Arabic Sentiment Ensemble NADI Shared Task 2

Abdelrahim Qaddoumi
New York University / amq259@nyu.edu

Abstract

This paper presents the 259 team's BERT ensemble designed for the NADI 2022 Subtask 2 (sentiment analysis) (Abdul-Mageed et al., 2022). Twitter Sentiment analysis is one of the language processing (NLP) tasks that provides a method to understand the perception and emotion of the public around specific topics. The most common research approach focuses on obtaining the tweet's sentiment by analyzing its lexical and syntactic features. We used multiple pretrained Arabic-Bert models with a simple average ensembling and then chose the best-performing ensemble on the training dataset and ran it on the development dataset. This system ranked 3rd in Subtask 2 with a Macro-PN-F1-score of 72.49%.

1 Introduction

Sentiment analysis (SA) is a process of computationally categorizing opinions expressed in a piece of text, especially in order to determine whether the attitude towards a particular product by labeling it positive, negative, or neutral. Sentiment analysis is a process of computationally categorizing opinions expressed in a piece of text, especially in order to determine whether the attitude towards a particular product by labeling it positive, negative, or neutral (Liu, 2012). As the world becomes increasingly digitized, sentiment analysis is becoming increasingly important. With the vast majority of people now using the internet and social media to communicate, NLP methods can help us analyze this massive amount of data to understand public opinion on various issues. Sentiment analysis can be extremely useful for businesses and governments, using sentiment analysis to track how the public's opinion (Liu, 2012).

One of the basic sentiment analysis approaches is a lexicon-based approach. This approach uses a list of words associated with neutral, positive, or negative sentiment. Then we use the generated list to score the sentiment of a text similar to what is done in (Neviarouskaya et al., 2010) and (Moreo et al., 2012). Another common approach is to use a machine learning algorithm to learn the sentiment of a text. However, this approach requires a training dataset of texts manually labeled with their sentiment for the machine learning algorithm to predict the sentiment of new texts such as (Chen and Tseng, 2011). Recently, transformers have been used for sentiment analysis, a deep learning method that learns the representation of text data for sentiment classification. This approach has been shown to outperform traditional machine learning methods such as support vector machines or deep learning models like long short-term memory or convolutional neural networks. In addition, the transformer approach can also handle a large amount of data, making it a scalable method for sentiment analysis like (Munikar et al., 2019).

While the accuracy of sentiment analysis in Arabic is still far from perfect, the current state of the art is much better than it was even a few years ago. The progress of sentiment analysis in Arabic has been significant in recent years. With the increasing availability of Arabic text data, there has been a corresponding increase in the development of methods and tools for sentiment analysis in Arabic. This progress will likely continue as more Arabic-specific data, and sophisticated methods become available (Al-Ayyoub et al., 2019).

There are a few reasons why sentiment analysis is complex with Arabic dialects. First, many variations in Arabic dialects make it difficult to identify patterns. Second, Arabic is a highly inflected language; words can have multiple meanings depending on their context in a sentence. This can make it difficult to determine the sentiment. Finally, Arabic dialects often use a lot of idiomatic expressions, which can also be challenging to interpret, as shown in (Laoudi et al., 2018).

The paper is structured as follows: Section 2

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 447 - 451
December 8, 2022 ©2022 Association for Computational Linguistics

concisely describes the used dataset. Section 3 describes the models used for the ensemble for Sentiment Analysis. Section 4 presents the results obtained for each combination. Section 5 presents related works. Section 6 presents a general discussion. Finally, section 7 contains the conclusion and points for future work.

2 Data

The competition provided a dataset of 5,000 tweets split into a training dataset of 1500 tweets, a development dataset of 500 tweets, and 3000 for the testing dataset. The tweets are labeled 'pos' for positive, 'neg' for negative, and 'neut' for neutral. We used the training dataset to decide on the best ensemble combination of models.

Class	Train	Development
pos	581	197
neg	579	190
neut	340	113

Table 1: Dataset description for Subtask 2 - Sentiment Analysis

3 System Description

We used Arabic pretrained language models that were fine-tuned for sentiment analysis publicly available on HuggingFace. There was no extra fine-tuning or preprocessing done after that. Instead, these models were used in a simple average ensemble by adding the logit values of the models' combination, using the maximum value for prediction, and then looping over the different combinations of the available models.

