

Profiling Hate Speech Spreaders on Twitter: SVM vs. Bi-LSTM

(Notebook for PAN at CLEF 2021)

Inna Vogel¹, Meghana Meghana¹

¹Fraunhofer Institute for Secure Information Technology SIT, Rheinstrasse 75, Darmstadt, 64295, Germany

Abstract

Hate speech is a crime that has been growing in recent years, especially in online communication. It can harm the individual or a group of people by targeting their conscious or unconscious intrinsic characteristics. Additionally, the psychological burden of manual moderation has necessitated the need for automated hate speech detection methods. In this notebook, we describe our profiling system to the PAN at CLEF 2021 lab “Profiling Hate Speech Spreaders on Twitter”. The aim of the task is to determine whether it is possible to identify hate speech spreaders on Twitter automatically. Our final submitted system uses character n -grams as features in combination with an SVM and achieves an overall average accuracy of 69.5% for the English and Spanish datasets. Additionally, we experimented with a Bi-LSTM model and trained it with Sentence-BERT, achieving slightly worse performance results. The experiments show that it is difficult to detect solidly hate speech spreaders on Twitter as hate speech is not only the use of profanity.

Keywords

Author Profiling, Hate Speech Spreaders, SVM, Bi-LSTM

1. Introduction

The Cambridge Dictionary defines *hate speech* as abusive or threatening speech or writing that expresses hate or prejudice towards a person or a particular group¹, especially based on ethnicity, religion, sex, or sexual orientation. Thus said, any characteristics of an individual can become the target of hate be it gender, nationality, or even educational background. The Internet and the possibility of communicating anonymously made it additionally an effective vehicle for spreading hateful and offensive content at an unprecedented rate [1]. Moreover, studies have highlighted a connection between the spread of hate speech and hate-related crimes [2]. That means, the spread of hate speech has the potential to damage our society, and cause severe harm to people or entire groups.

Currently, social media companies such as Twitter and Facebook use human annotators to manually detect hateful comments and posts². Additionally, users are encouraged to report offensive and potentially harmful content. Given the high volume of messages posted on social media websites, these methods are time-consuming, expensive, and depend on human

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

✉ inna.vogel@sit.fraunhofer.de (I. Vogel); meghana.meghana@sit.fraunhofer.de (M. Meghana)



© 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)

¹<https://dictionary.cambridge.org/de/worterbuch/englisch/hate-speech>

²<https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html>

judgment. The evident harm and volume of the uncontrolled spread of hate speech [3] and the psychological burden of manual moderation³ have necessitated the development of automated hate speech detection methods.

This problem of detecting hate speech is addressed in this year’s author profiling shared task of PAN at CLEF 2021 lab⁴ [4, 5]. Author profiling is the analysis of people’s writing in an attempt to identify demographic aspects such as age, gender, language variety, or psychographic aspects such as an author’s personality type [6, 7]. Given a Twitter feed, the final goal of this year’s challenge is to identify possible hate speech spreaders on Twitter as a first step towards preventing hate speech from being propagated among online users.

We propose two different learning experiments. Our final submitted system uses TF-IDF weighted character n -grams as features in combination with an SVM. As recurrent neural networks (RNN) can preserve sequence information over time, and thereby integrate contextual information better in classification tasks, we additionally experimented with a bidirectional LSTM (Bi-LSTM) and trained it with Sentence-BERT (SBERT), a modification of the BERT network. SBERT uses siamese and triplet network structures to derive semantically meaningful sentence embeddings [8]. Both models were trained on the PAN 2021 corpus provided by the organizers [9]. The corpus covers two languages: English (EN) and Spanish (ES). The performance of the systems is ranked by accuracy. Both models have achieved almost the same classification results. The SVM model performed slightly better than the Bi-LSTM model achieving an overall accuracy of 64% and 75% on the English and Spanish corpus, respectively (average 69.5%). The Bi-LSTM model achieved an overall average accuracy of 69%. The results show that it is not an easy task to differentiate solidly Twitter users who spread hate speech from those who for the most part follow the platform’s policies and guidelines.

In the following sections, we describe our approach for the author profiling task at PAN 2021. After a brief review of related work in Section 2, Section 3 details the Twitter data provided by the PAN 2021 organizers. Additionally, we show some key statistics observed in the tweets. Section 4 details the preprocessing steps and features used to train our models. The methodology and classification results are discussed in Section 5. The last Section 6 concludes our work.

2. Related Work

Mutanga et al. [10] investigated in their study different transformer-based methods for hate speech detection in Twitter texts. They used a publicly available multi-class hate speech corpus containing 24,783 tweets. The dataset is highly imbalanced with 77.4% of the tweets labeled as “neutral”, 16.8% as “Offensive”, and 5.8% as “Hate”. DistilBERT, a distilled version of BERT, outperformed all other trained methods such as XLNet, RoBERTa or attention-based LSTM achieving an $F1$ -score of 75%.

