Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect

Emna Fsih, Saméh Kchaou, Rahma Boujelbane and **Lamia Hadrich Belguith**
ANLP Research Group / Sfax, Tunisia
{emnafsih, samehkchaou4, rahma.boujelbane}@gmail.com
lamia.belguith@fsegs.usf.tn

Abstract

Arabic has a widely varying collection of dialects. With the explosion of the use of social networks, the volume of written texts has remarkably increased. Most users express themselves using their own dialect. Unfortunately, many of these dialects remain under-studied due to the scarcity of resources. Researchers and industry practitioners are increasingly interested in analyzing users' sentiments. In this context, several approaches have been proposed, namely: traditional machine learning, deep learning transfer learning and more recently few-shot learning approaches. In this work, we compare their efficiency as part of the NADI competition to develop a country-level sentiment analysis model. Three models were beneficial for this sub-task: The first based on Sentence Transformer (ST) and achieve 43.23% on DEV set and 42.33% on TEST set, the second based on CAMeLBERT and achieve 54.00% on DEV set and 43.11% on TEST set and the third based on multi-dialect BERT model and achieve 66.72% on DEV set and 39.69% on TEST set.

1 Introduction

Digital connectivity among Arab population has remarkably grown in the last few years. Apart from technological progress, the COVID-19 pandemic has been a factor for the increase of the penetration rate and consequently the increase in dialectal textual content in social networks. The dialect forms that differ from one region to another, have been considered for a long time to oral conversations of everyday life. They have neither standard nor sufficient resources for computational processing, unlike the mother language: MSA. As a result, there is a growing interest in dealing with this type of content. In this work we focus on developing a sentiment analysis model in the framework of shared task: Sentiment analysis of country-level Arabic. Several approaches have been proposed in the literature to build sentiment analysis models for poorly endowed languages. Deep learning has been proved as a very effective paradigm to classify sentiments in large data sets. However, this approach was not effective on small data sets and most of the time traditional machine learning algorithms get better scores.

In recent time transfer learning approaches has been shown to be beneficial to train a small data set and this by fine tuning a neural network model trained on a large data-set. BERT model (Devlin et al., 2018) based on transformer architecture is one of the effective transfer learning model. Indeed, (Moudjari et al., 2020) have used it to classify if an Algerian tweets is positive, negative or neutral. The model achieved an accuracy of 68%. Also, (Abdul-Mageed et al., 2020) have proposed another variant of BERT model baptized MARBERT that focused on both Dialectal Arabic (DA) and Modern Standard Arabic (MSA). The sentiment analysis model achieved an F-score of 71.50%. Using the same model, (Abuzayed and Al-Khalifa, 2021) have explored the effectiveness of augmenting data techniques proposed by (Abu Farha et al., 2021) to analyse the sentiments among a tweets corpus, the authors obtained an F1-score of 86%. Moreover, the Few-Shot Learning (FSL) approach has also been exploited in sentiment analysis. It is a sub-area of transfer learning which allows to classify new data when there is only a few training samples with supervised information which is the case in the present work. FSL is adapted with some success to NLP tasks, such text classification: (Bao et al., 2019) have proposed a meta-learning based method by using distributional signatures for few-shot text classification. (Luo et al., 2021) have presented few-shot text classification system upgraded by Label semantic augmented meta-learner (LaSAML) uses of label semantics.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 431 - 435
December 8, 2022 ©2022 Association for Computational Linguistics

	Num-Tweets	Num-Words	Num-Vocab	Num-Emojis
TRAIN positive	581	7934	4448	86
TRAIN negative	579	7900	4767	72
TRAIN neutral	340	4935	3064	49
DEV positive	179	2552	1696	42
DEV negative	190	2754	1897	40
DEV neutral	113	1670	1207	20
TEST positive	1179	16206	8477	139
TEST negative	1142	15963	8184	122
TEST neutral	679	9453	5357	85

Table 1: Data set statistics.

In order to build our system for NADI shared Task: Country-level sentiment analysis, we opted for transfer learning approach. In fact, we compared the effectiveness of three transfer learning models on different configurations of the corpus proposed by the organizers, this paper is structured as follows: Section 2 describes the NADI shared task's data set. Section 3 details the pre-processing and normalisation applied to the data set. Section 4 describes the data augmentation. Section 5 presents our proposed Sentiment analysis model for country-level Arabic. Finally, the conclusion is given in section 6.

2 Data

Three labeled data sets have been provided by the organizers (Abdul-Mageed et al., 2022) to build a sentiment analysis model:

TRAIN set for model training: it contains 1500 tweets, 23889 words and 10422 different words (including 207 emojis).

