Advances in Intelligent Systems Research, volume 175
2nd International Conference on Industry 4.0 and Artificial Intelligence (ICIAI 2021)

TunRoBERTa: A Tunisian Robustly Optimized BERT Approach Model for Sentiment Analysis

Chaima Antit[1,3,*], Seifeddine Mechti[1,3†], and Rim Faiz[1,2,†]
[1]Higher institute of management, 41 Av. de la Liberte, Tunis, 2000,Tunis.
[2]Carthage High Commercial Studies Institute, Rue Victor Hugo , Carthage-Présidence, 2016, Tunisia.
[3]Operational Research, Decision Making and Process Control Laboratory (Larodec), 41 Av. de la Liberté, Tunis, 2000, Tunis.
[*]Corresponding author(s). E-mail(s): chaimaantit@gmail.com
Contributing authors: Seif.mechti@isseps.usf.tn; Rim.faiz@ihec.rnu.tn
†These authors contributed equally to this work

ABSTRACT

Sentiment Analysis has grown in importance and popularity due to the proliferation of microblogging sites and the increase in posted comments, tweets, and posts, as it allows for the prediction of people's feelings, thoughts, impressions, and opinions. Sentiment analysis is regarded as one of the most active research areas in NLP. As a result, this tool has piqued the interest of marketing and business firms, government organizations, and society as a whole. Based on that, we propose a Tunisian model in this paper. A robustly optimized BERT approach was used to establish sentiment classification from the Tunisian corpus.

***Keywords**: Natural Language Processing, Sentiment Analysis, Deep Learning, Bidirectional Encoder Representations from Transformers, Robustly Optimized BERT Pretraining Approach, Tunisian Dialect, Social media networks*

1. INTRODUCTION

With the popularity gained by social media networks in recent years, Sentiment Analysis has become one of the most used tools of Natural Language Processing since it permits to detect a person's feelings, opinions and impressions. Most studies in the field of sentiment analysis are either applied on English or multilingual datasets. Actually, it has been proved that models which are trained on a specific language such as CamemBERT (French BERT model)[1] and BERTje (Dutch BERT model) [2] outperform those trained on a collection of multilingual corpora.

In this paper, we will be interested in Tunisian dialect as research for this language is still very limited. Tunisians social media users tend to express their opinions, reactions and thoughts towards major events using their local dialect (informal language) which is characterized by Code Switching (see Figure 1).

Figure 1 Extracted comments from Tunisian social media

BERT [3] has demonstrated state of the art results for eleven Natural Language Processing tasks. It permits the system to learn from input text in a bidirectional way instead of analyzing sentences in a unidirectional way. It is pre-trained on a large corpora using Masked Language Modeling (MLM) and Next Sentence Prediction(NSP). This model was later evaluated and amelio rated in the RoBERTa model introduced by the Facebook AI research team [4]. RoBERTa [4] achieves the best results on GLUE, RACE and SQuAD. The improved model could match or even outreach the achievement of all post BERT models in distinct NLP tasks especially when dealing with smaller datasets.

1.1 Contributions

The contributions of this paper are as follows:
1. We present TunRoBERTa, a trained Robustly optimized BERT pretraining approach (RoBERTa) language model for Tunisian dialect, for sentiment analysis.
2. We compare the results of TunRoBERTa model with existing models.
3. We study the effect of combining TunRoBERTa as an embedding technique with CNN.

1.2 Outline

This paper is organized as follows: Section 2 provides a literature review that examines related methods and techniques about Tunisian Sentiment Analysis. TunRoBERTa and the various steps we must take to build our model are presented in Section 3. A TunRoBERTa-CNN based sentiment analysis model was raised in section 4. Section 5 presents the evaluation section. In Section 6, we discuss the given results and we highlight some future works.

2. Related Works

Tunisia has witnessed an increase in the use of social media platforms especially after the revolution. For example, Facebook, the most popular microblogging site in Tunisia, accounted for 8 270 000 users in Tunisia in January 2021 which represents 68.6% of its entire population1. In fact, Tunisian users attempt to generate content on social networks using an informal language which mainly include: Modern Standard Arabic (MSA), Tunisian Dialect in Arabic script and Latin script and French. Tunisian Dialect Sentiment Analysis works could be divided into two types : Non-Code-Switched text and Code-Switched text.

• Based on Non-Code-Switched text:

An emotional dictionary for Tunisian Sentiment Analysis was proposed in [5] .The authors collected 60.000 political comments from Tunisian Facebook pages in order to create a new method to construct an emotional dictionary for sentiment analysis based on emoticons. In order to establish the given method they applied two steps: initial construction of emotional dictionaries then emotional dictionaries enrichment. As a result, they obtained 9 dictionaries. To evaluate their method, the authors applied a test corpus composed of 755 words manually labelled.

