Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

**RIST**

Information Processing at the Digital Age Journal

CERIST Natural Language Processing Challenge

March 29th, 2023

# XLM-T for Multilingual Sentiment Analysis in Twitter using oversampling technique

Mohammed E. Barmati[1,2] and Bachir Said[1,2]

[1]*Kasdi Merbah University, Ghardaia Road, BP.511, Ouargla, 30,000, Algeria*
[2] *Laboratory of Artificial Intelligence and Information Technologies (LINATI)*

**Abstract**

With the emergence of Pre-trained Language Models (PLMs) and the success of large scale, the field of Natural Language Processing (NLP) has achieved tremendous development such as Sentiment analysis (SA) that is one of the fast-growing research tasks in NLP. This paper describes the system that our team submitted to the CERIST NLP Challenge 2022 for task 1.b. The purpose of this task is to identify the sentiment polarity of the CERIST NLP Challenge 2022 datasets in English and Arabic languages comments collected from twitter. Our approach is based on a PL Model called XLM-T, and uses the Oversampling technique to solve the sentiment analysis problem of multilingualism in twitter. Experimental results confirm that this state-of-the-art model is robust achieving accuracy of 85%.

*Keywords:* NLP, Sentiment Analysis, pre-trained language model, oversampling technique.

## 1. Introduction

The development of web sector and the emergence of social media attracted more people due to its interactive nature. Hence, a massive amount of data was created during the last sixteen years. Social media

data and website comments contain significant information about people's opinions on a specific topic. Thus, it is necessary to take the benefit of this thesaurus information. Opinion Mining is a process of automatic extraction of knowledge from the opinion of others about a particular topic or problem Padmaja et al., 2013. According to Liu et Zang, 2012, Sentiment Analysis is defined as ''the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes and emotions concerning entities such as products, services, organizations, individuals, issues, events, topics, along with their attributes''. SA is classified into three distinct levels, specifically document, sentence and aspect levels based on specific classes such as positive, negative, or neutral Birjali et al., 2021.

A large number of tools and approaches, in the literature, are utilized to conduct the SA task. Most of them are designed to manage SA in English which is the science language Alotaibi, 2015. Arabic language has received less efforts compared with other languages Tsarfaty et Goldberg 2010 however, hundreds of studies have been proposed for ASA. Since its introduction since a decade, ASA has become one of the most popular forms of information extraction from the reviews. These reviews contributed in many benefits, such as showing the product brand or service valuable insights Khasawneh et al.; 2013, Mohammed et al., 2016, identifying potential product advocates or social media influencers Elouardighi et al., 2017, Kechaou et al., 2013, and detecting e-mail spam Hammad et al-halees, 2013.Consequently, ASA has been studied in various contexts, and a large number of studies has been published on the topic.

In this competition we focused on developing a solution for CERIST NLP Challenge 2022 sentiment analysis task1.b, we used a multilingual pre-trained language model called XLM-T Barbieri et al., 2022, before the training process we pre-processed the dataset and we used the Oversampling technique YAp et al., 2013 in order to improve the models performance. The rest of this paper is organized as follows. Section 2 introduces transformers-based Arabic sentiment analysis. Section 3 describes our methodology. Section 4 presents the experimental results. Finally, our conclusion is presented in Section 5.

## 2. Transformers-based Arabic sentiment analysis

There has been a lot of work on sentiment analysis. However, most of it is focused on English as it is the most widely used language. However Arabic language is also a widely spoken Semitic language and is an official language in 28 countries with around 400 million native speakers Darwish et Magdi, 2014.

At present, the study of multilingualism has become a new upsurge, and some related tasks organized recently have attracted a large number of researchers including Arabic language which has become a very interesting and challenging topic for researchers with its various topics and tasks NAssif et al., 2022. In addition to hate speech detection and spam detection, there are many important and related tasks to begin with such as Arabic sentiment analysis (ASA).

Given the effectiveness of transformer-based models, there have been various transformer models used in Arabic sentiment analysis. The widely utilized models are Multilingual BERT, AraBERT, and MARBERT Alamary, 2022. The author in Abuzayed et Al-khalifa, 2021 addressed sarcasm and sentiment detection using six BERT-based models including: MARBERT, QARiB, AraBERTv02, GigaBERTv3, Arabic BERT, and mBERT. MARBERT achieved promising results for both tasks. Authors in Alduailej et Alothaim, 2022 proposed AraXLNet, with Farasa segmenter and they have achieved good results in sentiment analysis task for Arabic using multiple benchmark datasets.

Authors in Barbieri et al., 2022 consisting of an XLM-RoBERTa introduced XLM-T, a model to train and evaluate multilingual language models for sentiment analysis in Twitter with eight different languages including English and Arabic, they compared the Twitter-based multilingual language model with a standard multilingual language model trained on general-domain corpora and find out that a single multilingual model is often more practical and versatile.

## 3. Methodology

### *3.1 Dataset*

In order to run our experiments, we used the provided datasets in two languages (Arabic and English) by the organizer, the data mainly comes from twitter comments. The Arabic dataset contains a training set of 3,355 Arabic tweets and test set of 6,100 Arabic tweets divided into three targets either "Positive 22.1%", "Neutral 43.8%" or "Negative 34%".While the English dataset contains a training set of 16,173 English tweets and test of 12,284 English tweets divided into three targets either "Positive 42.7%", "Neutral 42.6%" or "Negative 14.8%".

### *3.2 Data preprocessing*

In this section we employed three steps just to prepare the organizer datasets to our experiments model.

#### *3.2.1 Data cleaning*

In this step we cleaned the dataset by removing unwanted special characters and transform the targets type from words to integers ("Positive'' to 2, "Negative" to 0 and 'Neutral' to 1).

#### *3.2.2 Dataset sampling*

The class imbalance problem has been reported as a major obstacle to the induction of a good classifier in Machine Learning algorithms Batista et al., 2004. Since the organizer datasets is not balanced over the targets we used the Oversampling technique in order to improve the model performance. First, we need to determine the distribution of two classes (e.g. 'Positive', Negative or 'Neutral') before we proceed to balance out the data. Last, replicate the minority classes to achieve equal distribution with the majority class YAp et al., 2013.

To proceed in our experiment, we merge both training set of the English and Arabic dataset, and randomly shuffled their order. Finally, we get a new training dataset of 25,122 tweets of English and Arabic languages.

#### *3.2.3 Tokenization*

This step consists of preparing the organizer datasets as input for our experimental model by splitting the input text into tokens and designated them as the input_ids. These ids were padded to a fixed length to avoid variations per row.

### *3.3 Fine-tuning XLM-RoBERTa twitter based sentiment analysis*

For our experiments we fine-tune XLM-T model Barbieri et al., 2022, which is available in the HuggingFace repository[11] as cardiffnlp/twitter-xlm-roberta-base-sentiment. XLM-T was trained on a corpus of 198M sentiment analysis tweets in eight different languages including English and Arabic. This model was chosen to conduct our experiments based on its ability to perform NLP tasks for multilingual text based on the reviewed literature work.

[1]1 https://huggingface.co/

## 4. Experiments

### *4.1 Experiments setting*

Our Experiments were carried out using Google Colab's GPU hardware accelerator platform. The extracted features for the training set were fed to the input of the tuned models. The batch size and learning rate were set to 32 and 2e-5, respectively. The models were optimized using the ADAMW optimizer. The models were trained for five epochs.

### *4.2 Performance evaluation*

The experimental results that are obtained upon training the model on the given dataset are discussed and a comparison with the above-mentioned models is made. The models were then compared with and without using the Over sampling technique on the basis of Accuracy and F1 score on Table1.

Table 1.Experimental results of XLM-T and six other pre-trained transformer models.

| | Without oversampling | | Using oversampling | |
|---|---|---|---|---|
| Model | Accuracy (%) | F1-score (%) | Accuracy (%) | F1-score (%) |
| BERT | 69.94 | 68.56 | 82.17 | 81.92 |
| mBERT | 69.51 | 68.18 | 83.18 | 82.98 |
| Giga BERT-v4 | 73.04 | 72.38 | 84.96 | 84.83 |
| ROBERTA | 71.17 | 69.40 | 80.70 | 80.41 |
| XLM-ROBERTA | 72.73 | 71.64 | 83.64 | 83.37 |
| XLNET | 68.69 | 67.05 | 78.59 | 78.15 |
| XLM-T | 74.60 | 74.10 | 85.93 | 85.71 |

### *4.3 Discussion*

Based on the experiment results, Table1 shows that the XLM-T model achieved the highest performance on the organizer datasets when compared to the other transformers, this is because the model was originally trained on millions of tweets in over thirty languages including Arabic and English languages. Taking the imbalance of the organizer datasets problem as consideration, the Oversampling technique helped the models to improve the performance and obtain better results making the dataset more suitable. We also can conclude that the results highlight the potential of the domain-specific language model, as more suited to handle social media and specifically multilingual SA.

For the final submission, our technique based on the test data shows an F1-Score of 0.65186 and accuracy of 0.65180 for the Arabic dataset and an F1-Score of 0.6656 and accuracy of 0.6632 for the English dataset.

## 5. Conclusion

This paper introduces our approach for participating in the CERIST NLP Challenge task 1.b, aiming to identify the sentiment polarity of the organizers datasets annotated in English and Arabic languages collected from Twitter. We used in our system the Oversampling technique as a part of data pre-processing and a pre-training model based on XLM-T for classification. Our results demonstrate that multilingual domain-specific language models with the mentioned technique were able to achieve excellent performance for the sentiment analysis task on accuracy of 85.93%.

## References


Abuzayed, A., & Al-Khalifa, H., 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* (pp. 312-317).

Alammary, A. S., 2022. BERT models for Arabic text classification: a systematic review. *Applied Sciences*, *12*(11), 5720.

Alduailej, A., & Alothaim, A., 2022. AraXLNet: Pre-trained language model for sentiment analysis of Arabic. *Journal of Big Data*, *9*(1), 1-21.

AlGhamdi, M. A., & Khan, M. A., 2020. Intelligent analysis of Arabic tweets for detection of suspicious messages. *Arabian Journal for Science and Engineering*, *45*, 6021-6032.

Alotaibi, S. S., 2015. *Sentiment analysis in the Arabic language using machine learning* (Doctoral dissertation, Colorado State University).

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2021). Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *arXiv preprint arXiv:2104.12250*.

Batista, G. E., Prati, R. C., & Monard, M. C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.

Birjali, M., Kasri, M., & Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Darwish, K., & Magdy, W., 2014. Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, *7*(4), 239-342.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elouardighi, A., Maghfour, M., Hammia, H., & Aazi, F. Z., 2017. A machine Learning approach for sentiment analysis in the standard or dialectal Arabic Facebook comments. In *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)* (pp. 1-8). IEEE.

Hammad, A. A., & El-Halees, A., 2013. An approach for detecting spam in Arabic opinion reviews. *The International Arab Journal of Information Technology*, *12*.

Kechaou, Z., Wali, A., Ben Ammar, M., Karray, H., & Alimi, A. M., 2013. A novel system for video news' sentiment analysis. *Journal of Systems and Information Technology*, *15*(1), 24-44.

Khasawneh, R. T., Wahsheh, H. A., Al-Kabi, M. N., & Alsmadi, I. M., 2013. Sentiment analysis of arabic social media content: a comparative study. In *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)* (pp. 101-106). IEEE.

Lan, W., Chen, Y., Xu, W., & Ritter, A., 2020. An empirical study of pre-trained transformers for Arabic information extraction. *arXiv preprint arXiv:2004.14519*.

Liu, B., & Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.

Liu, B., 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mohammad, A. S., Qwasmeh, O., Talafha, B., Al-Ayyoub, M., Jararweh, Y., & Benkhelifa, E., 2016. An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study. In *2016 11th International conference for internet technology and secured transactions (ICITST)* (pp. 98-103). IEEE.

Nassif, A. B., Darya, A. M., & Elnagar, A. (2021). Empirical evaluation of shallow and deep learning classifiers for Arabic sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-25.

Padmaja, S., & Fatima, S. S., 2013. Opinion mining and sentiment analysis-an assessment of peoples' belief: A survey. *International Journal of Ad hoc, Sensor & Ubiquitous Computing*, *4*(1), 21.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Versley, Y., Candito, M., ... & Tounsi, L., 2010. Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages* (pp. 1-12).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N., 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)* (pp. 13-22). Springer Singapore.