DEV set to adjust the model parameters: it contains 500 tweets, 7893 words and 4290 different words (including 102 emojis).

TEST set for evaluation: it contains 3000 tweets, 47293 words and 17926 different words (including 346 emojis).

More statistics on word distribution in each class are given in Table 1.

3 Data Pre-processing and Normalisation

In order to enhance the quality of the tweets corpus before feeding them as an input to the classification models, we apply the following pre-processing treatments:

- Eliminate all useless units such as website links and superfluous characters between words for example "©", "®", "@".
- Remove redundant letters and punctuation's marks.

Normalisation: We used in this phase the set of CAMeL tools for Arabic Language Processing (Obeid et al., 2020) that allows to apply the following alterations: Normalization of few Arabic characters and spelling errors in order to unify them into one form. In fact, there are some letters in Arabic that can be described as confusing in some cases. We normalize words containing such letters having one representative letter. For example, the different representations of HAMZA (أ, آ, and إ) were converted into the letter Alif (ا). We remove unnecessary characters including those with no phonetic value, such as (. . . , ‹, ›).

4 Data augmentation

Different data augmentation techniques were proved to be useful to efficiently augment the corpus. Thus, due to the small size of the competition corpus, we propose to augment the corpus with other versions of the corpus generated by different augmentation method and test their effectiveness to improve the quality of the sentiment classification model.

Contextual augmentation: The first augmentation method consists on applying the contextual embeddings method to tweets. We use the BERT model of the library NLPAug tool in order to insert (aug-insert) or replace words using word embedding (aug-subst). For that, we use the multi-dialect-bert-base-arabic language model (Talafha et al., 2020). The chosen words for words substitution or insertion are selected randomly. This method allows to obtain 3000 sentences in addition to 1500 existing tweets. Table 2 describes a few examples after and before contextual augmentation method.

Sentences before augmentation	Sentences after augmentation
لجل حبه كلشي يهون. انسي الناس وانسي الكون	واللي حبه كلشي بعيد. له الدنيا و الكون بعينه
عسي موب هو اللي ساحب علي امك يا مسفر	عسي انك انت اللي ساحب لنا الخير يا مسفر
بعون الله هنكمل حلمنا :)	بعون الله هنكمل حلمنا :) ♥
ومن يتوكل على الله فهو حسبه ☺	ومن يتوكل على الله فهو حسبه ♥ ☺

Table 2: Examples of sentences before and after contextual augmentation.

Emojis exploration: The second augmentation method is based on the exploration of emojis (aug-emoj). We adapt several techniques. Firstly, we drop all emojis with the demoji package Python.

Conversely to the first technique, we choose to augment the tweets with emojis. We fix a list of positive emojis for positive tweets such as ♥, ☺. We adapt a manual passage on the corpus to select the emojis used only for positive tweets.

Subsequently, we add emojis attached to positive tweets that do not contain emojis. We also fix a list of emojis for negative tweets such as ☹, :’(. These emojis are only used to express negative comments.

Then, we add them to negative tweets without emojis. This method of exploring emojis makes it possible to obtain 1000 sentences added to the corpus.

5 Sentiment analysis model for country-level Arabic

5.1 Transformer models

Among transformer models language modeling architectures:

Multi-dialect BERT model: We have explored in this work a Multi-dialect Arabic BERT pre-trained language model (M1_Bashar). The latter, used the weights of the Arabic-BERT model (Safaya et al., 2020) trained on 10M Arabic tweets have been developed by (Talafha et al., 2020) for the task of Arabic dialect identification problem.

CAMeLBERT model for dialectal Arabic: CAMeLBERT developed by (Inoue et al., 2021), is a collection of BERT models pretrained on the dialectal Arabic (DA) data sets. It is intended to be fine-tuned on an NLP tasks, such as NER, POS tagging, sentiment analysis and dialect identification. In this work, we exploit it to build a sentiment analysis model (M2_CAMeL).

Sentence Transformer (ST): Is a very popular approach deployed for semantic similarity and clustering (Reimers and Gurevych, 2019). ST is a simple and efficient alternative for few-shot text classification. In this work, we adapt Sentence Transformer Fine-Tuning (SetFit) to solve Sentiment classification on NADI-2022 tweets (M3_SetFit).

5.2 Experiments

5.2.1 Baseline

We investigated at first the efficiency of transformer architecture on Baseline tweets for training and testing steps. We pretrained at first on the corpus proposed by the organizers, the models mentioned in the previous section. M1_Bashar model achieves greater F1-score of 66.72% on DEV set compared to other models. However, M3_SetFit achieved the highest F1-sore on testing data. Table 3 presents the obtained results.

