Profiling Hate Speech Spreaders on Twitter

Notebook for PAN at CLEF 2021

Kumar Gourav Das[1], Buddhadeb Garai[1], Srijan Das[1] and Braja Gopal Patra[2]

[1]*Department of Computer Science & Engineering, Future Institute of Engineering & Management, Kolkata, India*
[2]*Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA*

Abstract

Hate speech on social media has led to feelings of concern and distrust among individuals and communities. This has aggravated over time because spreading hate speech is often inconsequential to its authors. Hate speech detection can help organizations identify, monitor and manage hate speech spreaders. This paper describes the systems submitted to hate speech profiling task at PAN-2021 using tweets provided by its organizers. This task aims to identify whether the author of the tweets spreads hate speech. We investigated two popular embeddings; namely TF-IDF and BERT (Bidirectional Encoder Representations from Transformers) for converting tweets into vectors. The systems using TF-IDF features with Support Vector Machine (SVM) obtained the best results of 67% and 81% for both English and Spanish tweets (test datasets), respectively. We also used Convolutional Neural Network (CNN) on features generated using BERTweet (a pre-trained language model for English tweet) and it obtained an accuracy of 66% on English tweets.

Keywords

hate speech spreaders, author profiling, BERTweet, text analysis,

1. Introduction

Social media has become an important communication medium. Social media platforms such as Facebook, Twitter, LinkedIn, and WhatsApp enable users to interact with each other by both sharing and consuming information. Information can spread widely and even viral on social media very quickly. Unfortunately, hate speech can also spread easily that not only harm individual victims but also create an adverse impact on society.

Hate speech is commonly defined as any communication that uses offensive and threatening language that targets specific groups of people basis on some characteristics such as religion, ethnicity, nationality, race, color, gender, or some other characteristics. A huge amount of user-generated content on the web and social media has given rise to a variety of challenges including the spreading and sharing of hate speech messages. Advances in Natural language processing (NLP) and text analysis techniques provide an automated way to identify the hate speech posts by analyzing social medias' posts. However, classifying users as haters or not from their posts is a challenging task.

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

kumargouravdas18@gmail.com (K. G. Das); brajagopal.cse@gmail.com (B. G. Patra)
0000-0003-2997-5314 (B. G. Patra)

In this paper, we perform hate speech spreader identification from Twitter data, provided by the organizers of PAN-2021 [1]. The organizers provided tweets of users for both languages namely English, and Spanish. The training dataset consists of data obtained from 200 users of each of English and Spanish languages, while the test dataset contains only 100 users of both languages. We used term frequency-inverse document frequency (TF-IDF) to convert both English and Spanish tweets into vectors. Support vector machine (SVM) was used for classifying the hate speech spreaders. For the English dataset, we also used a pre-trained model BERT with Convolutional neural network (CNN) for classification.

The rest of the paper is organized in the following manner. Section 2 discusses related work briefly. Section 3 describes data and provides the detailed implementation of our hate speech spreader identification system. Section 4 describes results and detailed analysis of the results. Finally, conclusions and future directions are listed in Section 5.

2. Related work

The PAN-2021 task on profiling hate speech spreaders targets on identifying such users who spread hate speech in Twitter [1, 2]. In recent years, hate speech in social media has become a complex phenomenon, whose detection has recently gained a lot of attention in the domain of NLP. There have been many research performed on identifying hate speech in English, Spanish [3] and Indonesian language [4]. “Multilingual detection of hate speech against immigrants and women in twitter” task in SemEval 2019 [3] focuses on the detection of hate speech against immigrants and women in Spanish and English tweets extracted from Twitter.

Tah et al. [5] attempted to explore the use of profane words for the identification of hate speech. The authors classified 500 YouTube comments into 8 different categories of hate speech based on profane words. In another similar work, Malmasi et al. [6] discriminated hate speech and profanity. Many other similar experiments were conducted on identifying cyberbullying [7], abusive language on social media [8, 9], fake news in Twitter [10], and profane words in hate speech [5]. Multiple publications and shared tasks are available on profiling authors based on their tweets [10, 11, 12, 13]. Profiling Fake News Spreaders on Twitter shared task in PAN-2020 [10] was organized to identify whether the author of a Twitter feed is keen to spread fake news or not.

Different feature extraction techniques that had been used for hate speech detection such as lexicons [14, 15], bag-of-words [16], N-grams [9], TFIDF [17, 18], and word embedding [19].

Different machine learning [6, 18, 20] and deep learning-based [17, 21, 22] models have been used for classification of hate speech. Gaydhani et al [23] employed different machine learning algorithms on TF-IDF features for detecting hate speech and offensive language on Twitter. In another similar work, the authors investigated the application of n-gram representation for web content classification using a machine learning algorithm [18].

The above survey reveals that a variety of features extraction and classification models have been used for hate speech identification. The combination of lexicons and ML approaches obtained better results for classifying hate speech detection [24]. It was observed that the word embedding with deep learning-based model performed well for large datasets whereas language independent features like TF-IDF provides good result also for small datasets [19].

