

CERIST Natural Language Processing Challenge

March 29th, 2023

Hate speech detection model based on BERT for the Arabic dialects

Nourelhouda Chiker

Dr. Yahia Fares university, Medea, Algeria

Abstract

Hateful speech spread through social media has the potential to cause personal harm and suffering as well as social tension. Social media platforms, on the other hand, are unable to regulate all of the content that users post. As a result, there is a demand for automatic detection of hate speech. This demand is increased when the posts are written in complex languages, such as Arabic. This present study is dedicated to contributing to hate speech and offensive language detection tasks for Arabic dialects. This paper is about my participation on *CERIST Natural Language Processing Challenge 2022*. We propose an approach based on deep learning and a pre-trained BERT model. This approach is built by adding GRU and LSTM layers to BERT outputs. Additionally, to deal with the class imbalance issue in the dataset, two methods are proposed, the first is based on data augmentation by oversampling minority class using translation and back translation method and the second uses focal loss for training. The best results reached with focal loss training are 98.03% for accuracy and 98.02% for f1-score, and with data augmentation, 99.14% for both accuracy and f1-score.

Keywords: Hate speech; Offensive Language; Detection, Arabic, Deep Learning, BERT model, Class Unbalanced, CERIST.

1. Introduction

In social media platforms, there are uncontrollable number of comments and posts issued every second which make it impossible to trace or control the content of such platform. Therefore, social platforms are facing a problem in limiting these posts while balancing the freedom of speech. In addition, the diversity of people and their backgrounds, cultures and believes can ignite the flame of hate speech. In the other hand, each culture has its own different interpretations and characteristics of cyber-hate. So, every culture is assumed to act differently and have their own way of intervention in a manner which best suits the culture. For the Arab region, it is knowing that hate speech detection is more difficult due to the complexity of the language in terms of morphology, spelling, and dialects. This increase in usage and the openness in speech results in a public concern on existing practices in social networks. A vast number of posts that is hard and even impossible to control manually by platform owners. Hate and aggression through social networks should be rationed and regulated by policy makers and should be also countered by harnessing the power of artificial intelligence and machine learning algorithms to automate the detection of hate speech in social media. A lot of research studies have been presented in the last few years, only a few of these studies are dedicated to detect hate speech in Arabic dialects. Therefore, we contribute on hate speech and offensive language detection task by trying to design a textual hate speech detector based on deep learning algorithms and the pre-trained BERT model.

The rest of this paper is organized as follows. In Section 2 we briefly outline some previous work on hate and offensive language detection. The data is presented in Section 3, along with a description of our computational approach, features, and evaluation methodology. Results are presented in Section 4, followed by a conclusion and future perspectives in Section 5.

2. Related Work

This section reviews some earlier studies on hate speech and offensive language detection for different languages either by applying deep learning models or a transformer.

Duwairi et al., 2021, have examined the potential of CNN, CNN-LSTM, and BiLSTM-CNN deep learning networks to detect hate speech in social media, they reported three different sort of experiment: first is a binary classification with CNN they achieved 81% accuracy, second is a ternary classification CNN and BiLSTM-CNN model achieved an accuracy of 74%, last one is a multi-class classification which achieved an accuracy of 73% with CNN-LSTM and with BiLSTM-CNN models. De Paula et al., 2022, for detecting hate speech they used AraBert, AraElectra, Albert-Arabic, AraGPT2, mBert, and XLM-Roberta, six different transformer versions. They tested models built from encoder and decoder blocks as well as models that were just trained on Arabic and other languages. They also used the majority vote and highest sum ensemble approaches. In Roy et al., 2021, the authors examine the issue of determining whether or not a tweet is hateful or offensive with a pre-trained multilingual Transformer-based text encoder. They get Macro F1 score of 60.70, 53.28, and 49.74 for fine-grained, and 90.29, 81.87, and 75.40 for hate speech detection in English, German, and Hindi, respectively. In Brandes et al., 2018, a detailed dataset for religious hate speech in dialectical Arabic was presented. The dataset is made up of 6600 tweets gathered from Twitter using religious-related keywords. The authors looked at a number of features, including lexicon-based and n-gram features, as well as traditional machine learning algorithms. They have used neural networks like GRU and LSTM to test their theories. GRU had the highest estimation precision of 77%, according to the experiments. In Tang et al., 2020, they explore how to optimize BERT for multi-label sentiment analysis in code-switching text by

selecting a pre-trained BERT model, fine tuning methods and data augmentation method for unbalanced problem.

3. Data Description

CERIST* provided this dataset, which contains comments and posts written in different Arabic dialect from social networks as well as data from other datasets Hadj Ameer and Aliane, 2021. The posts and comments related to Covid-19 are the main subject of this dataset. It contains 8662 comments for train set and 2616 for test set. In the annotation, the supplied comment was either labeled as hate speech (Hateful) or not (Not hateful), which was a binary classification task. Table 1 provide the number of instances of different classification labels.

Table 1. Training set class distribution

Dataset	Hateful	Not Hateful
Comments	7676	986

4. Proposed Approach

In this section, we describe approach we created to detect hate speech. For the evaluation of these models, we utilized a different unseen testing dataset. Both training and testing datasets have been preprocessed.

4.1. Preprocessing

To obtain a practical dataset, we went through many rounds of preprocessing.

- First, we removed all the language except Arabic.
- We filtered all the emoticons.
- Replace the duplicate letters with a single instance of the letter.
- Letter normalization: unifying the letters that appear in different forms. Example we replace {أ إ ؤ ء آ ئ} with {ا}. Also remove all of the Arabic diacritics 'Tash keel' like (tashdid, Fatha, kesra ...etc.). and the elongation as (اجراءات become اجراءات)
- All non-Arabic characters, numbers, symbols, punctuation, hashtags, web addresses, diacritics are removed.
- Then we filtered both the Arabic and Darija stop words.
- Finally, deleting all the empty lines.

4.2. Proposed Solution

Creating a model that deals with hate speech detection is supposed to deal with unbalanced classes so it keeps performing better on real data. We try to solve this issue by putting two methods that have been suggested in the literature to the test.

* <http://www.nlpchallenge.cerist.dz/>

4.2.1. Training with focal loss

We use focal loss (Lin et al., 2017) suggested by different works (Lu et al., 2020) (Srivastava et al., 2018) as suitable for class imbalance problem confronted in toxicity datasets to train our proposed models. This loss is characterized by decreasing the contribution of the correctly classified instances which make loss implicitly focus in minority classes.

4.2.2. Data augmentation

For minority class data augmentation, we use translation and back translation to several languages. We first translate the comment text from its original language to one of the selected languages, and then we translate it back to its original language. If the resulted text does not match the original text, we add it to the dataset under the original comment class. To avoid possible data redundancy, we drop duplicates on the final augmented data. In literature, this strategy is used for augmenting small datasets like TRAC dataset having only 12k sample for training, Aroyehun and Gelbukh, 2018; Risch and Krestel, 2018.

4.3. Proposed approach

As a baseline, we propose to fine-tune Bert-base-multilingual-uncased3 model. We add to BERT model a dropout layer with the value 0.3 as rate having the aggregate sequence representation (i.e. [CLS] output) as input. And another dropout layer having the first dropout outputs as input and 0.3 rate. And a final layer is simply a dense layer having the number of classes as units (i.e. classification layer) Figure 1 describes this model. Then, we train the resulted BERTbase model on the training set using Early Stopping that allows us to specify 3 epochs. We use the value of $2e-5$ as learning rate (i.e. $lr = 0.00002$) with Adam optimizer as it is suggested by Sohn and Lee, 2019. Other than these hyper parameters choices, we keep the model parameters as default. For the loss function we chose to test the effectiveness of focal loss, Lin et al., 2017 on dealing with the problem of class imbalance. As it is mentioned in section (4.2.1). We use 0.25 and 2 as values for alpha and gamma parameters. These values are the same chosen in Lu et al., 2020; Srivastava et al., 2018. For the augmented dataset we use categorical cross-entropy with 'softmax' activation. Since this kind of losses calculates the error from logits.

As it is mentioned in the introduction, we propose building BERT based deep learning architectures by adding deep learning layers to BERT model. In this architecture, we were inspired by Mozafari et al., 2020 work by adding LSTMs to BERT and making the final BERT outputs of tokens as their inputs. But unlike this work we prefer to use one GRU as a first layer and one LSTM as a second layer rather than using only one, the GRU input are the final BERT outputs and the LSTM inputs are the GRU outputs. These hidden states resulted from the first GRU layer and the second LSTM layer are then concatenated and passed as input to global average and max pooling layers. We suppose that concatenating outputs from layers of different levels may construct hybrid features in the deep learning architecture, Lu et al., 2020. The outputs of pooling layers are concatenated to construct the input of the last dense layer dedicated for classification. As it is described below in the figure 2, the result of pooling layers outputs passes first by a dropout layer before being fed to the dense layer. This model was trained once with focal loss for the original dataset and once with categorical cross entropy for the augmented dataset similarly to how we handled the baseline model.