Messaoudi, Abir, et al.[6] proposed two deep learning models in order to establish sentiment analysis on Tunisian Romanized script. In the first model, they applied Word2vec or frWaC as initial

representation. Then, they used distinct classifiers which are CNN and Bi-LSTM followed by a fully connected layer with a softmax activation function for prediction . As for the second model, they adopted BERT multilingual followed by CNN or Bi LSTM and a 1https://napoleoncat.com/stats/facebook-users-in-tunisia/2021/01 fully connected layer with a softmax activation function for prediction. As a result, Multilingual BERT embedding's combined with CNN have outperformed other models.

• Based on Code-Switched text:

Sayadi, Karim, et al.[7] presented a new manually annotated dataset collected from twitter. The authors performed feature selection using Information Gain in order to use it while training six distinct classifiers for a binary classification and a multi-class classification. The experimental results showed that the SVM outperforms other classifiers with the set of one gram, two-grams and three-grams as features. In addition, an increase in the accuracy of different classifiers is observed when they are trained only with Tunisian Dialect features or when dealing with binary classification.

Although, a Tunisian Sentiment Analysis Corpus dataset called TSAC was introduced in [8]. The authors applied multiple machine learning techniques on several MSA and Multi-dialectal datasets including TSAC dataset in order to classify various comments written in Tunisian dialect. Best results were obtained while using the TSAC dataset.

Despite the existing research in the Sentiment Analysis field, Tunisian dialect sentiment analysis system deals with many challenges due to the limitation and lack of free available resources in this dialect. Besides, research conducted in this field for the Tunisian dialect are restricted and limited.

3. PROPOSED MODEL: TUNROBERTA

Two phases will be applied in our study in order to create our model : a pre-trained phase followed by a smaller fine-tuning phase. Our model will be pre-trained following the RoBERTa training-regime. The pre-trained phase occurs in an unsupervised way by providing a corpora of text in the Tunisian language. TunRoBERTa model is trained from scratch on a collection of Tunisian datasets using our trained tokenize which will preprocess, tokenize our corpus and train our model on dynamic Masked Language Modeling task. The Masked Language Modeling task offers our model the possibility to memorize textual patterns from our Tunisian unlabeled datasets.

Therefore, we propose a TunRoBERTa-CNN model which is a combination of two model TunRoBERTa as an embedding technique and Convolutional Neural Network (CNN).Our model TunRoBERTa will be used to convert comments to word embedding, it will be applied as a contextual embedding method. Our proposed model will map words to their indexes and representations in the embedding matrix depending on their context. Then, we will feed these representations to the CNN model in order to establish classification.

4. EVALUATION

In this section we present the datasets used to predict sentiments and we introduce our results on these corpus.

4.1 DATA SETS

Our proposed model was pretrained on 7 unlabeled Tunisian datasets publicly available. All datasets consist in social media related data:

• Tunizi dataset[2]: contains 100k comments collected from August 2019 to January 2020.
• Tunisian Dialect Corpus [3]: composed of 20k comments collected from June 2020 to October 2020.
• Tunisian Google Play Store Reviews[4]: A dataset of more than 70k reviews collected from popular Tunisian applications on Google play store. • T-HSAB[5]: Refers to a set of 6,024 Tunisian tweets posted between October 2018 and March 2019.
• TSAC[6]: extracted from January 2015 until June 2016 and containing 17k comments.
• Tunisia Social Media Corpus[7]: contains 16763 comments in the Tunisian dialect.
• Tunisian Arabic Corpus[8]: A dataset that contains 800 tweets.

All experiments were performed on the two datasets presented in Table 1. Distinct datasets were preprocessed by removing links, emoji symbols, and punctuations. They were divided into 80% train and 20% test and evaluated using different models such as: RoBERTa, CNN, CNN-LSTM, Multilingual BERT and our proposed model TunRoBERTa. Therefore, we applied our model TunRoBERTa as an embedding technique then we combined it with the Convolutional Neural Network (CNN).

Table 1 Overview of the evaluation datasets

Dataset	Size	Classes	Positive labels	Negative labels

Tunisian Dialect Corpus	20k	2	9006	11326
TSAC dataset	17k	2	8854	8199

[2]https://zindi.africa/competitions/ai4d-icompass-social-media-sentiment-analysis-for-tunisian arabizi/data
[3]https://github.com/BoulahiaAhmed/Tunisian-Dialect-Corpus
[4]https://zindi.africa/hackathons/sentiment-analysis-on-tunisian-google-play-store reviews/data
[5]https://github.com/Hala-Mulki/T-HSAB-A-Tunisian-Hate-Speech-and-Abusive-Dataset [6]https://github.com/fbougares/TSAC
[7]https://github.com/chiraz/Tunisian-social-media-corpus
[8]https://github.com/NadiaBMKarmani/Tunisian-Arabic-opinion-classification-Corpus

4.2 RESULTS

All models are trained using a Batch size=16 and Epochs=5. In order to validate previous classification results, experiments were also performed on the TSAC dataset. Table 2 reviews the sentiment classification performances for Tunisian Dialect Corpus dataset. Our proposed model TunRoBERta achieves an accuracy of 79.1% and outperforms CNN model (62.2%), RoBERTa model (71.6%), CNN-LSTM model (75.3%) and multilingual BERT (78%).