Kovács et al. [3] used a combination of Convolutional and Long Short-Term Memory (LSTM) neural networks to detect hate speech in social media. The model was applied to the HASOC2019 corpus and attained a macro $F1$ -score of 63%. The authors also conducted experiments with

³<https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator>

⁴PAN at CLEF 2021 “Profiling Hate Speech Spreaders on Twitter”: <https://pan.webis.de/clef21/pan21-web/author-profiling.html>

RoBERTa and FastText as feature extractors. As the training data was limited, different methods for expanding resources, such as leveraging unlabeled data or similarly labeled corpora, were explored. Their results show that classification results could be significantly increased by leveraging additional data.

A major challenge for the automatic detection of hate speech on social media is the separation between hate speech and instances of offensive language. Davidson et al. [11] first collected tweets using hate speech keywords. Crowdsourcing was used to label the tweets into the following three categories: “hate speech”, “offensive language”, and “neither”. A multi-class classifier was then trained to distinguish between the three categories. The best performing model achieved an overall $F1$ -score of 90%. However, the confusion matrix revealed that almost 40% of the hate speech tweets were misclassified.

3. Dataset and Corpus Analysis

To train our system, we used the PAN 2021 author profiling corpus⁵ proposed by Rangel et al. [9]. The corpus consists of 200 English (EN) and Spanish (ES) Twitter authors each. The tweets are stored in an XML file containing 200 tweets per author. Every tweet is stored in a `<document>` XML tag. The dataset is balanced, which means the data refers to an equal distribution of class instances. Half of the documents per language folder are authors that have been identified sharing hate speech. The other half are texts from users who may share offensive tweets but could not be identified as hate speech spreaders. Table 1 shows excerpts from the corpus⁶. Every author received an alphanumeric author-ID which is stored in a separate text file together with the corresponding class affiliation. For training and testing, we split the data in the ratio of 70/30. The gold standard can only be accessed through the TIRA [12] evaluation platform provided by the PAN organizers. The results are hidden from the participants and can only be unblinded by the organisers.

It is important to note that the classes are not predefined by the organisers. We assume that class 0 refers to hate speech spreaders. Nevertheless, since the organisers do not explicitly define classes 0 and 1, we have kept the class names as originally proposed. As can be seen in Table 1, the Twitter-specific tokens such as hashtags, URLs, and user mentions were replaced by the providers with the following placeholders: #HASHTAG#, #URL# and #USER#. The examples provided in Table 1 were chosen carefully to show that insults and profanities are used by hate speech spreaders as well as by other users. Additionally, Twitter-specific text significantly contributes to the difficulty of automatic hate speech detection, as the posts contain plenty of poorly written text and paralinguistic signals such as emoticons, @-mentions, and hashtags. Prior to feature engineering (described in Section 4), we analysed the distribution of different tokens. Table 2 shows some key insights for both languages.

We observed the distribution of specific tokens to see whether we could use these for the features engineering process. Unfortunately, we could not spot any significant differences between the classes. Therefore, to train our model, we did not use features mentioned in Table 2.

⁵<https://zenodo.org/record/4603578#.YKZKqKgzZaQ>

⁶The selected tweets are used for demonstration and research purposes only and do not reflect the opinion of the authors.

Table 1

English (EN) and Spanish (ES) excerpts from the PAN 2021 “Hate Speech Spreaders on Twitter” data.

Class 0 Tweets (EN & ES)	Class 1 Tweets (EN & ES)
“#USER# #USER# Trump, that mother-fucker is guilty of cowardice while being Commander-in-Chief #HASHTAG#.”	“Kappa They gon be beating my fodder ninjas asses weak ass punks and i wont even be laughing on the outside :-)”
“RT #USER#: If a nigga taking care of me i’m fasho taking care of him. it’s really that simple.”	“Shut your fucking mouth i have no ill will towards Kaep but he’s not even close lmao #URL#”
“RT #USER#: Celebrities are so useless and corny B*tch what the fuck does this even mean?”	“#USER# All the people shit talkin this are trippin, i’d pipe tf out if an old lady if she was payin for all my shit”
“#USER# #USER# Mordes la mano de quien de da. De comer eres un cancer para nuestro pais #URL#”	“#USER# Pos pa tu tierra sucnormal hi-jadeputa”
“Los varones opinando sobre el feminismo #HASHTAG#. Nos sorprende? No nos sorprende”	“#USER# Ostia tio que palo metió el jodido”
“Que pinches perras ganas de estar cogiendo con Ale”	“RT #USER#: Qué horror. Condenado a 15 años de prisión por dejar embarazada a su hija tras un año de violaciones #URL#”