3. Data and Methods

Fig. 1 describes the architecture of our proposed system of profiling hate speech spreader in Twitter data. As shown in this figure, our system has five key components (data pre-processing, feature extraction, data splitting, model development on train data, and classification model evaluation on test data). Each of the step is discussed in detail in the subsequent sections.

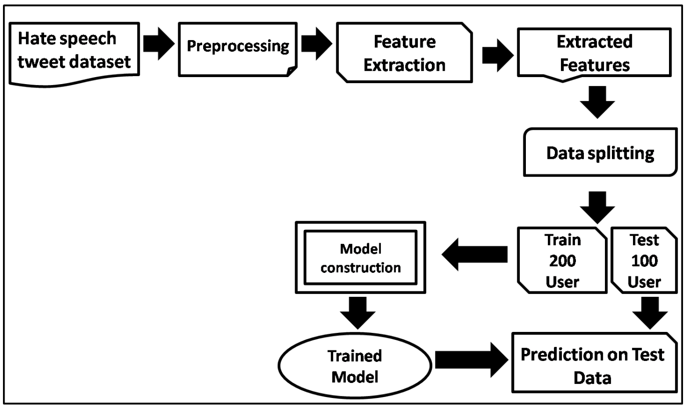

Figure 1: System Overview

3.1. Dataset

The organizers provided training datasets for both English and Spanish languages annotated with hate speech spreader or not at the user level. Both English and Spanish datasets contain tweets from 200 users and each user has 200 tweets. The organizers also provided test datasets that contain tweets from 100 users for both English and Spanish languages.

3.2. Preprocessing

We applied different preprocessing techniques to remove noisy and non-informative words from the tweets. We removed emoji from every tweet and converted it into lower case. We removed all stop words, punctuations, and junk words. We also performed tokenization and stemming. Porter stemmer was used to convert words into their root forms.

3.3. Feature Extraction

This section describes several features used in our experiments. We used several embedding techniques such as GloVe, Google News word2vec, latent semantic analysis, TF-IDF, Bidirectional Encoder Representations from Transformers (BERT) and BERTweet for our initial experiments on training data. However, we found TF-IDF and BERTweet performed better compared to others.

TF-IDF reduces the effect of less informative words that appear very frequently in data and helps to determine how relevant a given word is in a particular document. TF-IDF works by determining the relative frequency of a word in a specific document compared to the inverse proportion of that word over the entire document corpus. We used TF-IDF for both English and Spanish datasets. We performed several experiment on unigram, bigram, and trigram features, however, the system using unigram obtained the best result on training dataset.

We used BERTweet [25], a pre-trained language model for English tweets. It is trained on 850M English tweets. Due to the limitation of the max length limit of tokens in BERT (Bidirectional Encoder Representations from Transformer), we performed fine-tuning in BERTweet in order to fit 200 tweets to a single vector. Figure 2 provides word counts of every user in English data. In fact, the minimum word count was 1500 for a single user with 200 tweets. Thus, we divided the input tweets into smaller text and feed each of them into BERTweet. We experimented with different chunk sizes and the system obtained better results on training dataset when we splited 200 tweets of each user into a chunk of eight tweets (maximum of 200 word), with one tweet overlapped. For every eight tweets (200-length chunk), we extracted a representation vector from BERT of size 768 each. For each user, we obtained 29 vectors of dimension 768 each. To increase the number of instances for CNN, we treated these 29 vectors as separate user and labelled with users class. Further, we implemented the majority class voting to decide the final label.

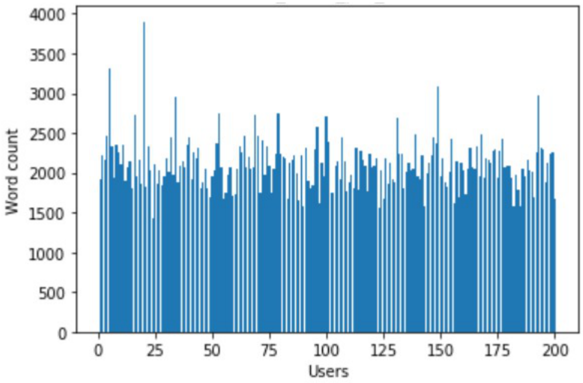

Figure 2: Wordcount statistics of the English dataset

3.4. Classifiers and Parameters

For tunning the performance of our developed systems, we divided the training data into training (60%) and validation sets (40%). A series of experiments were conducted using different machine learning and deep learning frameworks such as Naïve Bayes, Decision Tree, Random Forest, SVM, Logistic Regression, Adaboost, Long short-term memory (LSTM), Bidirectional-LSTM, and CNN. However, SVM and CNN classifier outperformed all others and we used these two classifiers for the final submission.

We used linear kernel in SVM for all our experiments. All the systems are evaluated based on accuracies. In our experiments, the CNN model was applied to BERTweet features. Hence different types of filters, padding, maxpooling, and dropouts were used according to the feature size. The best result was observed on the combined feature set where the input vector size was

768. For CNN, padding was used to convert the feature space into 28×28. Then maxpooling over 2×2 grid and a dropout of 0.5 were inputted to a ReLU activation function. Finally, softmax was performed in the output layer for classification. The deep learning models were implemented using Keras. We submitted SVM model for both English and Spanish datasets and CNN model only for the English language.

