akaBERT at SemEval-2022 Task 6: An Ensemble Transformer-based Model for Arabic Sarcasm Detection

Abdulrahman Mohamed Kamr and **Ensaf Hussein Mohamed**
Research Support Center in Computing and Informatics
Department of Computer Science,
Faculty of Computers and Artificial Intelligence,
Helwan University, Cairo, Egypt.
abdokamr94@fci.helwan.edu.eg, ensaf_hussein@fci.helwan.edu.eg

Abstract

Due to the widespread usage of social media sites and the enormous number of users who utilize irony implicit words in most of their tweets and posts, it has become necessary to detect sarcasm, which strongly influences understanding and analyzing the crowd's opinions. Detecting sarcasm is difficult due to the nature of sarcastic tweets, which vary based on the topic, region, the user's attitude, culture, terminologies, and other criteria. In addition to these difficulties, detecting sarcasm in Arabic has its challenges due to its complexities, such as being morphologically rich, having many different dialects, and having low resources. In this research, we present our submission of (iSarcasmEval) sub-task A of the shared task on SemEval 2022. In Sub-task A; we determine whether the tweets are sarcastic or non-sarcastic. We implemented different approaches based on Transformers. First, we fine-tuned the AraBERT, MARABERT, and AraELECTRA. One of the challenges that faced us was that the data was not balanced. Non-sarcastic data is much more than sarcastic. We used data augmentation techniques to balance the two classes, significantly affecting the performance. The performance F1 score of the three models was 87%, 90%, and 91%, respectively. Then we boosted the three models by developing an ensemble model based on hard voting. The final performance F1 Score was 93%.

1 Introduction

According to the Cambridge dictionary definition of Sarcasm, it is the use of remarks that mean the opposite of what they say, made to hurt someone's feelings or humorously criticize something, Like when you say Love this weather. (When the weather is horrible).

Sarcasm detection is determining whether or not a piece of text is sarcastic. Sarcasm is a significant challenge for sentiment analysis systems. This is because a sarcastic sentence usually contains an implicit negative sentiment that is expressed with positive expressions. This discrepancy between the surface and intended sentiments creates a difficult challenge for sentiment analysis systems.

Sarcasm detection is a difficult task for a variety of reasons. First of all, there aren't many labeled data resources for sarcasm detection. Moreover, any available texts that can be collected (for example, Tweets) contain many issues, such as an evolving dictionary of slang words and abbreviations, so it usually takes many hours for human annotators to prepare the data for any potential use. Furthermore, the nature of sarcasm detection adds to the task's difficulty, as Sarcasm can be considered relative and varies significantly between people, and it depends on many factors such as the topic, region, time, the events surrounding the sentence, and the readers/writers mentality and the; in other words, a sentence that one person finds sarcastic may sound normal to another. (Farha and Magdy, 2021)

In addition to these previous challenges, discovering irony in the Arabic language has its own set of challenges due to the complexities of the language, such as being formally rich, different dialects, lack of resources, and rapid development due to the inclination of the Arabic language. The Arab citizen makes fun of all his affairs, especially politics, which uses many words and terms that are implicit in it.

Transformers greatly assisted in significant advancements in NLP tasks. They are a new neural network that does not employ convolution or recursion. They instead use their attention to find correlations between words in the text. Transformers can process text in parallel, allowing them to learn much faster than sequential methods. They also outperform previous methods in terms of results.

Transformer-based language models have re-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 885 - 890

cently proven to be highly efficient at language understanding, giving promising results across various NLP tasks and benchmark datasets. The language modeling capability of these models aids in capturing the literal meaning of context-heavy texts. For Arabic NLP in particular, the best results for sentiment analysis are currently achieved by AraBERT, a language model proposed by (Antoun et al., 2020).

Despite recent advances, detecting sarcasm remains a difficult task because of implicit, indirect phrasing and the symbolic nature of language. When working with Twitter data, the task becomes even more difficult because the social media posts are often short and contain noise sources, code-switching, and the use of nontraditional dialectal variations. Furthermore, BERT-based models have struggled with rare words, which are more common in social media texts due to their informal nature and the prevalence of slang words. It is difficult for language models like AraBERT, which have been trained on structured corpora from Wikipedia.(Hengle et al., 2021)

In this study, we tackle (iSarcasmEval) sub-task A of the shared task on SemEval 2022.In SubTask A; we determine whether the tweets are sarcastic or non-sarcastic. We have proposed models based on transformers to discover the sarcasm that transformers have proven to excel in other NLP tasks such as sentiment analysis. We summarize what has been accomplished in this research in the following:

- Fine-tuning the state-of-the-art transformers-based models such as AraBERT, AraELECTRA, and MARBERT.
- One of the challenging problems in this task is that the dataset is unbalanced, with 2357 non-sarcastic tweets and 745 sarcastic. We solved this problem by using augmentation and balancing the two classes.
- We proposed an ensemble model based on hard voting between the three fine-tuned transformers, and it outperforms each of them.