3.1 CAMeLBERT

While pre-trained language models such as ArarBERT have shown significant success in many NLP tasks in various languages, including Arabic, it is unclear what these multilingual models learn in Arabic and their most important features. Thus, Inoue et al. (2021) worked on an experiment to see how different sized pre-training data sets and language variants affected the performance of pre-trained language models. The paper culminated with nine different models, but four models that are trained for sentiment analysis CAMeLBERT-MSA, CAMeLBERT-DA, CAMeLBERT-CA, and CAMeLBERT-Mix. For a full-list off the dataset (Inoue et al., 2021).

The main difference between the models is the different Arabic languages used in the dataset and there are three different types of Arabic: Modern Standard Arabic (MSA), Classical Arabic (CA), and Dialectical Arabic (DA) (Al-Saidat and Al-Momani, 2010). MSA is the Arabic form used in most written documents and media today, it is based on the grammar and vocabulary of the Qur'an, and is the language of Arab countries' governments and schools, Classical Arabic is the Arabic form used in the Qur'an and other early Islamic literature, Dialectical Arabic is the form of Arabic spoken in everyday life in Arab countries, and each dialect differs based on region, social class, and religion (Al-Saidat and Al-Momani, 2010).

3.1.1 CAMeLBERT MSA SA Model

The model is trained on dataset for Modern Standard Arabic. The size of the model is 107GB with 12.6 Billion words.

3.1.2 CAMeLBERT DA SA Model

The model is trained on dataset for Dialectal Arabic. The size of the model is 54GB with 5.8 Billion words.

3.1.3 CAMeLBERT CA SA Model

The model is trained on dataset for Classical Arabic. The size of the model is 6GB with 0.847 Billion words.

3.1.4 CAMeLBERT Mix SA Model

The model is the combination of the three models CAMeLBERT CA SA Model, CAMeLBERT DA SA Model, and CAMeLBERT MSA SA Model. The size of the model is 167GB with 17.3 Billion words.

3.2 Arabic-MARBERT-Sentiment Model

The model is the result of fine-tuning MARBERT (Abdul-Mageed et al., 2020a) on KAUST dataset (Alharbi et al., 2020) which contains 95,000 tweets.The size of the model is 0.655GB. This work is done by Ammar Alhaj Ali on Huggingface but unfortunately was not able to cite the model as the researcher did not add a way to cite it.

4 Results

We have validated the different ensemble combinations models on the training dataset. The ensemble with model CAMeLBERT Mix SA Model and CAMeLBERT MSA (Modern Standard Arabic) SA model based on BERT achieved the best results. We believe this is because the combination of different dialects in the Mix model with the modern standard Arabic version has the majority of text features among all other models. This ensemble probably achieved the best results because the task's data contains both MSA and dialects.

4.1 Submission Results

The final results that we achieved on the NADI Shared Task Subtask 2 - Sentiment Analysis:

1. Development Sentiment Analysis: Macro-F1-PN equal to 72.49%
2. Test Sentiment Analysis: Macro-F1-PN equal to 69%

4.2 Subtask 2 - Sentiment Analysis

Model	Precision	Recall	F1-PN
MSA	62.61%	61.92%	70.18%
Mix	60.65%	60.08%	69.70%
DA	57.91%	57.82%	67.07%
CA	52.97%	52.76%	61.92%

Table 2: Single Model Results for Subtask 2 - Sentiment Analysis

Models Ensemble	Dataset	Precision	Recall	F1-PN
Mix_MSA	Dev	63.31%	63.21%	72.49%
Mix_MSA	Test	61.80%	61.33%	69.86%
Mix_CA_MSA	Dev	62.50%	62.35%	71.94%
DA_MSA	Dev	63.14%	62.99%	71.63%
Mix_CA_DA_MSA	Dev	62.28%	62.11%	71.36%
Mix_DA_CA	Dev	62.07%	62.01%	70.98%

Table 3: Results for Subtask 2 - Sentiment Analysis

5 Related Work

Arabic Sentiment Analysis received more attention recently, with many approaches showing effectiveness on the task; however, while some surveys have summarised some of the approaches for Arabic SA in literature, most of these are reported on different datasets, making it challenging to identify the most effective approaches among them (Farha and Magdy, 2021). Therefore, the researchers Farha and Magdy (2021) present a comprehensive comparative study of the most effective approaches for Arabic sentiment analysis.