4. Results and Discussion

4.1. Results

SVM outperforms other classifiers and obtained a maximum accuracy of 67% and 81% for the English and Spanish datasets respectively. BERTweet pre-trained model and CNN classifier based system obtained an accuracy of 66% for only English dataset.

4.2. Discussion

The TF-IDF based model with SVM outperformed BERTweet based CNN model for the English dataset. Dataset provided by the PAN organizer contains only 200 users' tweet this may be the main reason for the low accuracy of the CNN based system. Deep learning systems are known as data hungry systems. For the Spanish dataset, TF-IDF based model with SVM classifier performed well as compared to other classifiers.

A total of 74 teams have participated in this task. Our system ranked 15th among 74 participants in this task at PAN with the average accuracy of 74% for both English and Spanish languages. The highest average accuracy of 79% was obtained by *SiinoDiNuovo* team across both languages. The maximum accuracies of 78% and 85% were obtained for identifying hate speech spreader in English and Spanish by *dukic* and *SiinoDiNuovo* teams, respectively.

5. Conclusion and Future Work

We presented classification systems to identify hate speech spreaders from the content of their tweets in English and Spanish languages. Among two languages, the TF-IDF with SVM based system for Spanish obtained better result than the English language.

In the future, lexicons for hate speech can be developed and can be used as features to improve the classification accuracies. It would be interesting to implement a fine grained tweet level classifier and then combine the results using majority voting. Further, external annotated data can be used to improve the accuracies. For Spanish data, pre-trained embeddings can be used in the future.

References

[1] M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, et al., Overview of pan 2021: Authorship verification,

profiling hate speech spreaders on twitter, and style change detection, in: Advances in Information Retrieval, 2021.

[2] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, Tira integrated research architecture, in: Information Retrieval Evaluation in a Changing World, Springer, 2019, pp. 123–160.

[3] V. Basile, C. Bosco, E. Fersini, N. Debora, 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] M. O. Ibrohim, I. Budi, Multi-label hate speech and abusive language detection in indonesian twitter, in: Proceedings of the Third Workshop on Abusive Language Online, 2019, pp. 46–57.

[5] P. L. Teh, C.-B. Cheng, W. M. Chee, Identifying and categorising profane words in hate speech, in: Proceedings of the 2nd International Conference on Compute and Data Analysis, 2018, pp. 65–69.

[6] S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, Journal of Experimental & Theoretical Artificial Intelligence 30 (2018) 187–202.

[7] Y. Chen, Detecting offensive language in social medias for protection of adolescent online safety, Master dissertation, The Pennsylvania State University, 2011.

[8] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proceedings of the ACM on Human-Computer Interaction 1 (2017) 1–22.

[9] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[10] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, in: CLEF, 2020.

[11] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter, Working Notes Papers of the CLEF (2018) 1–38.

[12] B. G. Patra, K. G. Das, D. Das, Multimodal author profiling for twitter - notebook for PAN at CLEF 2013, in: Notebook for PAN at CLEF, 2018.

[13] B. G. Patra, S. Banerjee, D. Das, T. Saikh, S. Bandyopadhyay, Automatic author profiling based on linguistic and stylistic features - notebook for PAN at CLEF 2013, in: Working Notes for CLEF 2013 Conference, Valencia, Spain, 2013.

[14] N. D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.

[15] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, W. Daelemans, A dictionary-based approach to racism detection in dutch social media, arXiv preprint arXiv:1608.08738 (2016).

[16] S. Malmasi, M. Zampieri, Detecting hate speech in social media, arXiv preprint arXiv:1712.06427 (2017).

[17] S. Köffer, D. M. Riehle, S. Höhenberger, J. Becker, Discussing the value of automatic hate speech detection in online debates, Multikonferenz Wirtschaftsinformatik (MKWI 2018):

Data Driven X-Turning Data in Value, Leuphana, Germany (2018).

[18] S. Liu, T. Forss, Combining n-gram based similarity analysis with sentiment analysis in web content classification., in: KDIR, 2014, pp. 530–537.

[19] M. Mozafari, R. Farahbakhsh, N. Crespi, A bert-based transfer learning approach for hate speech detection in online social media, in: International Conference on Complex Networks and Their Applications, Springer, 2019, pp. 928–940.

[20] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & internet 7 (2015) 223–242.

[21] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[22] S. Zimmerman, U. Kruschwitz, C. Fox, Improving hate speech detection with deep learning ensembles, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[23] A. Gaydhani, V. Doma, S. Kendre, L. Bhagwat, Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach, arXiv preprint arXiv:1809.08651 (2018).

[24] S. Abro, Z. A. Sarang Shaikh, S. Khan, G. Mujtaba, Z. H. Khand, Automatic hate speech detection using machine learning: A comparative study, Machine Learning 10 (????) 6.

[25] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.