	DEV	TEST
M1_Bashar	**66.72%**	39.69%
M2_CAMeL	47.85%	41.72%
M3_SetFit	43.23%	**42.33%**

Table 3: MACRO F1-PN SCORE for transformer models.

5.2.2 Impact of augmentation techniques

In order to test the effectiveness of each proposed augmentation method, we associated each one separately with the baseline corpus. Table 6 shows the obtained results. The augmentation based on the exploration of emojis achieves the greater result with an F1-score equal to 65.33% on DEV and 43.11% on TEST.

5.3 Discussion

To analyse the strengths and weaknesses of our model, we provide the confusion matrix for its performance on the NADI test set in Figure 1. The matrix highlights a number of issues stemming

Tweet	Actual	Predicted	Correct-Label
احبك و ابيك و اباك و ابغاك جمعتها لرضاك و اختار فيها ♥	Neutral	Positive	Positive
ثقلا تجبر احدا على محبتك ولو كنت تحبه	Negative	Positive	Positive
ثق في نفسك ثم لا احد	Negative	Positive	Positive
يسلمو على ذوقك	Negative	Positive	Positive
انت راحتي فهل لك ان تبقى معي الى الابد	Negative	Positive	Positive

Table 4: Examples of mislabeled tweets (1).

Tweet	Actual	Predicted	Correct-Label
وعليك مالسلام ورحمه الله وبركاته	Positive	Negative	Neutral
وعليكم السلام ورحمه الله وبركاته	Neutral	Positive	Neutral

Table 5: Examples of mislabeled tweets (2).

from the training data set itself. For instance, it can be clearly seen that the model is moderately effective in terms of positive and negative classes. The model predicts the true classes in almost 50% of cases.

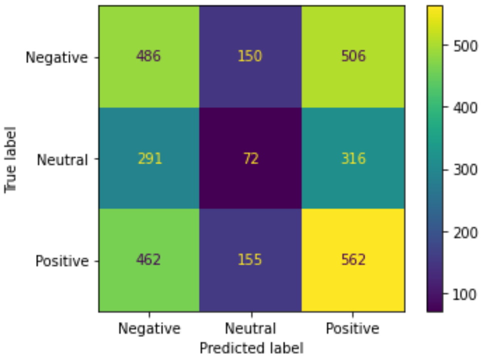

Figure 1: The confusion matrix of our SetFit model on NADI test set

However, the model is not at all efficient at the level of the neutral class this is expected given the number of neutral instances in the train. Some of the results shown in the confusion matrix have also led us to further investigate the data sets themselves. This resulted in finding that our model does in fact predict the correct class for certain tweets, which were somehow originally mislabeled. Some of these examples can be seen in Table 4. Another type of error was noticed possibly linked to spelling errors made by Internet users: for example the tweets cited in the Table 5 are two equivalent sentences but they are annotated differently in the test corpus. There is a small difference in the word وعليكم due to a typing error.

	DEV	TEST
M1_aug_insert	63.70%	39.04%
M1_aug_subst	64.16%	39.80%
M1_aug_emoj	**65.33%**	39.38%
M2_aug_insert	49.44%	39.78%
M2_aug_subst	47.13%	40.54%
M2_aug_emoj	54.00%	**43.11%**
M3_aug_insert	43.98%	40.84%
M3_aug_subst	48.21%	41.62%
M3_aug_emoj	46.09%	42.06%

Table 6: MACRO F1-PN SCORE for transformer models with augmentation.

6 Conclusion

In this paper, we presented our submitted method to the third NADI shared task. We proposed a transformer models for Sentiment analysis of country level Arabic. The experimental results shows that CAMeLBERT and SetFit models achieved an F1-score of 43.11% and 42.33% respectively on testing data set better than multi-dialect BERT model, while the latter achieved the best F1-score of 66.72% on development data-set.

References

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

Muhammad Abdul-Mageed, Chiyu Zhang, Abdelrahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. The nuanced arabic dialect identification shared task. In *Proceedings of the Seventh Workshop for Arabic Natural Language Processing at EMNLP 2022*, Abu Dhabi.

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.

Abeer Abuzayed and Hend 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*, pages 312–317.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2019. Few-shot text classification with distributional signatures. *arXiv preprint arXiv:1908.06039*.

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*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. 2021. Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2773–2782, Online. Association for Computational Linguistics.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1202–1210, Marseille, France. European Language Resources Association.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

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.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T. Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification.