AlexU-AL at SemEval-2022 Task 6: Detecting Sarcasm in Arabic Text Using Deep Learning Techniques

Aya Lotfy, Marwan Torki and Nagwa El-Makky
Computer and Systems Engineering Department
Alexandria University
Alexandria, Egypt
eng-aya.lotfy1419, mtorki, nagwa.elmakky@alexu.edu.eg

Abstract

Sarcasm detection is an important task in Natural Language Understanding. Sarcasm is a form of verbal irony that occurs when there is a discrepancy between the literal and intended meanings of an expression. In this paper, we use the tweets of the Arabic dataset provided by SemEval-2022 task 6 to train deep learning classifiers to solve the sub-tasks A and C associated with the dataset. Sub-task A is to determine if the tweet is sarcastic or not. For sub-task C, given a sarcastic text and its non-sarcastic rephrase, i.e. two texts that convey the same meaning, determine which is the sarcastic one. In our solution, we utilize fine-tuned MARBERT (Abdul-Mageed et al., 2021) model with an added single linear layer on top for classification. The proposed solution achieved 0.5076 F1-sarcastic in Arabic sub-task A, accuracy of 0.7450 and F-score of 0.7442 in Arabic sub-task C. We achieved the 2^{nd} and the 9^{th} places for Arabic sub-tasks A and C respectively.

1 Introduction

Sarcasm is ubiquitous phenomenon on the social web, and is difficult to be analysed automatically and manually by humans because of its nature. Sarcasm data can be very confusing to computer systems which use it to perform tasks such as sentiment analysis, opinion mining, author profiling, and harassment detection (Liu, 2012; Rosenthal et al., 2014; Maynard and Greenwood, 2014; Van Hee et al., 2018).

(Rosenthal et al., 2014) show that the sentiment polarity classification performance on non-sarcastic tweets is much better than on sarcastic ones, in the context of SemEval. Sentiment polarity classification is used widely in industry, driving marketing, administration, and investment decisions (Hassan Yousef et al., 2014). So it is important to create models for sarcasm detection.

A comparatively small dataset is a challenge we faced when working on the Arabic dataset that makes it difficult to train complex models. We used transfer learning to treat this issue. By using a transfer learning, a pre-trained model for some task on a large dataset can be used as a starting point in another task which improves the performance. It is used in a wide range of natural language processing (NLP) tasks.

Word embeddings such as word2vec (Mikolov et al., 2013), FastText (Joulin et al., 2016) and Glove (Pennington et al., 2014) can be used to initialize vectors learnt form large dataset. Recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) became the most popular NLP approach to transfer learning. Google AI Language team pretrained BERT model and fine-tuned it for a large range of tasks, such as question answering and language inference where it achieved state-of-the-art performance. MARBERT is built using the same network architecture as $BERT_{Base}$ (Devlin et al., 2019), without the next sentence prediction (NSP). MARBERT is trained on a large Twitter dataset. Therefore, we utilize fine-tuned MARBERT to solve this challenge.

This paper is organized as follows. In Section 2, we discuss relevant related works in sarcasm detection. We describe our proposed system in Section 3. The models implementation details are explained in Section 4. Section 5 describes the dataset for the shared task. Then we report and analyze the evaluation results in Section 6. Finally, we provide our conclusions in Section 7.

2 Related Work

A weak supervision and manual labelling methods can be used for the annotation process. A weak supervision method is to consider the texts sarcastic, if they meet predefined criteria, like including specific tags (e.g. #sarcasm, #irony) (Ptáček et al.,

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 891 - 895
July 14-15, 2022 ©2022 Association for Computational Linguistics

2014; Khodak et al., 2018). As pointed out by (Oprea and Magdy, 2020), noisy labels can be induced by this labelling method. Manual labelling is collecting texts and presenting them to human annotators for labelling (Filatova, 2012; Riloff et al., 2013; Abercrombie and Hovy, 2016). This can lead to a problem when annotator perception differs from author intention, as further outlined by (Oprea and Magdy, 2020).

Compared to English, only a few studies have been made on Arabic sarcasm detection. Among the studies completed in this area are research by (Riloff et al., 2013; Oprea and Magdy, 2019; Joshi et al., 2016; Bamman and Smith 2015; Campbell and Katz, 2012; Amir et al., 2016; Hazarika et al., 2018). (Oprea and Magdy, 2020) show the effect of sociocultural variables on sarcasm communication online, which makes the performance of models trained on English unpredictable, if they are trained on other languages. (Benamara et al., 2017; Abbes et al., 2020; Abu Farha and Magdy, 2020) relay on the two labelling methods mentioned above. Recently, efforts have been made by (Abu Farha and Magdy, 2020; Abu Farha et al., 2021 and Abbes et al., 2020) to create standard datasets to support sarcasm detection. In the SemEval-2022 Workshop, a shared task ('iSarcasmEval: Intended Sarcasm Detection In English and Arabic') (Abu Farha et al., 2022) was organised to contribute to the development of this area using a new labelling method that avoids the limitations of previous labelling methods.