The following sections are organized as follows; Section 2 presents the background, section 3 describes the task and dataset description, section 4 gives an overview of the proposed system, section 5 explores the results and discussion, and section 6 concludes and gives possible directions for the future.

2 Background

There are few attempts to work on Arabic sarcasm. The workshops in the field of Natural language processing for the Arabic language revived the interest in detecting sarcasm, as this workshop provided annotated datasets, which was considered a challenging obstacle in front of researchers. Such as the previous shared tasks on irony detection(Ghanem et al., 2019) along with the participants' submissions and dialectal sarcasm datasets by (Abbes et al., 2020); (Farha and Magdy, 2021);(Abu Farha et al., 2021).

This survey mainly focuses on the WANLP 2021 workshop and its shared task on sarcasm in Arabic. This shared task seeks to promote and draw attention to Arabic sarcasm detection, which is critical for improving performance in other tasks such as sentiment analysis. The dataset used in this collaborative task, ArSarcasm-v2, comprises 15,548 tweets that have been labeled for sarcasm, sentiment, and dialect. Subtask 1 on sarcasm detection received 27 submissions. The majority of approaches relied on using and fine-tuning pre-trained language models like AraBERT and MARBERT (Abdel-Salam, 2021); (Abuzayed and Al-Khalifa, 2021); (Alharbi and Lee, 2021);(Bashmal and AlZeer, 2021);(Faraj et al., 2021);(Gaanoun and Benelallam, 2021);(Hengle et al., 2021);(Husain and Uzuner, 2021);(Israeli et al., 2021);(Naski et al., 2021);(Wadhawan, 2021). Deep learning and traditional machine learning approaches were used by a few of the participants(Nayel et al., 2021).

The BhamNLP(Alharbi and Lee, 2021) team was ranked best in the sarcasm detection task, with an F1-Score 0.6225. They deployed a multi-task learning architecture trained for sarcasm and sentiment classification in their approach. The model is based on both a MARBERT and a CNN-LSTM model, with each model's output being concatenated and supplied to the final output layer. The CNN-LSTM utilized both word and character embeddings.

3 Task And Dataset Description

In SemEval 2022 Task 6: iSarcasmEval (Intended Sarcasm Detection In English and Arabic)(Abu Farha et al., 2022). The organizer offered two datasets, Arabic and English. There are three subtasks. SubTask A its aim is to determine whether the given text is sarcastic or non-sarcastic; SubTask B is applied in (English only): it aims to determine which ironic speech category the given

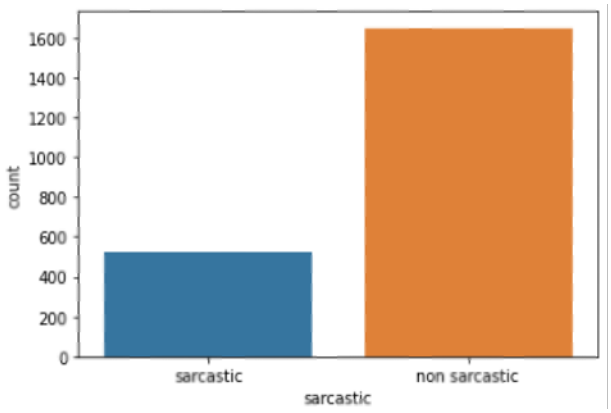

Figure 1: The distribution of the original dataset.

text belongs to, if an. It's a binary multi-label classification task. SubTask C: it aims to determine which text is the sarcastic one; if it is given two texts that convey the same meaning,
we have to clarify that our proposed system is a solution to Subtask A on the offered Arabic dataset. The Arabic dataset consists of 3,102 tweets with 2357 non-sarcastic tweets and 745 sarcastic ones. The test set contains 1400 tweets. It was released without labels for evaluation purposes. The dataset is provided with dialects and sarcasm labels. Table 1 shows some statistics of the released training set. Figure 1 shows the distribution of sarcastic and non-sarcastic tweets. Table 2 shows a sample of sarcastic and non-sarcastic tweets.

