Profiling Hate Speech Spreaders by Classifying Micro Texts Using BERT Model

Notebook for PAN at CLEF 2021

Alzahrani Esam[1,2], Jololian Leon[1]

[1] *University of Alabama at Birmingham, 1720 University Blvd, Birmingham, AL 35294, USA*
[2] *Al-Baha University, Alaqiq, 65799, Saudi Arabia*

Abstract

Hate speech detection has lately gained considerable attention from researchers. To profile authors effectively, we consider classifying all tweets for a specific author independently. We used BERT, the pretrained model, to classify all individual tweets for each user. Then, we added an extra layer, called a confidence layer, by which we calculate the percentage of classified hateful tweets by the model and decide whether this author is spreading hate speech or not. We found this approach simple, yet effective in determining those considered haters. Our approach achieved 77% accuracy for the Spanish test dataset and 63% accuracy for the English test dataset.

Keywords

Author profiling, hate speech, transfer learning, digital forensics

1. Introduction

The field of data mining and text analysis has recently received considerable attention. The enormous amount of text on the internet makes analyzing such content a necessity. The purposes of textual content analysis vary, e.g., sentiment analysis, cyberbullying detection, and hate speech detection. One of the major topics to be investigated in this field is hate speech spreaders profiling, which is the focus of the author's profiling task of PAN at CLEF 2021. Hate speech is sometimes mistaken as freedom of speech. However, any speech that might elicit tension between different groups of society and spread hate among them is considered hate speech. Generally, text analysis is intertwined with the field of Natural Language Processing (NLP). NLP is a broad field where humans' language choices are investigated and analyzed using AI techniques, such as machine learning and deep learning. The latter has received more attention recently, especially, after the emergence of transformers.

The goal of this task is to profile authors who are seen as hate speech spreaders on Twitter. The datasets of the PAN committee were collected from Twitter within two different languages, English and Spanish. Each dataset consists of 100 authors with 200 tweets for each author. The rest of the paper covers related previous work, dataset handling, method, results and discussion, and conclusion [1]–[3].

2. Related work

Efforts in hate speech studies vary as some researchers present annotated datasets for other researchers to work with. Corpora are provided in different languages such as Arabic, English, German, and others. English is the most frequent language used in hate speech tasks. As a result, some papers offer ready datasets along with detection techniques for different hate speech tasks [4]–[8]. The source

[1] CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: esam@uab.edu (A. 1); leon@uab.edu (A. 2)
ORCID: 0000-0002-9859-6302 (A. 1)

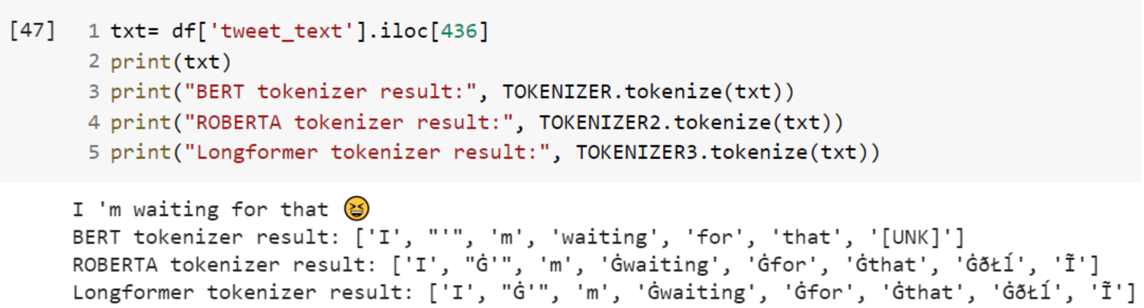
of corpora could be Facebook, Twitter, and other textual platform resources. The categorization of hate speech varies because hate speech could be directed to a certain group of people who share the same quality or characteristic. Hate speech against a religion would be, for example, Islamophobia. Another example is Cyberbullying, where some users bully others and leave a pronounced negative impact on them. Therefore, determining the type of hate speech will affect the strategy of collection, annotation, feature selection, and classification techniques.

