SarcasmDet at SemEval-2022 Task 6: Detecting Sarcasm using Pre-trained Transformers in English and Arabic Languages

Malak Abdullah and **Dalya Faraj** and **Safa Swedat**

Jumana Khrais and **Mahmoud Al-Ayyoub**
Computer Science, Jordan University of Science and Technology, Irbid, Jordan
mabdullah@just.edu.jo
dfalnore18, saswedat20, jdikhries162@cit.just.edu.jo
maalshbool@just.edu.jo

Abstract

This paper presents solution systems for task 6 at SemEval2022, iSarcasmEval: Intended Sarcasm Detection In English and Arabic. The shared task 6 consists of three sub-task. We participated in subtask A for both languages, Arabic and English. The goal of subtask A is to predict if a tweet would be considered sarcastic or not. The proposed solution SarcasmDet has been developed using the state-of-the-art Arabic and English pre-trained models AraBERT, MARBERT, BERT, and RoBERTa with ensemble techniques. The paper describes the SarcasmDet architecture with the fine-tuning of the best hyperparameter that led to this superior system. Our model ranked seventh out of 32 teams in subtask A- Arabic with an f1-sarcastic of 0.4305 and Seventeen out of 42 teams with f1-sarcastic 0.3561. However, we built another model to score f-1 sarcastic with 0.43 in English after the deadline. Both Models (Arabic and English scored 0.43 as f-1 sarcastic with ranking seventh).

1 Introduction

In recent years, sarcasm has received remarkable research attention due to its importance and extraordinary impact on society, especially with the significant increase in its use in social media (Ghosh et al., 2020; Van Hee et al., 2018; Faraj and Abdullah, 2021). However, sarcasm is saying something, and what is meant is the opposite. Adding to the difficulty of detecting sarcasm is that all the words in the sentence are positive, but the intended meaning is just the opposite (Channon et al., 2005). Therefore, the detection of sarcasm depends mainly on the general meaning of the sentence, which makes us move away from traditional methods and use artificial intelligence, especially natural language processing.

Natural Language Processing (NLP) is a branch of artificial intelligence that gives machines the ability to read and understand the meanings of humans (Nadkarni et al., 2011). As a result, machines became able to understand the underlying meanings of words and not just rely on keywords. Moreover, it helps reveal forms of speech, such as emotion analysis, humor, ridicule, abuse, etc (Abujaber et al., 2021; Qarqaz et al., 2021).

Task 6 at SemEval-2022, "iSarcasmEval: Intended Sarcasm Detection In English and Arabic" (Abu Farha et al., 2022) suggested three main subtasks in both Arabic and English languages: SubTask A is a binary classification problem that determines whether a tweet is sarcastic or not, SubTask B is a multi-label classification problem which determines the category of tweets sarcastic if it is found in English only, SubTask C is two sentences, one sarcastic and the other paraphrased to be non-sarcastic, define which one is sarcastic.

We only participated in subtask A. Our solution combines four state-of-the-art pre-trained NLP models: AraBERT and MARBERT for Arabic and BERT and RoBERTa for English. SarcasmDet placed seventh out of 32 teams in the subtask A-Arabic and seventh (after the competition deadline) out of 42 teams. We have experimented with the pre-trained language models with different hyperparameters using the simple transformers library. It is worth mentioning that using the ensemble technique has increased our score remarkably.

The paper is constructed as follows: Section 2 provides the related works. Section 3 describes the shared task and the provided dataset. Section 4 describes our system solution. Section 5 shows our experiments. Section 6 provides the results, and finally, the conclusion is in Section 7.