4 System Overview

The proposed model comprises two main phases: Data augmentation and the Ensemble module, consisting of three pre-trained transformers(AraBERT, MARABERT, AraELCTRA) and the Voting module, as shown in this Figure 2.

4.1 Data Augmentation

One of the challenging problems in this dataset is its small size, and it's unbalanced. The number of non-sarcastic tweets is 2357, while sarcastic is 745. We used data augmentation to balance the minority classes, Sarcastic. We used the NLPAug tool for augmentation. It is a python library based on non-contextual embeddings like Glove, Word2Vec, etc. And also for contextual embeddings like BERT and RoBERTa. By inserting or replacing words using word embedding, we used AraBERT for that purpose. Tables 3 shows a sample of tweets before and after augmentation.

4.2 Ensemble Module

This module consists of two phases. Firstly, apply three of the state of art pre-trained transformers, then use an ensemble hard voting technique.

(I) **Transformer Based Models:**

Deep learning methods have shown promising results in many machine learning domains, including natural language processing, computer vision and speech recognition. Due to architectures inspired by the human brain, deep learning techniques have recently outperformed traditional machine learning methods in terms of performance. Most deep learning techniques in the context of NLP use word vector representations to represent textual inputs (Mikolov et al., 2013a),(Mikolov et al., 2013b). These traditional techniques are being replaced by transformer-based techniques and significantly improve most NLP tasks, such as classification. As a result of the pre-training process, transformer-based techniques can generate efficient word embedding, making them powerful language models.

- **AraBERT** (Antoun et al., 2020) Pre-trained to handle Arabic text, AraBERT is a language model that is inspired by Google's BERT architecture. Six variants of the same model are available for experimentation: AraBERTv0.2-base, AraBERTv1- base, AraBERTv0.1-base, AraBERTv2-large, AraBERTv0.2-large, and AraBERTv2-base.
- **AraELECTRA**
(Antoun et al., 2020) With reduced computations for pre-training the transformers, ELECTRA is a method aimed toward the task of self-supervised language representation learning. ELECTRA models are inspired by the two primary components of Generative Adversarial Networks: generator and discriminator. They aim at distinguishing between real input tokens and fake ones. These models have shown convincing state-of-the-art results on Arabic QA data.

- **MARBERT**
provided by (Abdul-mageed et al., 2020). These models are based on the BERT-

Table 1: Distribution of tweets among different dialects

SARCASTIC/DIALECT	EGYPT	GULF	LEVANT	MAGHREB	MSA	NILE	TOTAL
NON-SARCASTIC	472	67	81	12	1470	255	2357
SARCASTIC	0	17	35	77	49	567	567

Table 2: Sample of tweets with Dialect and classified to sarcastic and non-sarcastic.

DIALECT	SARCASTIC	NON-SARCASTIC
Egypt	- -	احمد الشيخ احسن لاعب في الدوري و المصري احسن كورة لحد دلوقتي Ahmed Al-Sheikh is the best player in the league and Al-Masry team is the best football so far.
Gulf	الف مبروك الانتصار لفريقنا، مدرب الفريق الثاني ماخذ وضعية سوي نفسك ميت ! Congratulations to our team for wining , the coach of the another team didn't take anyting expect the week breath	اغنية رامي عياش اللي بحكي فيها على عيني و راسي والله عنك لحظة ما أخلي كتير حلوة Ramy Ayash song's which is about on my eyes and my head I swear I will not let you it's awesome
Levant	مافي حدا مثله يضحك عليك There is no one like him laughs of you	اللهم انتقم منه انتقم من كل من فرح بدماء المسلمين حسرك على ابناءك Oh God, take revenge on him, take revenge on everyone who rejoiced in the blood of Muslims
Magreb	واحد قال لو بغيت نتزوج بوحدة بيضاء وطويلة وخدامة قالت له خد الثلاجة Someone said to her mom: I wanted to marry a white woman and tall and serve me, she told him to take the refrigerator instade	طاح القدر يا ليبيا عارفين OH MY GOD, Libya, you know
MSA	رياض محرز ينقذ سمكة من الغرق، الله على اخلاقك يا فخر العرب Riyad Mahrez saves a fish from drowning, may God bless your morals, the best arabian player	بعد غيبة و طول انتظار ها قد عاد الاستقرار من جديد :)(النور قطع الهرم After an absence and a long wait, stability has returned again :)) The light cut agine
Nile	ضبط شخص بدبلوم انتحل صفة طبيب The arrest of a person with a diploma who pretended to be a doctor without getting in faculty of medicine	دراكولا لو كان عايش معانا لغاية دلوقت كان زمانه بيمص قصب If Dracula had been living with us until now, his time would have been sucking a cane

Table 3: Sample of the original and augmented tweets.

