. *Child Development*, 61(6):1824–1841.

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. Association for Computational Linguistics.

Megan L. Dress, Roger J. Kreuz, Kristen E. Link, and Gina M. Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

Suzana Ilić, Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, Brussels, Belgium. Association for Computational Linguistics.

Stacey L. Ivanko, Penny M. Pexman, and Kara M. Olineck. 2004. How sarcastic are you?: Individual differences and verbal irony. *Journal of Language and Social Psychology*, 23(3):244–271.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Jihen Karoui, Farah Banamara Zitoune, and Véronique Moriceau. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168. Arabic Computational Linguistics.

Roger J. Kreuz and Gina M. Caucci. 2007. Lexical influences on the perception of sarcasm. FigLanguages '07, page 1–4, USA. Association for Computational Linguistics.

L. Liu, J. L. Priestley, Y. Zhou, H. E. Ray, and M. Han. 2019. A2text-net: A novel deep neural network for sarcasm detection. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, pages 118–126.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tomer R. Aharon-Peretz J. Shamay-Tsoory, S G. 2005. The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology*, 19(3):288–300.

Jayashree Subramanian, Varun Sridharan, Kai Shu, and Huan Liu. 2019. Exploiting emojis for sarcasm detection. In *Social, Cultural, and Behavioral Modeling - 12th International Conference, SBP-BRiMS 2019, Proceedings*, pages 70–80. Springer Verlag.