iCompass at Shared Task on Sarcasm and Sentiment Detection in Arabic

Malek Naski **Abir Messaoudi** **Hatem Haddad**
Moez Ben HajHmida **Chayma Fourati** **Aymen Ben Elhaj Mabrouk**
iCompass, Tunisia
{malek,abir,hatem,moez,chayma,aymen}@icompass.digital

Abstract

We describe our submitted system to the 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic (Abu Farha et al., 2021). We tackled both subtasks, namely Sarcasm Detection (Subtask 1) and Sentiment Analysis (Subtask 2). We used state-of-the-art pretrained contextualized text representation models and fine-tuned them according to the downstream task in hand. As a first approach, we used Google's multilingual BERT and then other Arabic variants: AraBERT, ARBERT and MARBERT. The results found show that MARBERT outperforms all of the previously mentioned models overall, either on Subtask 1 or Subtask 2.

1 Introduction

Sentiment Analysis and especially Sarcasm Detection in Arabic are challenging tasks, not only because of the labelled data scarcity, but also because of the complexity of sarcasm detection for models. Sarcasm is highly dependent on the culture, gender and other aspects like dialects. Indeed, the Arabic language has many variants and dialects, other than the Modern Standard Arabic (MSA). Even if some dialects share some vocabulary, they still differ according to countries, where each dialect has its own specifications which makes an impact on the task. For example, the sentence ”ما فيه إمكانيه تقربون المذيع والمذيعه بالأخبار اكثر من بعض!! مرره بعيدين بشكل يشتت الانتباه” represents a sarcastic text from the gulf region (Abu Farha et al., 2021) and reflects conservatism. The following example in gulf Arabic means ”Where can I find dresses for Ramadan and elegant cloaks?”: ”وين احصل جلبيات لرمضان وعبايات كشخة؟”. For a Tunisian speaker for example, this can be very confusing as ”كشخة” means ”ugly” in the Tunisian dialect! This can be used to express sarcasm if a Tunisian speaker addresses a gulf Arabic speaker using this word to express something negative while knowing that it means something positive in the gulf region.

Similarly, some words or concepts in dialects can be interpreted differently from one to another. As an example, the sentence ” فضيعين فضيعين ” presents a positive sentiment for some local dialects and a negative one for others.

In the sarcasm detection task, it is difficult to extract the information needed to detect if the sentence is sarcastic or not. A sarcastic one is defined by the shared task as the following: ”a sarcastic sentence usually carries a negative implicit sentiment, while it is expressed using positive expressions”. However, in some cases, sentences may contain contradictory sentiments expressed at the beginning and at the end. As examples:

- حلو أوي شعور الإهانة لما أقول لسواق التاكس قصر يقوم مشوحلي بإيده وسايق كإني خدت لطشة على وشي كده فعلا أبدعتي
- نفاق من الدرجة الأولئ ممتاز استمري ولكي كرسي في أسفل الدرك من النار

In other examples like ”كلها بقت احمد الشيخ اهو” , the task becomes even harder since the sentence is extracted from its context. For instance, people usually use positive words to express negative sentiments.

Working on Arabic sentiment analysis was first initiated by (Abdul-Mageed et al., 2011), and afterwards many contributions took place (Mulki et al., 2018), such as the open dataset on sentiment analysis and sarcasm by (Abu Farha and Magdy, 2020), followed by the dataset of this competition, ArSarcasm-v2 (Abu Farha et al., 2021).

The paper is structured as follows: Section 2 provides a concise description of the used dataset.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 381–385
Kyiv, Ukraine (Virtual), April 19, 2021.

Section 3 describes the used systems and the experimental setup to build models for Sentiment Analysis and Sarcasm Detection. Section 4 presents the obtained results. Section 5 presents a general discussion. Finally, section 6 concludes and points to possible directions for future work.

2 Data

The provided training dataset of the competition, ArSarcasm-v2 (Abu Farha et al., 2021), consists of **12548 tweets**, labelled with the tweet's sentiment class (POS: positive, NEG: negative and NEU: neutral), its sarcasm class (TRUE: sarcastic, FALSE: non-sarcastic), and its dialect (namely ”msa”, ”gulf”, ”egypt”, ”levant” and ”magreb”). The initial released training data was provided without the validation dataset. For this reason, we had to split the data to be able to validate our models. We chose a 80:20 split given the size of the dataset. We chose to have a balanced validation set. The distributions for the two subtasks are as shown in Table 1 and Table 2 respectively.
To preprocess the data, we removed emojis, URLs, diacritics, punctuation and any non-UTF characters.

Class	Train	Validation	Total
Sarcastic	913	1255	2168
Non-sarcastic	9125	1255	10380
Total	10038	2510	12548

Table 1: The dataset description for Subtask 1 - Sarcasm Detection.