Original Text	Augmented Text
الليله دي دموع تماسيح	دي دموع تماسيح
ياعم مبارك دا جلده و مش بيطلع منه جنيه	ياعم دا جلده و مش بيطلع جنيه
الشحات مين له نص دستور البلد يا بلد	الشحات له نص البلد يا بلد

base and trained on a set of books and news articles. AraBERT was trained on 66GB of text-only news articles. MARBERT was trained on a larger dataset (128 GB), 50% of which is tweets. The variation in MARBERT's training data gives it the ability to handle better the variations in colloquial Arabic, which is very useful for sarcasm detection

(II) **Voting Module:** The three classifier outputs are fed into the voting module based on hard voting. It selects the majority prediction, whether sarcastic or non-sarcastic. For example, the ensemble module predicts sarcastic if two models predicted sarcastic, as shown in Figure 3.

5 Results and Discussion

5.1 Performance Evaluation Metrics

We used a variety of metrics to evaluate the proposed model quality. To evaluate the model's performance, we have to compute its Accuracy, Recall, Precision, and F1-Score. The equations for these evolution metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

5.2 Experimental Results

We applied two experiments, one without augmentation and the other with augmentation. We split the data into 60% for training, 20% for validation, and 20% for testing.

- **First experiment,** After splitting, we fine-tuned the three pre-trained transformers AraBRT, MARBERT, AraELECTRA on the original data and the ensemble model. The results show that F1-Score for AraBERT, AraELECTRA, MARABERT, and the Ensemble model were 74%, 69%, 77%, 77% respectively. As shown in Table 4.

- **Second experiment,** after data augmentation, and the number of tweets in sarcastic and non-sarcastic are equal. Experiment 2 shows that F1-Score for AraBERT, AraELECTRA, MARABERT, and the Ensemble model were 87%, 90%, 91%, 93% respectively. As we see in Table 5, there is an improvement in the performance of ARAELECTRA, MARABERT, and the Ensemble model.

The results show that our proposed model ranked sixth on the official competition. F1-score for sarcastic tweets is 0.4438, while F1-score for non-sarcastic tweets is 0.6222.

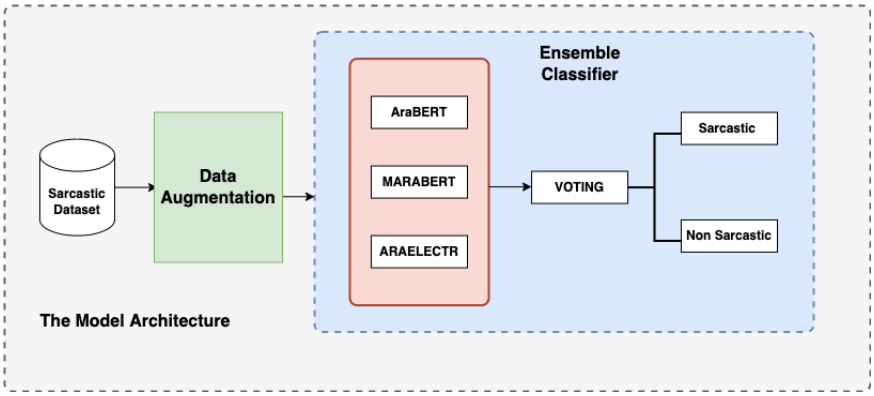

Figure 2: The proposed model architecture

Table 4: AraBert, MARBERT, ARAELECTRA, and Ensemble models on the original data.