Shared tasks are an important source of hate speech research. Usually, a dataset and an objective are shared by an organizer to encourage researchers to work on a demanding area of research. The results and approaches offered by participants provide a versatile package that represents a rich source of different methods and techniques. There are hate specific tasks that offer the above-mentioned benefits, such as HaSpeeDe [9], AMI [10], HatEval [11], OffensEval organized by SemEval, TRAC-2 [12], and MEX-A3T [13]

3. Method

As the first step in most text analysis tasks, preprocessing is an important step that can significantly affect model performance. In our method of handling the datasets, we have tried different techniques with controlling the method of classification to ensure the best use of the given datasets. We have found that the English dataset gives a better performance when hashtags, URLs, mentions, RT tags, and punctuation are removed. However, the Spanish dataset gave better accuracy when we kept the text unchanged.

Because we used the transfer learning method for our classification, we have conducted experiments about the tokenization of different transfer learning models such as BERT, ROBERTA, and Longformer. BERT uses WordPiece tokenizer1; one of the downsides of WordPiece is the lack of emojis support. Emojis are heavily used in social media contexts and they can reveal useful information that can boost model performance. On the other hand, ROBERTA and Longformer use the same tokenizer, Byte-Pair Encoding (BPE)1. BPE can handle emojis and assign a token to them. Conversely, WordPiece assigns an unknown token [UNK] to all emojis as shown in Figure1.



```
[47]  txt= df['tweet_text'].iloc[436]
      print(txt)
      print("BERT tokenizer result:", TOKENIZER.tokenize(txt))
      print("ROBERTA tokenizer result:", TOKENIZER2.tokenize(txt))
      print("Longformer tokenizer result:", TOKENIZER3.tokenize(txt))
```

```
I 'm waiting for that 😆
BERT tokenizer result: ['I', "'", 'm', 'waiting', 'for', 'that', '[UNK]']
ROBERTA tokenizer result: ['I', "Ġ'", 'm', 'Ġwaiting', 'Ġfor', 'Ġthat', 'ĠðŁĺ', 'Ĩ']
Longformer tokenizer result: ['I', "Ġ'", 'm', 'Ġwaiting', 'Ġfor', 'Ġthat', 'ĠðŁĺ', 'Ĩ']
```

Figure 1: How different tokenizers handle emojis

As mentioned before, we considered using transfer learning techniques for this task. Transfer learning has been proven to be state-of-the-art for many tasks [14]. There are different choices we could have considered, but in the interest of time and simplicity, we have considered using BERT, ROBERTA, and Longformer Huggingface implementations. With the use of Huggingface, we conducted experiments with different parameters for better accuracy. After conducting multiple experiments, we found that BERT gave us the best accuracy for the target task. For the English dataset task, we used bert-base-cased version with the following hyperparameters.

- Batch_size = 32
- Epochs= 50
- Learning rate = 2-5
- Max_length = 50

For the Spanish dataset task, we used bert-multilingual-cased with the following hyperparameters.

- Batch_size = 16
- Epochs= 50
- Learning rate = 2-5
- Max_length = 60

The hyperparameters are within the range of suggested values by BERT authors [14].
After visualizing and understanding the target datasets, we decided to handle the data differently and observe the outcome. We built three scenarios: 1) using the preprocessed dataset as is; 2) combining all tweets for each user; and 3) considering all tweets for each user to be in the same group after splitting the dataset into training, validation, and testing datasets (Fig2 shows the steps). The first scenario gave the best accuracy among all three scenarios. However, we could not test the accuracy for each user because the tweets were randomly selected and not all users have the same number of tweets. The second choice was the worst among all the scenarios as the dataset became small, which limited the ability of the model convergence and representation.

