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

RIST

Information Processing at the Digital Age Journal

CERIST Natural Language Processing Challenge

Arabic Hate speech and social networks offensive language detection

Hakim Bouchal[a], Ahror BELAID[b,c]

[a]*Laboratory of Medical Informatics (LIMED), Faculty of Technology, University of Bejaia, 06000, Algeria,*
[b]*LIMED Laboratory, Faculty of Exact Sciences, University of Bejaia, 06000 Bejaia, Algeria*
[c]*Data Science & Applications Research Unit - CERIST, Bejaia, 06000, Algeria*

Abstract

The containment measures caused by the coronavirus pandemic have stimulated the use of social networks as a means of exchanging information, communication, and combating social distancing. This paper presents our participation in the NLP Challenge2022 competition initiated by RESEARCH CENTRE FOR SCIENTIFIC AND TECHNICAL INFORMATION (CERIST). The competition focuses on the task of detecting Arabic hate speech and offensive language on social networks, specifically analyzing Twitter messages related to the COVID-19 pandemic and classifying users' sentiments as either hateful or not. In the present work, we propose a model based on recurrent neural networks, more precisely the Bidirectional long-term memory (Bi-LSTM). We trained the model using a dataset constructed by the authors of this challenge. As a result, we achieves an accuracy of 96.35 %.

Keywords: Sentiment Analysis; Offensive Language Detection; Arabic Social Media; Arabic Text Classification;

1. Introduction

Natural language processing (NLP) is the ability of a computer to understand human language as it is spoken or written. Sentiment analysis and opinion mining are techniques that analyze human opinions, thoughts, emotions, beliefs, attitudes, and comments that people tend to share, Zhang and Liu, 2016, generally in the form of textual corpora on social networks. Sentiment analysis can be applies to many domains, such as marketing, health, banking, and politics, among others. For instance, it allows for the monitoring of social media to see the public opinion before taking an appropriate decision. Additionally, it facilitates the assessment and appraisal of customer preferences and perceptions concerning various services and commercial products. Sentiment analysis can take place at various levels, depending on the scope and context of the analysis, the three primary levels for sentiment analysis include, Zhang and Liu, 2016 ; Wankhade et al., 2022:

- *Document level:* Classifies the sentiment of the whole document if it expresses a positive or negative sentiment.
- *Sentence level:* Each sentence in the document is analyzed to determine if it expresses an opinion, then evaluates the polarity of the opinion.
- *Aspect level:* This level allows for a more comprehensive analysis. It consists of analyzing the feeling or opinion on specific aspects or elements of the sentence.

Social media platforms, being less controlled, makes it easy to exploit them to spread hate speech and offensive comments about individuals and groups. This can lead to emotional distress and affect the mental health of social media users. It is recommended that more measures be taken to monitor and mitigate hate speech and incitement on social media to prevent tensions that damage the entire social fabric Izsák, 2014. Therefore, automatic detection of offensive language on social media will help filter out this content before it reaches and hurts social media users. Moreover, it represents an important step towards the regulation of this type of content.

Deep learning models, particularly recurrent neural networks (RNNs) with word embaddings, improved the performance of various NLP tasks by capturing complex linguistic patterns and semantic relationships between words, making them more effective and invaluable for tasks such as text classification, sentiment analysis, and language generation.

In recent years, with the mass of textual data available on the Internet and the computational power of machines. Deep learning models, particularly recurrent neural networks (RNNs) with word embedding Jiao and Zhang, 2021, improved the performance of various NLP tasks by capturing complex linguistic patterns and semantic relationships between words, making them more effective and invaluable for tasks such as text classification, sentiment analysis, and language generation. In this paper, we seek to classify the sentiments for Arabic textual data related to the COVID-19 pandemic posted on social media.

2. Challenges in Sentiment Analysis

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

- *Ambiguity in language or Domain dependence:* This challenge refers to the fact that language is often ambiguous, and a sentence or words can have multiple interpretations based on the context and domain in which they are employed.
- *Cultural differences:* A sentence that is considered positive in one culture may be interpreted differently in another culture. Moreover, the high variety of Arabic dialect.
- *Sarcasm and irony:* A sentence or expression that evokes the opposite of what is meant, depending on the context. They are considered one of the most critical challenges.
- *Negation Detection:* Sentences considered very similar by most commonly used similarity measures, the only difference being the negation term, forcing the two sentences into opposite classes.
- *Unstructured text:* Is text that does not have a defined format or structure.

3. Related work

Sentiment analysis for the Arabic language is far behind other languages like English, whether at the sentence-level or document-level Shoukry and Rafea, 2012. Recently, it has become a very active research area, particularly for automatic hate speech detection in social media and many techniques have been proposed for automatic detection.

As for classical machine learning approaches, Shoukry and Rafea, 2012 apply Support Vector Machines (SVM) and Naive Bayes (NB) at the sentence level for Arabic tweet sentiment polarity classification. In Abozinadah and Jones, 2017, authors used a statistical learning approach for feature selection and analyze Twitter content to detect whether or not accounts are "abusive" using Support Vector Machine (SVM) as a classifier. Also Alakrot et al., 2018 utilized SVMs to classify comments which contain obscene or offensive words and phrases as either positive or negative. The authors experimented using a variety of N-gram features, word-level features, and preprocessing approaches.