2 Related Work

Sarcastic texts over social media are among the most researched issues of importance for sentiment analysis. Authors of (Rajadesingan et al., 2015) proposed Sarcasm Classification Using a Behavioral modeling Approach. They investigate users'

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1025 - 1030

past tweets and use psychological studies to better sarcasm detection of tweets. They used three models to perform their evaluation (Logistic Regression (LR) Decision Tree (DT) and support vector machine (SVM)). The evaluation is performed on a dataset obtained from Twitter and provided on Kaggle with 9104 sarcastic tweets. They used different class distributions (1:1, 10:90, and 20:80) and 10-fold cross-validation techniques. The best performance was obtained from the LR model and class distribution (1:1) with 83.46% accuracy. In 2016, researchers of (Bouazizi and Ohtsuki, 2016) provided pattern-based features approach to detect sarcasm expressions posted on Twitter which extracted with the hashtag sarcasm. They applied four models (Random Forest (RF), SVM, K-Nearest Neighbour (KNN), and Max-Entropy (Max.Ent.)) on three different balanced datasets with different sizes (1000, 2256, and 6000 instances). RF model outperformed the rest with 81.3% in terms of F1-score. In this paper (Pawar and Bhingarkar, 2020), researchers almost did the same work; they used the Pattern-based approach and the same classifiers (RF, SVM, and KNN). The used dataset consists of 9104 sarcastic tweets. RF outperforms the rest in accuracy and F1-score with 81% and 79%, respectively.

The focus on detecting sarcasm in Arabic has emerged in recent years, and researchers have become increasingly interested in this field. In (Karoui et al., 2017), the authors were among the beginning people to supply a dataset to detect sarcasm in Arabic. First, they collected the dataset from Twitter of the Arabic tweets with different Arabic dialects, such as the Egyptian, Syrian, and Saudi dialects. Then, they cleaned the dataset, 5,479 tweets, including 1733 irony, and added four features for each tweet: surface, sentiment, shifter, and internal context features. Finally, they applied several machine learning algorithms, and experiments showed that the Random Forest classifier achieved a high accuracy of 72.76% for detecting sarcasm in Arabic tweets.

On the other hand, another dataset is provided for sarcasm in Arabic by (Farha and Magdy, 2020). This dataset contains three topics: sarcasm, sentiment, and dialect labels include 10,547 rows of tweets where 16% are sarcastic tweets. The researchers applied a Bidirectional LSTM deep learning approach (biLSTM) to achieve an F1-score of 0.46, which indicates the difficulty of detecting sarcasm in the Arabic language.

The researchers in (Faraj and Abdullah, 2021) participated in the WANLP 2021 Shared Task for subtask 1 (Sarcasm Detection) competition. First, they used the AraSarcasm-v2 dataset and then cleaned data by using NLTK library (Bird, 2006), such as normalized and removed emojis, links, and Html tags. Next, they implemented several pre-trained models such as AraBERt, multilingual BERT cased and uncased, and XLM-Roberta. Their solution, called SarcasmDet, is based on fine-tuning of the large AraBERT and base AraBERT, and they got first place with an accuracy of 0.7830 and fourth place with f1-sarcastic 0.5989.

3 Task and Data Description

In Shared Task on SemEval 2022 - Task 6 (iSarcasmEval): Intended Sarcasm Detection In English and Arabic (Abu Farha et al., 2022), has three subtasks in two languages: Arabic and English, and each task solves different requirements. We participated in SubTask A: a binary classification problem that determined whether a tweet contains sarcasm or not.

The dataset consists of nine columns: the tweet, the rephrase, sarcastic, which defines whether the tweet is sarcastic or non-sarcastic, irony, stair, understatement, overstatement, and the rhetorical question. The last five columns are the type of sarcasm contained in the tweets. Moreover, the dataset contains 3467 non-null rows. Regarding our task, a binary sarcasm classification of the tweets, the tweet, and the sarcasm are the only required columns to solve the task. Table 1 shows a sample of the utilized dataset in English subtask A.