The paper (Abdul-Mageed et al., 2011) kicked off the work to partially fill this gap of the lack of work on sentiment analysis in Arabic. They present a newly developed manually annotated Modern Standard Arabic (MSA) corpus with a new polarity lexicon, and investigate the impact of different levels of preprocessing settings on the classification task (Abdul-Mageed et al., 2011).

The newly generated data from Internet users on social media can be processed to extract useful information, such as users' opinions, by two main approaches: corpus-based and lexicon-based; (Abdulla et al., 2013) addresses both approaches to SA for the Arabic language using social media data. In (Abdul-Mageed et al., 2014), the researchers presented SAMAR, a system that uses lemma and the two parts of speech tagsets for sentiment analysis of Arabic social media, and addresses four issues: lexical representation, standard features for Arabic, handling of Arabic dialects, and genre-specific features.

Following the current trend of using transformers in English sentiment analysis, the introduction of AraBERT also promoted the usage of transformers for Arabic sentiment analysis. In (Wadhawan, 2021), the researchers present a strategy to identify the sentiment of the Arabic tweet. Their approach was two steps, the first step involved preprocessing using Farasa Segmentation, and the second step involved transformer-based models, AraELECTRA and AraBERT. This trend was also accompanied by a few tasks that work on cultivating the work of Arabic sentiment analysis such as (Abdul-Mageed et al., 2020b), (Abdul-Mageed et al., 2021), and lastly, (Abdul-Mageed et al., 2022).

6 Discussion

It is interesting to see that the models alone can achieve good results. However, the model with just classical Arabic was able to achieve no trivial result with only 5% of the size of the Mix model, which is worth investigating to understand the reason behind the result. The current pretraining models available freely have a relatively good performance on this task even with such a limited dataset and no training, which shows researchers' quality and hard work in the Arabic NLP field.

Future work would involve fine-tuning these models on the training dataset, trying different ensemble methods, and trying few-shot learning. Another thing to explore would be to try AraT5, the most recent state-of-the-art Arabic natural language understanding system (Nagoudi et al., 2022). The paper (Nagoudi et al., 2022) proposed a simple and effective transfer learning approach and evaluated the approach on Arabic, finding that the approach outperforms the state-of-the-art on all tasks in the benchmark.

7 Conclusion

Tables 2 and 3 show the results obtained over development data for the NADI task (Pos being positive, Neg being negative, and Neut being neutral) for the single model and ensemble model. Five language models were used to classify sentiment (mix, ca, da, MSA, and MARBERT). A simple two models ensemble (Mix and MSA) obtained the best results for the task and was selected for the final submission.

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.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

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

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Seven Arabic Natural Language Processing Workshop (WANLP 2022)*.

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Information processing & management*, 56(2):320–342.

Emad Al-Saidat and Islam Al-Momani. 2010. Future markers in modern standard arabic and jordanian arabic: a contrastive study. *European journal of social sciences*, 12(3):397–408.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. Asad: A twitter-based benchmark arabic sentiment analysis dataset. *arXiv preprint arXiv:2011.00578*.

Chien Chin Chen and You-De Tseng. 2011. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, 50(4):755–768.

Ibrahim Abu Farha and Walid Magdy. 2021. A comparative study of effective approaches for arabic sentiment analysis. *Information Processing & Management*, 58(2):102438.

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. *arXiv preprint arXiv:2103.06678*.

Jamal Laoudi, Claire Bonial, Lucia Donatelli, Stephen Tratz, and Clare Voss. 2018. Towards a computational lexicon for moroccan darija: Words, idioms, and constructions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 74–85.

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

Alejandro Moreo, M Romero, JL Castro, and Jose Manuel Zurita. 2012. Lexicon-based comments-oriented news sentiment analyzer system. *Expert Systems with Applications*, 39(10):9166–9180.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 806–814.

Anshul Wadhawan. 2021. Arabert and farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets. *arXiv preprint arXiv:2103.01679*.