The results are based on the first scenario for both the Spanish language and the English language. The third scenario gave us the ability to test the model within 10 balanced authors, including all the tweets for all selected authors. Figure 2 illustrates the third scenario steps where all tweets for the same user are included in the assigned dataset (Training, testing, validation) with a random shuffle. This scenario allows us to add an extra layer for confidence, which calculates the number of tweets that are classified as hateful for each user. In our initial testing, we achieved 80% accuracy when we determined the threshold for the number of hateful tweets for each use at 95. If more than 95 tweets of any user were classified hateful, the user will be classified as a hate speech spreader.

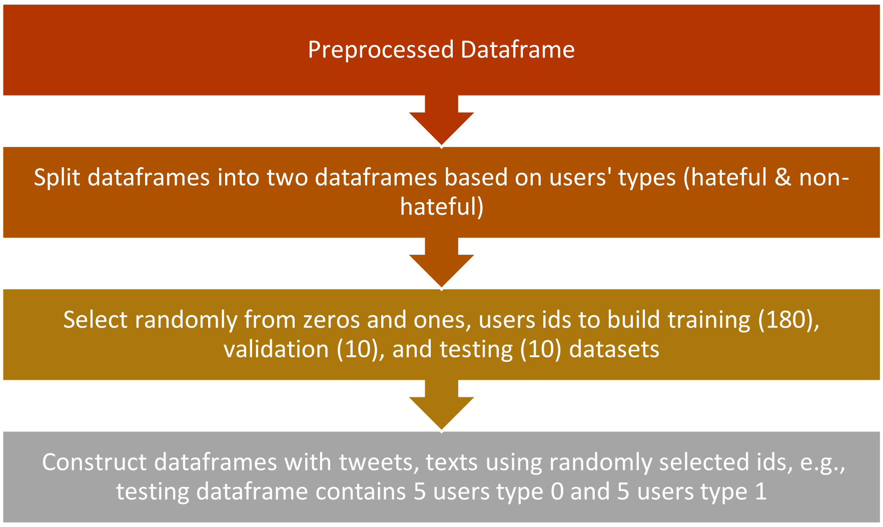

Preprocessed Dataframe
Split dataframes into two dataframes based on users' types (hateful & non-hateful)
Select randomly from zeros and ones, users ids to build training (180), validation (10), and testing (10) datasets
Construct dataframes with tweets, texts using randomly selected ids, e.g., testing dataframe contains 5 users type 0 and 5 users type 1

Figure 2: The 3rd scenario dataset splitting

4. Results and discussion

Our best results for the Spanish and English datasets were 77% and 63%, respectively. Table1 shows the results and the setup we used to achieve those results. We found that using BERT multilingual cased model with the determined hyperparameters in table 1 for the Spanish language achieved the best results among all the three models, BERT, ROBERTA, and Longformer. The experiments on the English dataset showed that using the BERT base cased model with the determined hyperparameters in table1 achieved the best results among all different conducted experiments which differ in the type of models and the values of hyperparameters. In our experiments, we tried using all three models with different hyperparameters and different preprocessing techniques. Nevertheless, the best results were achieved when we did not apply any preprocessing technique with the determined hyperparameters. In both languages, the use of cased models always yielded better accuracies. ROBERTA and Longformer tokenizers can handle emojis that exist in most tweets nowadays. The problem we faced in using both models was when they converged and became stuck at the same starting loss and accuracy. Therefore, our choice fell on BERT because it started to converge in the early epochs. However, as mentioned earlier, the BERT tokenizer cannot handle emojis which creates a failure of using an important part of the data. The shortness of tweets is another factor that contributes to the low accuracy. In this task specifically, not all hate speech spreaders' tweets are considered hateful. If a user is considered a hateful speech spreader, there might be only a few number of hateful tweets. As a result, this creates confusion for the model as the author is classified as a hate speech spreader, but not all of his/her tweets are classified as hateful. We also found that some users repeat the same tweets more than once.