Models	Precision	Recall	Accuracy	Macro-F1 score	F1 sarcastic
ARABERT	83%	84%	88%	83%	74%
ARAELECTRA	84%	79%	87%	81%	69%
MARABERT	89%	83%	90%	86%	77%
ENSEMBLE	88%	84%	90%	85%	77%

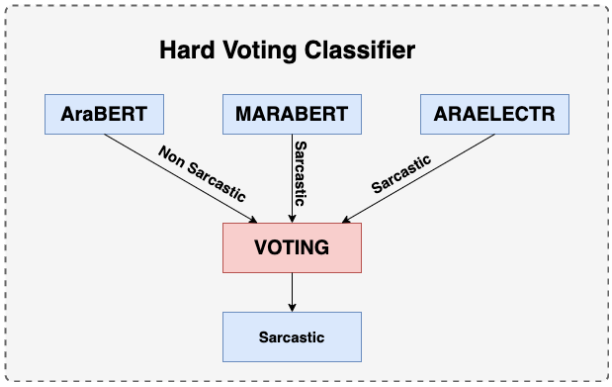

Figure 3: Hard voting Module.

6 Conclusion

We presented our submission on the shared Subtask A on Arabic sarcasm detection iSarcasmEval 2022 to tackle the problem of detecting sarcasm in Arabic. We explored different pre-trained models based on BERT. We noticed that the performance of AraELECTRA, MARABERT, and the Ensemble model is greatly improved after data augmentation and balancing the sarcastic and non-sarcastic tweets. The best submission model was the ensemble model; it applies hard voting on the AraBERT, MARABERT, and AraELCTRA. In the future, we plan to improve the performance of the proposed model by understanding and identifying which features are essential that contribute to enhancing the model prediction and troubleshooting unexpected model outputs.

References

I. Abbes, W. Zaghouani, O. El-Hardlo, and F. Ashour. 2020. Daict: A dialectal arabic irony corpus extracted from twitter. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, May*, page 6265–6271.

R. Abdel-Salam. 2021. Wanlp 2021 shared-task: Towards irony and sentiment detection in arabic tweets using multi-headed-lstm-cnn-gru and marbert. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 306–311.

M. Abdul-mageed, A. Elmadany, E. Moatez, and B. Nagoudi. 2020. Arbert marbert : Deep bidirectional transformers for arabic.

I. Abu Farha, W. Zaghouani, and W. 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*, page 296–305.

Ibrahim Abu Farha, Silviu Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Table 5: AraBert, MARBERT, ARAELECTRA, and Ensemble models on the augmented data.

Models	Precision	Recall	Accuracy	Macro-F1 score	F1 sarcastic
ARABERT	87%	87%	86%	86%	87%
ARAELECTRA	90%	90%	90%	90%	90%
MARABERT	91%	91%	91%	91%	91%
ENSEMBLE	93%	93%	93%	93%	93%

A. Abuzayed and H. Al-Khalifa. 2021. Sarcasm and sentiment detection in arabic tweets using bert-based models and data augmentation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 312–317.

A.I. Alharbi and M. Lee. 2021. Multi-task learning using a combination of contextualised and static word embeddings for Arabic sarcasm detection and sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 318–322.

W. Antoun, F. Baly, and H. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding.

L. Bashmal and D. AlZeer. 2021. Arsarcasm shared task: An ensemble bert model for sarcasmdetection in arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 323–328.

D. Faraj, D. Faraj, and M. Abdullah. 2021. Sarcasmdet at sarcasm detection task 2021 in arabic using arabert pretrained model. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 345–350.

I.A. Farha and W. Magdy. 2021. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Arabic Natural Language Processing Workshop*, page 21–31.

K. Gaanoun and I. Benelallam. 2021. Sarcasm and sentiment detection in arabic language a hybrid approach combining embeddings and rule-based features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, volume Figure 1, page 351–356.

B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, and P. Rosso. 2019. Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *ACM International Conference Proceeding Series*, page 10–13.

A. Hengle, A. Kshirsagar, S. Desai, and M. Marathe. 2021. Combining context-free and contextualized representations for arabic sarcasm detection and sentiment identification.

F. Husain and O. Uzuner. 2021. Leveraging offensive language for sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 364–369.

A. Israeli, Y. Nahum, S. Fine, and K. Bar. 2021. The idc system for sentiment classification and sarcasm detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 370–375.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed representations ofwords and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, page 1–9.

M. Naski, A. Messaoudi, H. Haddad, M. BenHajhmida, C. Fourati, and A. Ben Elhaj Mabrouk. 2021. In *iCompass at Shared Task on Sarcasm and Sentiment Detection in Arabic. Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 381–385. [link].

H. Nayel, E. Amer, A. Allam, and H. Abdallah. 2021. Machine learning-based model for sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop, Ml*, page 386–389.

A. Wadhawan. 2021. Arabert and farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets.