

# Identify Hate Speech Spreaders on Twitter using Transformer Embeddings Features and AutoML Classifiers

Notebook for PAN at CLEF 2021

Talha Anwar<sup>1</sup>

<sup>1</sup>*Independent Researcher*

## Abstract

Hate speech against other communities, religions and countries is getting more common on social media. There is a need to control the spread of hate and offensive language on social media. Most studies identify whether a sentence is a hatred or not. This paper deals with the identification of whether a user spreads hate on Twitter or not by analyzing hundreds of tweets from the user. Feature embeddings of hundreds of tweets of a user are extracted using different transformers techniques such as BERT, BERTTweet, and RoBERTa for the English language and BETO for the Spanish language. An AutoML classifier is used to classify these embedding features. An accuracy of 75% and 85% is achieved using five-fold cross-validation and Accuracy of 72% and 82% is obtained for gold standard test data for English and Spanish, respectively. This paper secured 4th position in PAN competition.

## Keywords

hate speech, transformers, AutoML, Twitter

## 1. Introduction

Hate speech and offensive language against ethnicity, race, color, nationality, gender, sexual orientation, religion, or other characteristics are common on social media, particularly on Twitter. If a particular group raises hate Twitter trend against one specific group, the supporter of that group reply with a more intensive and offensive hate trend. This is more common among political party supporters. Recent studies also show that not only among political parties, this behavior is also observed among countries such as China received a lot of hate comments based on COVID spread from USA [1]. A lot of studies are conducted to identify the hatred, offensive tweets, and fake news [2] in different languages [3, 4] on twitter as well as youtube [5]. These studies focused on the identification of hatred tweets instead of hate spreaders. Profiling hate speech spreaders on Twitter is current challenge to identify hate spreaders on user base instead of a single tweet [6, 7]. This competition is organized at TIRA Integrated Research Architecture to maintain the essence of reproduce-ability of the results [8].

---


CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ [chtalhaar@gmail.com](mailto:chtalhaar@gmail.com) (T. Anwar)

🌐 <https://github.com/talhaanwarch> (T. Anwar)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This paper deals with profiling hate speech spreaders on Twitter using transformers embeddings as features and AutoML as classifiers. Tweet preprocessing, classification based on tweet and user are both studied this paper.

## 2. Methodology

### 2.1. Dataset

Dataset consists of English and Spanish tweets. For each language, data of 300 users is collected; 200 for training and 100 for testing. Each user data comprised of 200 tweets. The rating as hate speech or not is based on the user instead of the tweet. It meant that 200 tweets for each user are analyzed to label that user as a hate spreader.

### 2.2. Pre-processing of tweets

Text data is pre-processed to remove any noise. First digits are removed, followed by 'RT', '#USER#', and '#URL#'. Next, tweets are converted to lower case. Stop-words were not removed from the tweet.

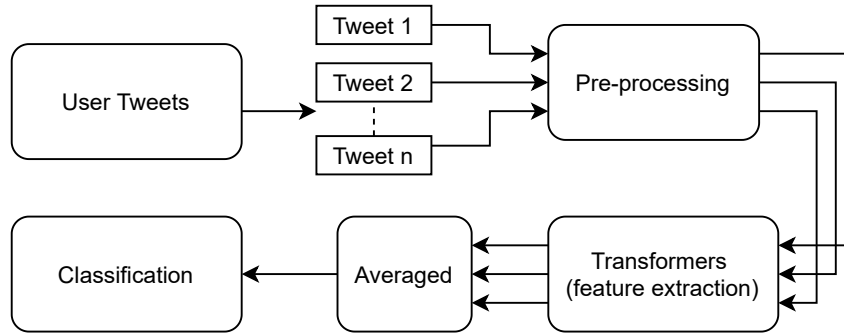
### 2.3. Tweet level training

Most NLP algorithms work best on the sentence level. So tweet level model training is used as our baseline approach. A label is assigned to each tweet from the user label. All the tweets of a user are given the label of that user. This way 40,000 tweets are obtained and split into train, validation and test at a ratio of 80:10:10. Train set consists of 32400 tweets, validation and test set comprised of 4000 and 3600 tweets, respectively. This test data is not the same as gold standard test data. This is created for internal evaluation of the system. The aim is that once the model is trained on tweet level, we will predict on tweet level for each user and then average the prediction of all tweets for a particular user and assigned it as user label. For the English baseline model, BERT base uncased model is implemented on tweet level to solve the problem. Ktrain, a lightweight wrapper of deep learning libraries, is used to implement this [9]. The maximum length of the transformer token is set to 20, batch-size to 16, number of epochs to 5. Fit-one-cycle is used to train the model with a learning rate of  $1e^{-5}$ .

### 2.4. User level training

In user-level model training, instead of training the BERT model [10] on our data, we extracted the BERT embeddings from the pre-trained BERT model. The embeddings are extracted from the last hidden layer of the BERT model. Features are also obtained by concatenating the last four layers of BERT model as mentioned in BERT paper. We want to test from which layer the better results can be achieved. These feature embeddings are extracted on tweet level and then averaged at a user level. These user-level features are then fed to autogluon tabular predictor for classification. [11]. As there are two features set one from the last layer and the other from last 4 layers, so two separate AutoML models are trained. We also used BERTTweet [12] and RoBerta [13] model to extract features of English data. In BERTTweet model preprocessing

step is changed '#USER#' is replaced with '@USER' and '#URL#' is replaced by 'HTTPURL'. Emojis are converted to text. For Spanish language, BETO (BERTSpanish) model is used [14]. HuggingFace framework is used to implement all of these transformers models.



**Figure 1:** Flow diagram of user level hate spread classification

For submission of results to competition, we used 5-fold cross validation such that gold test data (unlabeled) is evaluated in each fold and finally the prediction is weighted averaged for each model. No train, validation and test splitting as in tweet level training is applied as in user level training data is limited because of averaging.

### 3. Results

The baseline approach is based on tweet-level training resulted in over-fitting. The training accuracy achieved for English task using BERT is 85%, whereas validation accuracy is 49% and test accuracy is 53%. As initial results are not good, so further 5-fold cross-validation is not applied. For the Spanish task, train, validation and test accuracy achieved is 92%, 64% and 56% respectively.

In user-level classification, the average 5 fold accuracy obtained is 75%, 73.5%, and 70% using embeddings from the last hidden layer of BERT, BERTTweet and RoBERTa, respectively. From the last 4 hidden layers of BERT, BERTTweet and RoBERTa, the accuracy of 72.5%, 70% and 73% is achieved. For the Spanish language, BETO resulted in an accuracy of 85% and 85% from embeddings of the last hidden layer and last four hidden layers, respectively. On gold standard test data, the accuracy of 72% and 82% is obtained for English and Spanish, respectively.

### 4. Discussion

Profiling hate speech spreaders on Twitter should be based on user profiles containing multiple tweets. It cannot be identified using a single tweet of a user. So deep learning methods based on sentence/tweet classification failed to classify the user as a hate spreader or not. There needs an approach to consider all tweets of a user at once while training. One way to do is to combine all tweets of user and do documentation classification. This technique produces large text documents which are not feasible to classify as transformers' self-attention mechanism

computational expense increase quadratically. One way is to use transformers which have a linear scale able self-attention mechanism such as Longformer [15], but the issue is that in document sentence have some connection with each other. In our case, the tweets have no connection with each other, so Longformer is not a feasible option. To resolve this issue, we extracted embeddings of each tweet and average to make one embedding sequence for one user.

## 5. Conclusion

Identify hate speech spread on Twitter is a need of the day to make social media a peaceful platform. This paper proposed a feature extraction technique using transformers embeddings and AutoML classifiers to classify hate spread users on Twitter. A 5 fold validation accuracy of 75% and 85% is obtained for English and Spanish language, which reduced by 3% on gold standard test data. This paper also discussed the disadvantage of classification techniques based o tweets instead of users and also highlighted the impact of using long sequence transformers.

## 6. Acknowledgments

I would like to express my special thanks to all organizers specially Francisco Manuel Rangel Pardo, Paolo Rosso and Elisabetta Fersini

## References

- [1] L. Fan, H. Yu, Z. Yin, Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter, *Proceedings of the Association for Information Science and Technology* 57 (2020) e313.
- [2] M. S. Javed, H. Majeed, H. Mujtaba, M. O. Beg, Fake reviews classification using deep learning ensemble of shallow convolutions, *Journal of Computational Social Science* (2021) 1–20.
- [3] V. Basile, C. Bosco, E. Fersini, N. Dehora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: *13th International Workshop on Semantic Evaluation, Association for Computational Linguistics*, 2019, pp. 54–63.
- [4] T. Anwar, O. Baig, Tac at semeval-2020 task 12: Ensembling approach for multilingual offensive language identification in social media, in: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 2177–2182.
- [5] T. Tehreem, Sentiment analysis for youtube comments in roman urdu, *arXiv preprint arXiv:2102.10075* (2021).
- [6] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, 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)*, Springer, 2021.

- [7] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
- [8] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1\_5.
- [9] A. S. Maiya, ktrain: A low-code library for augmented machine learning, arXiv preprint arXiv:2004.10703 (2020). arXiv:2004.10703.
- [10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [11] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, arXiv preprint arXiv:2003.06505 (2020).
- [12] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2020).
- [13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
- [14] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
- [15] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).