Therefore, TunRoBERTa embedding's combined with CNN demonstrates state of the art results and performs better than other models for all performance measures with an accuracy of 80.6%, a precision equal to 83.9%, a recall metrics of 80.6% and F1 measure of 81.1%.

Table 2 Tunisian Dialect Corpus dataset Results

Proposed models	**Accuracy**	**Precison**	**Recall**	**F1-measure**
CNN	0.622	0.414	0.539	0.486
RoBERTa	0.716	0.639	0.685	0.661
CNN+LSTM	0.753	0.759	0.822	0.789
m-BERT	0.780	0.710	0.767	0.737
TunRoBERTa	0.791	0.732	0.768	0.749
TunRoBERTa+CNN	0.806	0.839	0.806	0.822

Table 3 TSAC dataset Results

[8]		Proposed models				
Model	Accuracy	Model	Accuracy	Precision	Recall	F1 measure
SVM	0.77	RoBERTa	0.51	0	0	0
MLP	0.78	CNN	0.638	0.366	0.823	0.506
		CNN-LSTM	0.816	0.959	0.753	0.834
		Multilingual BERT	0.887	0.904	0.877	0.890
		TunRoBERTa	0.892	0.906	0.884	0.854
		TunRoBERTa-CNN	0.908	0.913	0.908	0.910

Table 3 displays the results of the different proposed models on the TSAC dataset. We observe that our proposed models CNN combined with LSTM, Multilingual-BERT, TunRoBERTa and TunRoBERTA combined with CNN outperforms the results of the existing research of Medhaffar et al. models [8]. Also, we can observe that TunRoBERTa-CNN model presents the best accuracy with a 90.8% compared to 89.2% scored by TunRoBERTa, 88.7% recorded by the multilingual BERT model, CNN-LSTM with an accuracy of 81.6%, CNN with a 63.8% of accuracy and last 51% scored by RoBERTa model. This is also the case for the F1 metric performances.

For the precision measure, the best performance was recorded by CNN LSTM model with 95.9% followed by TunRoBERTa-CNN (91.3%), TunRoBERTa (90.6%), Multilingual-BERT (90.4%), CNN (36.6%) and finally RoBERTa(0%). Best recall performance was presented by TunRoBERTa-CNN with a value of 90.8% proceeding ahead TunRoBERTa (88.4%), Multilingual BERT(87.7%), CNN (82.3%) CNN-LSTM model (75.3%) and RoBERTa (0%).

5. Conclusion and Future work

In this work, we tackled sentiment analysis tasks on code-switched Tunisian dialect. We proposed a Tunisian Robustly optimized BERT approach model called TunRoBERTa which outperformed Multilingual-BERT, CNN ,CNN combined with LSTM and RoBERTa. Furthermore,we combined our proposed model TunRoBERTa as an embedding technique with Convolutional Neural Networks(CNN). Subsequently, results showed that TunRoBERTa combined with CNN outperformed other models. In the future, we will apply Tun RoBERTa to various NLP tasks such as Dialect Identification, Next Sentence Prediction, Reading Comprehension Question-Answering.

References

[1] Martin, L., *et al.*: CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 7203–7219 (2020)

[2] de Vries, W., et al.: Bertje: A dutch BERT model. CoRR (2019)

[3] Devlin, J., *et al.*: 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)

[4] Liu, et al.: Roberta: A robustly optimized bert pretraining approach. CoRR (2019)

[5] Ameur, *et al.*: Exploiting emoticons to generate emotional dictionaries from facebook pages. In: Intelligent Decision Technologies 2016, pp. 39–49 (2016). Springer

[6] Messaoudi, *et al.*: Learning word representations for Tunisian sentiment analysis. In: Mediterranean Conference on Pattern Recognition and Artificial Intelligence, pp. 329–340 (2020). Springer

[7] Sayadi, *et al.*: Tunisian dialect and modern standard arabic dataset for sentiment analysis: Tunisian election context. In: Second International Conference on Arabic Computational Linguistics, ACLING, pp. 35–53 (2016)

[8] Mdhaffar, *et al.*: Sentiment analysis of Tunisian dialects: Linguistic resources and experiments. In: Third Arabic Natural Language Processing Workshop (WANLP), pp. 55–61 (2017)