Class	Train	Validation	Total
POS	1343	837	2180
NEG	3784	837	4621
NEU	4911	836	5747
Total	10038	2510	12548

Table 2: The dataset description for Subtask 2 - Sentiment Analysis.

3 System description

Pretrained contextualized text representation models have shown to perform effectively in order to make a natural language understandable by machines. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is, nowadays, the state-of-the-art model for language understanding, outperforming previous models and opening new perspectives in the Natural Language Processing (NLP) field. As a first approach, we used multilingual cased BERT model (hereafter mBERT) (Pires et al., 2019). Recent similar work have been conducted for Arabic which is increasingly gaining attention. As a second approach, we used three BERT Arabic variants: AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2021) and MARBERT (Abdul-Mageed et al., 2021).

3.1 mBERT

Following BERT's success, large pretrained language models were extended to the multilingual setting such as mBERT (Pires et al., 2019). mBERT was trained on 104 languages and has been used for fine-tuning on languages other than English.

3.2 AraBERT

AraBERT (Antoun et al., 2020), was trained on 70 million sentences, equivalent to 24 GB of text, covering news in Arabic from different media sources. It achieved state-of-the-art performances on three Arabic tasks including Sentiment Analysis. Yet, the pre-training dataset was mostly in MSA and therefore can't handle dialectal Arabic as much as official Arabic. A while afterwards, GigaBERT (Wuwei et al., 2020), a bilingual language model for English and Arabic, outperformed AraBERT on several tasks.

3.3 ARBERT

ARBERT (Abdul-Mageed et al., 2021) is a large-scale pretrained language model using BERT base's architecture and focusing on MSA. It was trained on 61 GB of text gathered from books, news articles, crawled data and the Arabic Wikipedia. The vocabulary size was equal to 100k WordPieces which is the largest compared to AraBERT (60k for Arabic out of 64k) and mBERT (5k for Arabic out of 110k).

3.4 MARBERT

MARBERT, also by (Abdul-Mageed et al., 2021) is a large-scale pretrained language model using BERT base's architecture and focusing on the various Arabic dialects. It was trained on 128 GB of Arabic Tweets. The authors chose to keep the Tweets that have at least 3 Arabic words. Therefore, Tweets that have 3 or more Arabic words and some other non-Arabic words are kept. This is because dialects are often times mixed with other foreign languages. Hence, the vocabulary size is equal to 100k WordPieces. MARBERT enhances the language variety as it focuses on representing the previously underrepresented dialects and Arabic variants.

3.5 System submission

We use these pretrained language models and build upon them to obtain our final models. Other than outperforming previous techniques, huge amounts of unlabelled text have been used to train general purpose models. Fine-tuning them on much smaller annotated datasets gives good results thanks to the knowledge gained during the pretraining phase, which is expensive especially in computational power. This is why, given our relatively small dataset, we chose to fine-tune BERT pretrained models. The fine-tuning actually consists of adding an untrained layer of neurons on top of the pretrained model and only tweaking the weights of the last layers to adjust them to the new labelled dataset.

We chose to train our models on a Google Cloud TPU of 8 cores using Google Colaboratory. The average training time of one model is around 5 minutes. We experimented with mBERT, AraBERT, ARBERT and MARBERT with different hyperparameters.

The final models that we used to make the submissions were:

- For Sarcasm Detection: a model based on MARBERT, trained for 8 epochs with a learning rate of 2e-5, a batch size of 128 and max sequence length of 128

- For Sentiment Analysis: a model based on MARBERT, trained for 3 epochs with a learning rate of 2e-5, a batch size of 128 and max sequence length of 128

4 Results

We have validated our models through balanced validation sets as mentioned in the data section, after splitting the training dataset into training and validation sets with a 80:20 ratio. The models based on MARBERT achieved the best results. We believe this is because MARBERT was trained mostly on dialectal Arabic which was underrepresented in previous pretrained models. Since this task's data is multi-dialectal, this model is expected to achieve the best performances.

4.1 Subtask 1 - Sarcasm Detection

The best results that we obtained for sarcasm detection on the sarcastic and non-sarcastic class are shown in Table 3. The accuracy achieved was **68.8%** and the F1-sarcastic was **57.7%**.

For reference, and to compare with other models, we also showcase the results obtained with mBERT and AraBERT. For AraBERT, the best results were found using a learning rate of 2e-5 and training for 3 epochs. The accuracy achieved was **64.1%** and the F1-sarcastic was **47%**. The results are shown in Table 4.

Class	Precision	Recall
Sarcastic	89.7 %	42.5 %
Non-sarcastic	62.3 %	95.1 %

Table 3: The best results using MARBERT for Subtask 1 - Sarcasm Detection.