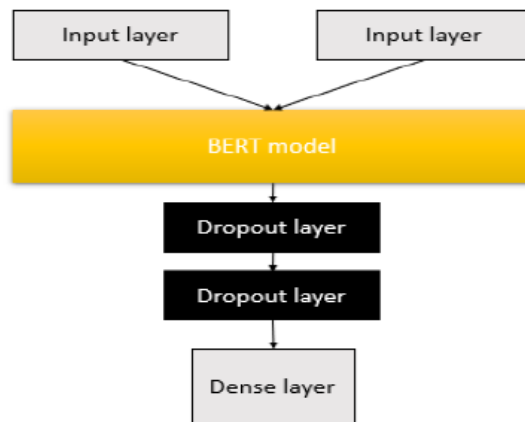


Fig. 1. BERT baseline model.

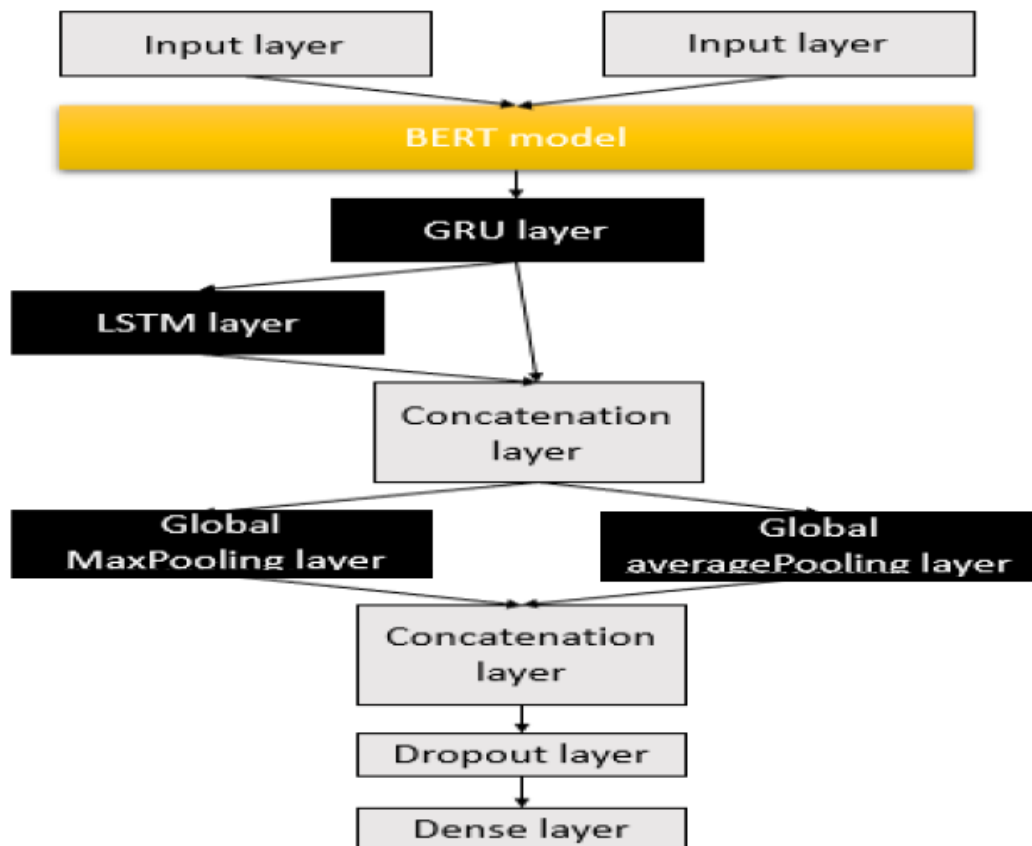


Fig. 2. Proposed model.

5. Results and Discussion

In this section, we provide the performance of our approaches on the original dataset and the augmented one. We first present the new augmented dataset, and then, the train and validation loss curves, after that, we provide the results of test step.

The comparison between the new data set's instances and the old one is shown in the following table.

Table 2. Data augmentation.

Dataset	Original dataset		Augmented dataset	
Class	Hateful	Not hateful	Hateful	Not hateful
Instances	7676	986	6320	7662

For judging the quality of learning process on the training set and checking possible overfitting, we prefer firstly expose the curves presenting the progression of loss during the three epochs of training on train and the validation sets. Figure 3 below shows the progression of loss function for BERT fine tuning and the proposed model on the original training dataset trained with focal loss.