Table 1
Models settings for the achieved results

Language	SPANISH	ENGLISH
SCORE	77%	63%
SCENARIO	1st	1st
PREPROCESSING	None	None
MODEL	bert-multilingual-cased	bert-base-cased
EPOCHS	50	50
BATCH SIZE	16	32
MAX LEN	60	50
LEARNING RATE	2^{-5}	2^{-5}

5. Conclusion

In this task, we considered studying the problem closely to understand the nature of the dataset and the purpose of the task. We considered using the transfer learning technique because it has been proven to perform well in NLP tasks. Even though it has been tested in the context of language modeling, recently, researchers started to apply transfer learning for classification tasks. Even though our results are near the average of all participated teams [1], we were always aiming for a better result. We believe the result did not match the sophistication of the approach we used but using transfer learning leaves you with many choices for future optimization. Due to time constraints, we provided what we have achieved to comply with the timeline provided by the organizers. However, we consider this as an ongoing learning experience which we will keep investigating. We faced some challenges with the dataset such as the limited size, the repetitive tweets or words, and the amount of noise compared to the short tweets. In some cases, the markers are unclear as some words can have different meanings depending on the context or the user's intended meaning. Tokenizers play an important role in language analysis tasks. Therefore, for future work, we would love to spend more time customizing the tokenizer we want to use to better represent the dataset. Moreover, we would consider spending more time exploring the tweets and eliminating any undesired noise. We would also consider using other pretrained models for the task.

6. References

[1] F. Rangel, G. L. D. L. P. Sarracén, Bert. Chulvi, E. Fersini, and P. Rosso, "Profiling Hate Speech Spreaders on Twitter Task at PAN 2021," in *CLEF 2021 Labs and Workshops, Notebook Papers*, 2021.

[2] J. Bevendorff *et al.*, "Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection," in *12th International Conference of the CLEF Association (CLEF 2021)*, 2021.

[3] M. Potthast, T. Gollub, M. Wiegmann, and B. Stein, "TIRA Integrated Research Architecture," in *Information Retrieval Evaluation in a Changing World*, N. Ferro and C. Peters, Eds. Berlin Heidelberg New York: Springer, 2019.

[4] S. Akhtar, V. Basile, and V. Patti, “A New Measure of Polarization in the Annotation of Hate Speech BT - AI*IA 2019 – Advances in Artificial Intelligence,” 2019, pp. 588–603.

[5] N. Albadi, M. Kurdi, and S. Mishra, “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)*, 2018, pp. 69–76, doi: 10.1109/ASONAM.2018.8508247.

[6] Y. L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, “ConaN - Counter narratives through nichesourcing: A multilingual dataset of responses to fight online hate speech,” *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 2819–2829, 2020, doi: 10.18653/v1/p19-1271.

[7] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, “Peer to peer hate: Hate speech instigators and their targets,” *12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018*, pp. 52–61, 2018.

[8] L. Gao and R. Huang, “Detecting Online Hate Speech Using Context Aware Models,” no. 2015, pp. 260–266, 2017, doi: 10.26615/978-954-452-049-6_036.

[9] C. Bosco, F. Dell’Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi, “Overview of the EVALITA 2018 Hate Speech Detection Task,” in *EVALITA@CLiC-it*, 2018.

[10] E. Fersini, P. Rosso, and M. Anzovino, “Overview of the task on automatic misogyny identification at IberEval 2018,” *CEUR Workshop Proc.*, vol. 2150, pp. 214–228, 2018.

[11] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, “Resources and benchmark corpora for hate speech detection: a systematic review,” *Lang. Resour. Eval.*, 2020, doi: 10.1007/s10579-020-09502-8.

[12] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, “Evaluating Aggression Identification in Social Media,” *Proc. Second Work. Trolling, Aggress. Cyberbullying*, no. May, pp. 1–5, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.1.

[13] M. E. Aragón, M. Álvarez-Carmona, M. Montes-Y-Gómez, H. J. Escalante, L. Villaseñor-Pineda, and D. Moctezuma, “Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets,” *CEUR Workshop Proc.*, vol. 2421, pp. 478–494, 2019.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.