3 Proposed System

SemEval 2022 task 6 focuses on detecting sarcastic tweets. The task supports Arabic and English languages and we tackle the problem on Arabic language. This task has three sub-tasks.

1. Sub-task A: Given a text, determine whether it is sarcastic or non-sarcastic.
2. Sub-task B (English only): A binary multi-label classification task. Given a text, determine which ironic speech category it belongs to, if any.
3. Sub-task C: Given a sarcastic text and its non-sarcastic rephrase, i.e. two texts that convey the same meaning, determine which is the sarcastic one.

We convert the input tweet to a fixed length sequence of words by padding shorter tweets and truncating longer ones. Then each word is replaced by its representation vector obtained from the pre-trained word embeddings model.

Dialect	Non-Sarcastic	Sarcastic	Total
MSA	1470	49	1519
Egypt/Nile	727	567	1294
Gulf	67	17	84
Levant	81	35	116
Magreb	12	77	89
Total	2357	745	3102

Table 1: Arabic dataset statistics for sarcasm detection over the dialects.

A MARBERT model is fine-tuned for sub-task A and then used for sub-tasks A and C. For sub-tasks C, the input tweets are passed to the model simultaneously and we consider the class of the tweet with the higher predicted score. We perform hyper parameters tuning to find the best parameters configuration.

4 Data Description

As mentioned above, annotator perception may differ from author intention. To overcome this problem, the authors annotated the data themselves, which is a new method to collect data introduced by the task's organisers.

Arabic and English datasets are collected using this method. For each sarcastic text, they provide a non-sarcastic rephrase to convey the same intended message, for English and Arabic datasets. Eventually, for English dataset, linguistic experts label each tweet to one of the ironic speech categories outlined by (Leggitt and Gibbs, 2000): sarcasm, irony, satire, understatement, overstatement, and rhetorical question. The dialect label of the text is included for the Arabic dataset.

Table 1 shows statistics of the Arabic training set, where we can find that 24% of the data is sarcastic (745 tweets). Most of the data is either in Modern Standard Arabic (MSA) or the Egyptian/Nile dialects, while there are few examples of the Magreb and Gulf dialects.

5 Implementation

We trained the proposed solution model using the given Arabic dataset. We divided its training data into 80% for training, 10% for validation and 10% for testing. In this section, we discuss the details of the different deep learning models we built or fine-tuned[1].

For our solution, we fix the tweets length to 64 by truncating longer tweets and padding shorter ones. This length is selected as the max value of the Arabic training tweets lengths after tokenization. After that each token in the input tweet is replaced with its vector representation obtained from a pretrained word embeddings model. We used pretrained MARBERT to initialize the words embeddings but these representations are then updated during the training of the deep learning models. The huggingface[2] pytorch implementation includes a set of interfaces designed for a variety of NLP tasks. Though these interfaces are all built on top of a trained BERT models, each has different top layers and output types designed to accomodate their specific NLP task. We used BertForSequenceClassification which is the normal MARBERT model with an added single linear layer on top for classification that we used as a sentence classifier.

5.1 MARBERT Model

Language models (LMs) exploiting self-supervised learning such as BERT (Devlin et al., 2019) which became a popular NLP approach to transfer learning. Transfer learning is used to reduce the time of the training and provide a better performance. This uses a pre-trained model as a starting point for training. Monolingual LMs pre-trained with larger vocabulary and bigger language-specific datasets usually perform better than multilingual models such as mBERT (Devlin et al., 2019; Virtanen et al., 2019).

Arabic has a large number of diverse dialects. Multilingual and Monolingual models such as mBERT and AraBERT (Antoun et al., 2020), respectively, are trained on mostly MSA datasets. The Arabic dataset used in this task has multiple dialects. This motivated us to use MARBERT which is trained on a large Twitter dataset (1B Arabic tweets), which involves both MSA and diverse dialects. The authors used the same network architecture as $BERT_{Base}$ (Devlin et al., 2019) to build the model, without the NSP objective which was found not crucial for model performance (Liu et al., 2019).