In deep learning approach, Albadi et al., 2018 studies religious hate speech on Arabic Twitter and publishes the first Arabic Twitter dataset for this task. Also they introduced an Arabic lexicon consisting of terms commonly used in religious discussions. They train several classification models using lexicon-based, ngram-based, and deep learning techniques. They conclude that Gated Recurrent Units (GRU) and pre-trained word embeddings can adequately detect religious hate speech. Authors in Mohaouchane et al., 2019 compare the performance of four different neural network architectures for detecting offensive language on Arabic Social Media. Convolutional neural network (CNN), bidirectional long-term memory (Bi-LSTM), Bi-LSTM with attention mechanism and CNN combined with the LSTM architecture. They report that CNN-LSTM the gives the best results. Anezi, 2022 constructs a unique dataset consisting of 4203 manually annotated comments in seven categories: anti-religion, anti-racism, anti-gender, violent, offensive, insulting, normal positive comments, and normal negative comments. They propose an architecture based on Deep Recurrent Neural Networks for classification and detection of hate speech, they achieves 84.14%.

4. Methodology and materials

4.1. Dataset and Pre-processing

In this study, we use the dataset "Ar-hatespeech-off", which consists of tweets from various social networks, along with data from other datasets. This dataset is used for offline Arabic hate speech detection. It contains 8662 posts related to COVID-19 and they are labelled as either hateful or not. Table 1 and figure 1 illustrate the statistics and class distribution of the dataset.

Table 1. Statistics Ar-hatespeech-off dataset.

# Sentences	Vocabulary size	# Hateful classes	# Not Hateful classes
8662	30522	986	7676

The number of samples with the ’not hateful’ class is significantly higher compared to the hateful’ class. Indicating a noticeable class imbalance in the dataset. This can lead to biased results in machine learning models, as the model may become more accurate at predicting the majority class while performing less effectively with the minority class Sun et al., 2009. To address this issue, techniques such as oversampling or undersampling can be employed to balance the data set. However, it’s worth noting that collecting more data for the minority class can be a time-consuming and labor-intensive process, often requiring meticulous labeling. In this study, we have chosen to mitigate this imbalance by subsampling the majority class.

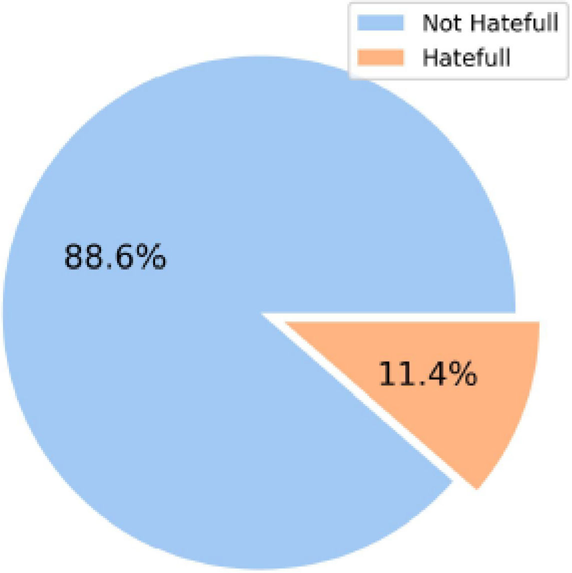

Fig. 1. The distribution of the two classes within the dataset.

A tweet consists of a collection of opinions or information expressed in various ways by different users. To obtain a refined set and reduce irrelevant content, we employ the following preprocessing steps:
- Remove all URLs (e.g. https://www.azerty.com)
- Remove all emoticons.
- Remove the dot succession (...).
- Remove targets (@username).

4.2. Feature Representation

Natural language processing models do not directly process textual input. It is essential to have a numerical representation of input text. There are many ways to get this representation, such as one-hot encoding, the count-based method (e.g. Bag-of-word, N-gram). In One-hotencodin, each word is represented by a vector with a single ’1’ in the position that corresponds to its category, and ’0s’ otherwhere. However, this dimension corresponds to the size of the vocabulary, it is not practical for a very large vocabulary, because it increases the computational cost during training, Cerda et al., 2018. In contrast, word embedding, a language modeling and feature learning technique Zhang et al., 2018, provides a dense vector representation of real values in which features are learned contextually. Therefore, words that have similar meanings will be closer together in the vector space.