Table 1: Sample of English data

Tweet	**Sarcastic**
The only thing I got from college is a caffeine addiction	1
I love it when professors draw a big questionmark next to my answer on an exam becauseI'm always like yeah I don't either	1
The population spike in Chicago in 9 months is about to be ridiculous	0
You'd think in the second to last English classof the year my prof would stop calling meSean	0

The data is highly imbalanced as there are 2600 tweets classified as non-sarcastic and 867 as sarcastic.

id	tweet	Sarcastic	rephrase	dialect
3	اية المهلبية دي يصحبي	1	ما هذا الجمال	nile
2805	أينما وجد العدل فثم شرع الله	0	NaN	msa

Table 2: sample from the training Arabic dataset

The Arabic training dataset contains 3102 tweets and five columns: id, tweet, sarcastic, rephrase, and dialect. They have categorized the text into sarcastic or non-sarcastic based on the author himself. Table 2 shows an example of training the Arabic dataset for subtask A- Arabic. The Arabic dataset in subtask A- Arabic is imbalanced due to class sarcasm and a clear difference between sarcastic and non-sarcasm tweets. The number of sarcastic tweets was 745, and the number of non-sarcastic tweets was 2357.

4 SarcasmDet Description

Texts are sequential, so they must be trained by models supporting data in which the order of its features is an important factor. Transformers are deep learning techniques that utilize the idea of self-attention mechanism (Potamias et al., 2020). In this work, two transformer-based pre-trained models are fine-tuned to achieve the requirements of the sarcasm detection task for each language. The two models are the BERT and the RoBERTa for English, where we combined two pre-trained models, AraBERT and MARBERT, for Arabic.

Subtask A - English Figure 1 shows the architecture of the proposed model for English. The ensemble technique is adopted by performing a weighted sum for the predictions of BERT and RoBERTa. One BERT machine with weight one and four RoBERTa machines, each with weight one, is deployed except for one machine with a weight of 0.5.

Regarding the BERT model, the dataset is tokenized by a pre-defined tokenizer of the model. The BERT model originally consists of 12 layers, the first ten layers are chosen to be untrainable and the last two layers to be trainable. Moreover, the model is trained using a batch size equal to 6 and the number of epochs equal to 3. As to the RoBERTa model, the dataset is tokenized using its pre-defined tokenizer. The layers are fine-tuned. The first nine layers are untrainable, the last three layers are trainable, and the model is trained using a batch size equal to 12 and the number of epochs equal to 3.

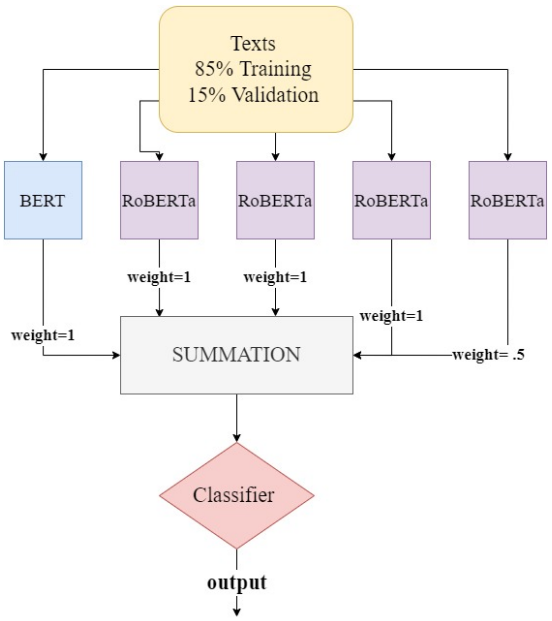

Figure 1: Architecture of SarcasmDet-English

Subtask A - Arabic The ensemble is a technique of combining several different models in the prediction process. In addition, we used hard voting that depends on the highest vote from all the model predictions. We combined two pre-trained models, AraBERT and MARBERT. The Arabic pre-trained model called AraBERT (Antoun et al.). AraBert is a pre-trained model that focuses directly on the Arabic language, and it is based on the BERT architecture (Devlin et al., 2018). There are two versions of AraBERT(v01 and v02). The first version, AraBERT-v01, was trained on 77M sentences, with a size of 23GB and 2.7B of words. The second version, AraBERT-v02, was trained on 200M sentences with 77GB and 8.6B words. MARBERT (Abdul-Mageed et al., 2020) is also based on the BERT architecture without the next sentence prediction and focuses on Dialectal Arabic and MSA.