Metric	Non-Sarcastic	Sarcastic
Precision	$\frac{TN}{TN+FP}$	$\frac{TP}{TP+FP}$
Recall	$\frac{TN}{TN+FN}$	$\frac{TP}{TP+FN}$

Table 2: Precision and recall with respect to the sarcastic and non-sarcastic classes.

6 Experiment Results

In this section we report and discuss the results of the proposed solution when evaluated on our testing data. Moreover, we show the results of our submission to SemEval 2022 task 6 on Arabic data.

Our model ranked the second out of 32 participants for sub-task A and the 9^{th} out of 13 participants for sub-task C, SemEval 2022 task 6 (iSarcasmEval: Intended Sarcasm Detection In English and Arabic).

6.1 Results and Evaluation

We divided the training data into 80% for training, 10% for validation and 10% for testing. The official evaluation metric for sub-task A was the F-score of the sarcastic class (F1-sarcastic), the macro average of the F-score for sub-task B and the accuracy for sub-task C. F1-sarcastic is calculated using the following equation:

$$F1^{sarcastic} = 2 \times \frac{P^{sarcastic} \times R^{sarcastic}}{P^{sarcastic} + R^{sarcastic}}$$

Where $P^{sarcastic}$, $R^{sarcastic}$ are the precision and recall with respect to the sarcastic class. Table 2 presents the equations to calculate the precision and recall with respect to the sarcastic and non-sarcastic classes.

Table 3 presents the models' performance for sub-task A, the best result (0.9) was obtained by BertForSequenceClassification, which is the normal MARBERT model with an added single linear layer on top for classification that we used as a sentence classifier. The proposed system has also been submitted for the sub-task C, and was ranked the 9^{th} out of 13 participants. For sub-task C, the input tweets are passed to the model simultaneously and we consider the class of the tweet with the higher predicted score. We believe this result should be studied as a future work by investigating different

[1] The source code for the developed models can be found through: https://github.com/AyaLotfy/iSarcasmEval.

[2] https://huggingface.co/UBC-NLP/MARBERT.

Model	F1-sarcastic
MARBERT	**0.90**

Table 3: Performance of the model using our testing set for sub-task A on Arabic dataset.

Task	Main Metric	Result	Rank
A	F1-sarcastic	0.5076	2^{nd}
C	Accuracy	0.7450	9^{th}

Table 4: Main metric results obtained by the proposed model on the official test set for both sub-tasks A and C on Arabic dataset.

setup settings and, more importantly, to analyse errors on the test dataset. The model was fine-tuned for 4 epochs using the initial learning rate $2e^{-06}$, a batch size of 32, Adam weight decay optimizer and cross-entropy loss function. As well, we built other models but MARBERT outperforms them. The submitted model achieved an F1-sarcastic of 0.5076 on the official testing set for sub-task A.

6.2 Submission Results

For both aforementioned sub-tasks, we (AlexU-AL team) submitted the predicted classes based on the MARBERT model. For sub-task A, the proposed model achieved the second rank compared with the other systems proposed by other 31 participants. The submitted model achieved an F1-sarcastic of 0.5076 on the official testing set for sub-task A. For sub-task C, the model achieved an F-Score of 0.7442 and an accuracy of 0.7450 on the official testing set. Table 4 presents the official results achieved by our proposed model on the official testing set for sub-tasks A and C.

7 Conclusion

We used the fine-tuned MARBERT model in our submissions to SemEval 2022 task 6. We participated in the A and C sub-tasks for sarcasm detection in Arabic tweets. Our proposed approach is ranked the 2^{nd} and the 9^{th} in sub-tasks A and C, respectively. For future work, we explore the impact of building deeper neural networks with multiple convolutions or recurrent layers applied sequentially on the input text.

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. DAICT: A dialectal Arabic irony corpus extracted from Twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6265–6271, Marseille, France. European Language Resources Association.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.

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, Marseille, France. European Language Resource Association.

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.

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*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

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.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 574–577.

Farah Benamara, Cyril Grouin, Jihen Karoui, Véronique Moriceau, and Isabelle Robba. 2017. Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du défi fouille de textes deft2017.

John D Campbell and Albert N Katz. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480.

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*.

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.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

Ahmed Hassan Yousef, Walaa Medhat, and Hoda Mohamed. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. *arXiv preprint arXiv:1805.05388*.

John Leggitt and Raymond Gibbs. 2000. Emotional reactions to verbal irony. *Discourse Processes - DISCOURSE PROCESS*, 29:1–24.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Diana G Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Lrec 2014 proceedings*. ELRA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. *arXiv preprint arXiv:1910.11932*.

Silviu Vlad Oprea and Walid Magdy. 2020. The effect of sociocultural variables on sarcasm communication online. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 213–223.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. pages 73–80.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.