Class	Precision	Recall
Sarcastic	89.9 %	31.8 %
Non-sarcastic	58.6 %	96.4 %

Table 4: The results using AraBERT for Subtask 1 - Sarcasm Detection.

For mBERT, the best results were performed using a learning rate of 2e-5 and training for 5 epochs. The accuracy achieved was **58.2%** and the F1-sarcastic was **31.2%**. Results are presented in Table 5.

Class	Precision	Recall
Sarcastic	88.5 %	19.0 %
Non-sarcastic	54.6 %	97.5 %

Table 5: The results using mBERT for Subtask 1 - Sarcasm Detection.

4.2 Subtask 2 - Sentiment Analysis

The best results that we obtained for sarcasm detection on the sarcastic and non-sarcastic class are shown in Table 6. The accuracy achieved was **72%** and the F-PN was **72.95%** .

For reference, and to compare with other models, we also showcase the results obtained with mBERT and AraBERT.

For AraBERT, the best results were found using a learning rate of 2e-5 and training for 3 epochs. The accuracy achieved was **69%** and the F-PN was **68.7%**. The results are shown in Table 7.

Class	Precision	Recall
POS	85.0 %	58.3 %
NEG	76.1 %	77.3 %
NEU	61.8 %	80.3 %

Table 6: The best results using MARBERT for Subtask 2 - Sentiment Analysis.

Class	Precision	Recall
POS	82.2 %	54.1 %
NEG	69.9 %	74.6 %
NEU	61.4 %	78.2 %

Table 7: The best results using AraBERT for Subtask 2 - Sentiment Analysis.

For mBERT, the best results were also found using a learning rate of 2e-5 and training for 3 epochs. The accuracy was **63.4%** and the F-PN was **60.15%**. The results are shown in Table 8.

Class	Precision	Recall
POS	74.6 %	42.4 %
NEG	63.9 %	68.6 %
NEU	58.4 %	79.3 %

Table 8: The best results using mBERT for Subtask 2 - Sentiment Analysis.

4.3 Official submission results

The final results that we achieved on the test set of the 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic were:

- Sarcasm Detection: F1-sarcastic equal to **48.6%**
- Sentiment Analysis: F-PN equal to **70.85%**

5 Discussion

Table 9 shows the results obtained over development data for the Sarcasm Detection task (True being sarcastic and False being non-sarcastic) and Table 10 shows the results obtained over development data for the Sentiment Analysis task. The development sets are as described in the data section. The results shown in Table 9 are clearly due to the training data being very unbalanced. The provided dataset has 2168 sarcastic examples and 10380 non-sarcastic ones. After the 80:20 split, the training set is left with 913 sarcastic examples and 9125 non-sarcastic ones.

It's also important to emphasise the difference between MARBERT's results and the other models' results. MARBERT was pretrained on various Arabic dialects and therefore works better with dialectal data.

	Predicted	
	False	True
False	1194	61
True	722	533

Table 9: Confusion matrix over development data - Sarcasm Detection

	Predicted		
	NEG	NEU	POS
NEG	647	162	28
NEU	107	671	58
POS	96	253	488

Table 10: Confusion matrix over development data - Sentiment Analysis

Table 11 and Table 12 review the official results of iCompass system for Sentiment Analysis and Sarcasm Detection against the top three ranked systems.

Team	Rank	F-PN
CS-UM6P	1	0,7480
DeepBlueAI	2	0,7392
rematchka	3	0,7321
iCompass	8	0,7085

Table 11: Ranking and results on the Sentiment Analysis Test set.

Team	Rank	F1-sarcastic
BhamNLP	1	0,6225
SPPU-AASM	2	0,6140
DeepBlueAI	3	0,6127
iCompass	21	0,4860

Table 12: Ranking and results on the Sarcasm Detection Test set.

6 Conclusion

Four language models were used to classify sentiment and to detect sarcasm (mBERT, AraBERT, ARBERT and MARBERT). The best results were obtained by MARBERT for both tasks with different hyperparameters, which was selected for the final submission. Future work would involve working on bigger contextualized pretrained models and enriching the existing Sarcasm Detection and Sentiment Analysis datasets.

References

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert marbert: Deep bidirectional transformers for arabic. *ArXiv*, abs/2101.01785.

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, 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*.

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.

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.

Hala Mulki, Hatem Haddad, and İsmail Babaoğlu. 2018. Modern trends in arabic sentiment analysis: A survey. *TAL Traitement Automatique des Langues*, 58(3):15–39.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Lan Wuwei, Chen Yang, Xu Wei, and Ritter Alan. 2020. Gigabert: Zero-shot transfer learning from english to arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.