In SarcasmDet, tweets are fed to the AraBERT and MARBERT. Next, we added the final layer, which is fine-tuning with the best hyperparameters as shown in **table 3** to classify the Arabic tweet into a sarcastic tweet or not. Then we applied the

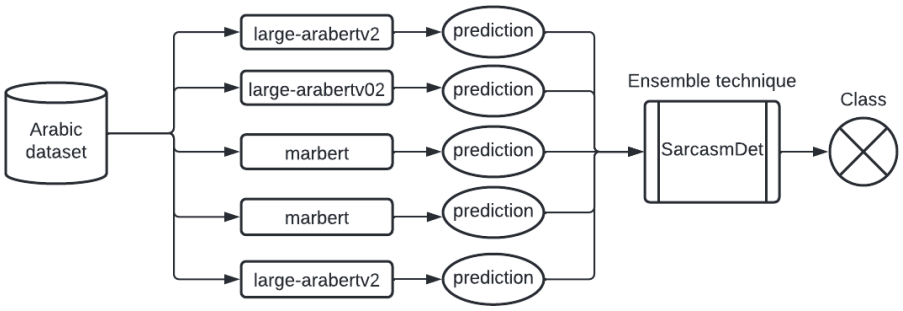

Figure 2: Architecture of SarcasmDet-Arabic

hard voting technique to select the final label in all tweets, as shown in Figure **2**.

Model	LR	Epoch	Batch Size
large-arabertv2	2e-5	5	8
marbert	1e-5	5	8

Table 3: Best the hyperparameter that we used in our models

5 Experiments

5.1 English

Unfortunately, the number of sarcastic sentences is much smaller than the number of the non-sarcastic ones, so using the accuracy to evaluate the model performance is not a good choice. Thus, the metric used to describe the version of the model is the F1-Score. F1-score is the harmonic mean of the precision and the recall. It is the average of the model's ability to find all sarcastic sentences and how accurate it is when classifying sentences as sarcastic.

The transformer-based models, BERT and RoBERTa, are trained using different hyperparameters. Table 4 and Table 5 show the results of the models with different fine-tuning.

Table 4: Result of Bert Model

BERT Parameters	**Results F1-sarcastic**
Batch size: 6 Number of epochs: 3 Number of trainable layers:4	0.38
Batch size: 6 Number of epochs: 3 Number of trainable layers:1	0.35
Batch size: 6 Number of epochs: 3 Number of trainable layers:2	0.39

Table 5: Result of RoBert Model

RoBERTa Parameters	**Results f1-sarcastic**
Batch size: 6 Number of epochs: 3 Number of trainable layers:3	0.36
Batch size: 4 Number of epochs: 2 Number of trainable layers:4	0.29
Batch size: 7 Number of epochs: 3 Number of trainable layers:4	0.37

5.2 Arabic

We have experimented with fine-tuning for two pre-trained models: AraBERT and MARBERT with different hyperparameters. Also, we attempted to increase the dataset using the augmentation technique with the ArSarcasm-v2 dataset to improve the results. We used the data set after the increment and fed the tweet to AraBERT and MARBERT, then fine-tune with hyperparameters to classify the Arabic tweet into a sarcastic tweet or not. **table 6** shows the hyper-parameters we have used in our experiments for the tested models.All of the models have been implemented using the HuggingFace library and SimpleTransforner pre-trained package.

6 Results

6.1 English

The model achieved a recall value equal to 0.37, a precision value equal to .51, and an F1-Score equal to 0.43. Table 7 shows the parameters and the results of the BERT model, RoBERTa model, and the ensemble model. It is noteworthy that the higher F1-Score in the SemEval competition in subtask A of task 6 is 0.6052, and the seventh rank is 0.4342. So the proposed architecture results, which is 0.4322, are equivalent to the eighth rank

Experiment	Model	LR	Epoch	Batch Size
1	large-arabertv2	1e-5	5	8
2	large-arabertv2	2e-5	5	8
3	large-arabertv02	2e-5	5	8
4	marbert	2e-5	5	8
5	marbert	1e-5	5	8
6	large-arabertv2	2e-5	5	8
7	marbert	1e-5	5	8

Table 6: all the hyper-parameter that we used in our experiments

in the competition.