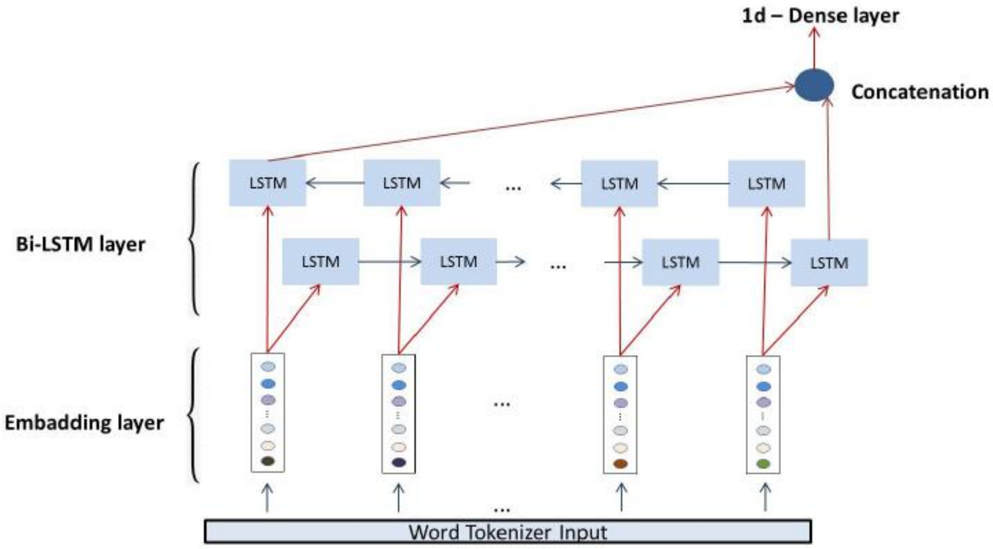

Fig. 2. Bidirectional LSTM Model.

4.3. Model

We propose to use a learning model based on recursive networks, in particular bidirectional long-term memory (Bi-LSTM), which is able to learn from previous and future context. Unlike to LSTM, Hochreiter and Schmidhuber, 1997, which depends only on previous inputs.

The first input layer consists of an embedding layer that will learn to represent words by a 16-dimensional vector, passed into a Bi-LSTM layer with dropout (rate of 0.4) to mitigate the overfetting problem. The output is then fed into a 1-dimensional dense layer with sigmoid activation. Ours network architecture is illustrated in figure below.

4.4. Experiment and Result

Table 2. The final data distribution of the 'Ar-hatespeech-off' dataset after subsampling the majority class and the performance of the proposed model.

-	Train	Validation	Test
Samples	1774	198	2166
Accuracy	99.83%	96.46%	96.35%
F1 score	99.6%	96.6%	85.82%

After applying the sub-sampling operation to the majority class, we use 10% for the validation set and the remainder for the training set. Then, the training is conducted with the following configuration: We utilize the AI framework called TensorFlow, running on Colab with a single GPU, Tesla T4 (16GB). We employ the Adam optimizer with a fixed learning rate of 0.01 and binary cross-entropy loss to train the model for 50 epochs, using a batch size of 32. The model is trained on a balanced dataset composed of 986 samples of each class. The evaluation on the test set composed of 2166 tweets gives a promising result with an accuracy of 96.35% and 85.82% F1 score, as shown in table 2.

5. Conclusion

In this study, we address the challenge of automatically detecting offensive language on Arabic social networks using the ’Ar-hatespeech-off’ dataset. We exploited the power of Bidirectional Long Short-Term Memory (Bi-LSTM) networks to process input data in both forward and backward directions, enabling us to classify tweets as hateful or not. To ensure a balanced dataset, we apply a subsampling technique. As a result, we achieve offensive language detection with an F1_score and accuracy of 85.82% and 96.35% respectively.

References

Abozinadah, E. A., & Jones Jr, J. H., 2017. A statistical learning approach to detect abusive twitter accounts. In Proceedings of the international conference on compute and data analysis. pp. 6-13.

Alakrot, A., Murray, L., & Nikolov, N. S., 2018. Towards accurate detection of offensive language in online communication in arabic. Procedia computer science, 142, 315-320.

Albadi, N., Kurdi, M., & Mishra, S., 2018. Are they our brothers? Analysis and detection of religious hate speech in the arabic twittersphere. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) pp. 69-76.

Anezi, F. Y. A., 2022. Arabic hate speech detection using deep recurrent neural networks. Applied Sciences, 12(12), 6010.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. Machine Learning, 107(8-10), 1477-1494.

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.

Hochreiter, S., & Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), 1735-1780.

Izsák, R. (2014). Report of the Special Rapporteur on minority issues.

Jiao, Q., & Zhang, S., 2021. A brief survey of word embedding and its recent development. In 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) Vol. 5. pp. 1697-1701.

Mohaouchane, H., Mourhir, A., & Nikolov, N. S., 2019. Detecting offensive language on arabic social media using deep learning. In 2019 sixth international conference on social networks analysis, management and security (SNAMS). pp. 466-471.

Shoukry, A., & Rafea, A., 2012. Sentence-level Arabic sentiment analysis. In 2012 international conference on collaboration technologies and systems (CTS) pp. 546-550.

Sun, Y., Wong, A. K., & Kamel, M. S., 2009. Classification of imbalanced data: A review. International journal of pattern recognition and artificial intelligence, 23(04), 687-719.

Wankhade, M., Rao, A. C. S., & Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780. Doi: 10.1007/s10462-022-10144-1.

Zhang, L., Liu, B., 2016. Sentiment Analysis and Opinion Mining. Springer US, 1-10. Doi: 10.1007/978-1-4899-7502-7_907-1.

Zhang, L., Wang, S., & Liu, B., 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.