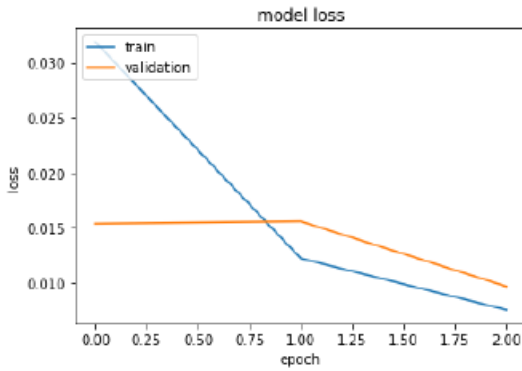


Fig. 3 . (a)

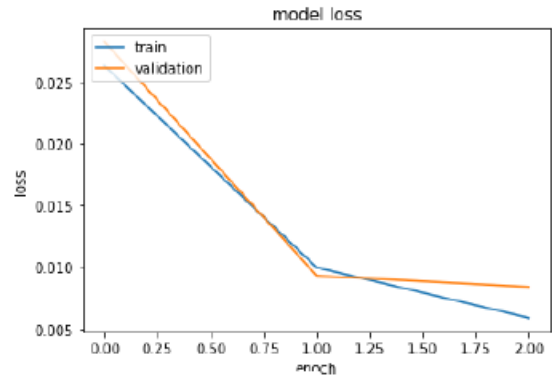


Fig. 3 . (b)

Fig. 3. Training and validation loss (a) for BERT model fine tuning and (b) for proposed BERT model both of them trained with focal loss.

The next figure 4 shows the curve of train and validation loss figure 4.1 for the baseline model on the augmented dataset and figure 4.2 for the proposed model on the augmented dataset both of them are trained with categorical cross entropy.

Fig. 4. Training and validation loss for Bert baseline and proposed BERT model trained on augmented dataset.

Now, we provide the performance evaluation results of all the models measured by using Accuracy, and F1-measure scores on both datasets.

Table 3. Test results.

Dataset	Original dataset	Augmented dataset
---------	------------------	-------------------

Models	Baseline Bert	Proposed Bert	Baseline Bert	Proposed Bert
Accuracy	87.09%	88.51%	87.09%	89.61%
F1-Score	97.04%	97.46%	97.04%	97.78%

Data augmentation followed in our approach for dealing with the problem of class imbalance caused the overfitting of the proposed models on training data. These curves on the augmented dataset do not progress similarly on the training and validation sets during epochs of training. This is related to that when we applied data augmentation technique on the training dataset, we excluded the validation and we only augmented the training set using translation and back translation strategy. This choice comes for maintain equivalent validation and test sets.

The Proposed BERT model curves present a gap starting from the 2nd epoch which means that this model has a weaker ability in avoiding overfitting caused by class imbalance in training data.

Based on the curves below, we can notice that the training on the original dataset with focal loss is better for avoiding overfitting. Since the gap produced on the third epoch seems to be more maintainable.

The data augmentation of the dataset with translation and back translation technique is not useful in our case for enhancing the performance of models.

The best model was the proposed Bert trained with focal loss which reach 98.03% accuracy and 98.02% F1-score.

6. Conclusion

In this study, we aim to enhance the automated textual hate speech detection ability and performance. We proposed an approach that takes advantage of deep learning algorithms and pre-trained BERTBASE model. In addition, to deal with class imbalance problem faced on datasets used for training and test, we investigated which solution is better between resampling the dataset, and using a loss dedicated for working on imbalanced data. Resampling is about adding instances to the minority class by using translation and back translation strategy.

References

- Aroyehun, S. T., & Gelbukh, A. (2018, August). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (pp. 90-97).
- Brandes, U., Tagarelli, A., & Reddy, C. K. (2018). ASONAM 2018: proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: Barcelona, Spain, 28-31 August, 2018.
- De Paula, A. F. M., Rosso, P., Bensalem, I., & Zaghouani, W. (2022). Upv at the Arabic hate speech 2022 shared task: Offensive language and hate speech detection using transformers and ensemble models. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection* (pp. 181-185).
- Duwairi, R., Hayajneh, A., & Quwaider, M. (2021). A deep learning framework for automatic detection of hate speech embedded in Arabic tweets. *Arabian Journal for Science and Engineering*, 46, 4001-4014.
- Hadj Ameur M.S, Aliane, H., 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. *Procedia Computer Science* 189, 232-241.2021.
- Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).
- Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K. K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurrency and Computation: Practice and Experience*, Dec. 2020, vol. 32, no. 23. doi: 10.1002/cpe.5627.
- Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8* (pp. 928-940). Springer International Publishing.

- Risch, J., & Krestel, R. (2018, August). Aggression identification using deep learning and data augmentation. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018) (pp. 150-158).
- Roy, S. G., Narayan, U., Raha, T., Abid, Z., & Varma, V. (2021). Leveraging multilingual transformers for hate speech detection. arXiv preprint arXiv:2101.03207.
- Srivastava, S., Khurana, P., & Tewari, V. (2018, August). Identifying aggression and toxicity in comments using capsule network. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018) (pp. 98-105).
- Sohn, H., & Lee, H. (2019, November). Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 551-559). IEEE.
- Tang, T., Tang, X., & Yuan, T. (2020). Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text. IEEE Access, 8, 193248-193256.