Table 7: Best Parameters of BERT, RoBERTa and Ensembling.

Model	**Parameters**	Results (F1-Score)
BERT	Batch size: 6 Number of epochs: 3 Number of trainable layers: 2	.39
RoBERTa	Batch size: 12 Number of epochs: 3 Number of trainable layers: 3	.41
Ensemble	BERT model with weight :13 RoBERTa models with weight:11 RoBERTa with weight : .5	.43

6.2 Arabic

First, we applied AraBERT and MARBERT with different finetuning. AraBERT outperformed on MARBERT with a score of an f1-sarcastic 0.4255. Then, we augmentation the dataset and applied AraBERT and MARBERT, in this case, MARBERT outperformed AraBERT with a score of an f1-sarcastic 0.4065, although the data size increased, this did not lead to an improvement in the results, the best result was for ARABERT using Arabic data for the task. But SarcasmDet using the ensemble technique specifically hard vote significantly outperformed both AraBERT and MARBERT with an f1-sarcastic score of 0.4304, Table **table 8** shows the organizers' final result and table **table 9** shows all experiments that we have implemented.

7 Conclusion

Sarcasm is an influential issue in human life, whether written or spoken. However, sarcasm detection in texts is a challenging task. This paper presented our model SarcasmDet for solving subtask A- English and Arabic in task6 at SemEval 2021 - iSarcasmEval: Intended Sarcasm Detection in English and Arabic. SarcasmDet is based on the fine-tuning of two pre-trained NLP models for each language and then applied ensemble technique to improve the model. The models trained on a dataset obtained from a competition SemEval 2022 subtask A of task 6. For English subtask A, the dataset consists of 3467 records, 866 of them are sarcastic, and the rest are non-sarcastic. The results showed that the RoBERTa model outperformed the BERT model, where the Ensembling technique outperformed both in the f1-sarcastic score, which was 0.43. On the other hand, the Arabic dataset consists of 8K tweets divided into training and testing sets. Our solution SarcasmDet is ranked 7th out of 32 teams with an f1-sarcastic score of 0.4305.

8 Acknowledgement

We gratefully acknowledge the Deanship of Research at the Jordan University of Science and Technology (JUST) for supporting this work via Grant #20210200.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

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.

Dia Abujaber, Ahmed Qarqaz, and Malak A Abdullah. 2021. Lecun at semeval-2021 task 6: Detecting persuasion techniques in text using ensembled pretrained transformers and data augmentation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1068–1074.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language under-

Rank	F1-sarcastic	F-score	Precision	Recall	Accuracy
7th	0.4305	0.6114	0.6240	0.7412	0.6957

Table 8: Official results on subtask 1 (sarcasm Detection) test set.

Experiment	Model	F1-sarcastic
1	large-arabertv2	0.3847
2	large-arabertv2	0.4255
3	large-arabertv02	0.4166
4	marbert	0.3650
5	marbert	0.4015
6	large-arabertv2	0.3276
7	marbert	0.4065
8	SarcasmDet	0.4305

Table 9: Results in all experiments, SarcasmDet is an ensemble for experiments 2,3,4,5 and 7

standing. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Mondher Bouazizi and Tomoaki Otsuki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

Shelley Channon, Asa Pellijeff, and Andrea Rule. 2005. Social cognition after head injury: Sarcasm and theory of mind. *Brain and Language*, 93:123–134.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Ibrahim Abu Farha and Walid Magdy. 2020. 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.

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

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

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Neha Pawar and Sukhada Bhingarkar. 2020. Machine learning based sarcasm detection on twitter data. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 957–961. IEEE.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Ahmed Qarqaz, Dia Abujaber, and Malak Abdullah. 2021. R00 at nlp4if-2021 fighting covid-19 infodemic with transformers and more transformers. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 104–109.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.