Table 2

Feature distribution of the PAN 2021 “Hate Speech Spreaders” dataset

Features	English		Spanish	
	Class 0	Class 1	Class 0	Class 1
Unique Tokens	20,280	19,298	28,806	28,761
Emojis Total	8,465	7,201	7,942	7,949
Emojis Unique	531	540	546	449
Uppercased Tokens Total	44,316	42,135	34,172	41,950
Uppercased Phrases Total	1,026	1,243	1,792	1,871
#URL# Token	8,556	6,759	5,865	6,897
#HASHTAG# Token	3,644	3,290	1,864	1,658
#USER# Token	17,250	17,585	16,014	22,088
Retweets (RT)	7,731	6,159	6,824	7,084

4. Preprocessing and Feature Extraction

The preprocessing pipeline to clean and structure the data was performed for both languages (EN and ES) and models as follows:

- The text from the original XML document was extracted and all 200 tweets per author were concatenated to one text.
- The white-space between the tokens has been reduced to a single space.
- The placeholders #USER#, #URL#, #HASHTAG#, and RT were removed.

- HTML characters were converted to Unicode characters (e.g.: “>”, “<”, “&” to “>”, “<”, “&”).
- Emojis were converted to text format by using Python’s `emoji` library.
- The text was lowercased.
- Irrelevant signs, e.g. “+”, “*”, “/” were deleted.
- Alphanumeric tokens were separated (e.g. “Berlin2018” to “Berlin 2018”).
- Sequences of repeated characters with a length greater than three were normalized to a maximum of two letters (e.g. “LOOOOOOOOL” to “LOOL”).
- Words with less than three characters were ignored (except for the Bi-LSTM model for the English language).
- Stopwords were deleted (except for the Bi-LSTM model for the English language).
- As the last step, we lemmatized the English tweets for the TF-IDF character n -gram SVM model using `wordnetlemmatizer`.

Besides the different preprocessing steps, we also experimented with different vectorization techniques and hyperparameter tuning by employing scikit-learn’s grid search function. The hyperparameters were tuned separately for English and Spanish. We experimented with emotional signals and lists of hate words as handcrafted features as well as with automatically learned features. The best results were achieved by using Scikit-learn’s term frequency-inverse document frequency (TF-IDF) vectorizer and Sentence-BERT (SBERT), a BERT model modification that uses siamese and triplet network structures to generate semantically meaningful sentence embeddings [8]. For the English language, we used the sentence transformer model `stsb-distilbert-base`⁷ and for Spanish `distiluse-base-multilingual-cased-v1`, a multilingual knowledge distilled version of multilingual Universal Sentence Encoder [13]. The models were trained with a maximum of 200 sentences per author, based on the 200 tweets per author and file.

For the SVM model, we employed TF-IDF weighted character n -grams. In English, the best results were achieved using a maximum of 1,250 features (`min_df=5`) and character n -grams with range [3;7]. For Spanish, we used top 2,350 features (`min_df=5`) and character n -grams with range [2;7].

5. Methodology

We defined this year’s PAN author profiling task “Hate Speech Spreaders on Twitter” as a binary classification problem. For each language (EN and ES) we trained two different models. We tested different features and vectorization techniques with a Support Vector Machine (SVM). Additionally, we experimented with bidirectional LSTM (Bi-LSTM) models as recurrent neural networks (RNN) have shown that they can preserve sequence information over time and thereby integrate contextual information in classification tasks.

For the final SVM model, we trained a linear kernel and set the penalty parameter $C=10$ for the English data. For the Spanish corpus, we trained the SVM with the radial basis function

⁷<https://huggingface.co/sentence-transformers/stsb-distilbert-base>

Table 3

Hyperparameters for the Bi-LSTM model

Bi-LSTM	English	Spanish
Bi-LSTM layer memory units	32	7
Dropout	0.2	0
First dense layer memory units	32	7
Activation function	ReLU	ReLU
Second dense layer memory units	24	5
Activation function	ReLU	ReLU
Activation function in output layer	Sigmoid	Sigmoid
Loss Function	Binary crossentropy	Binary crossentropy
Optimizer	Adam	Adam

Table 4

Accuracy (Acc.) scores of the final systems on the official PAN 2021 test dataset on Tira

Model	Features	Language	Acc.	Av. Acc.
SVM	TF-IDF char n -grams [3;7], 1,250 features	EN	64%	69.5%
SVM	TF-IDF char n -grams [2;7], 2,350 features	ES	75%	
Bi-LSMT	SBERT stsb-distilbert-base	EN	59%	69%
Bi-LSTM	SBERT distiluse-base-multilingual-cased-v1	ES	79%	

kernel (RBF) and $C=5$. The performance was ranked by accuracy. Table 4 shows the scores for our final system performed on the official PAN 2021 test set on the TIRA platform [12]. Accuracy scores are calculated individually for each language by discriminating between two classes. Each model was trained on 70% of the training data provided by the organizers. On the remaining 30% split hyperparameters were tuned. The highest accuracy on the test set using SVM with TF-IDF weighted character n -grams was 64% for the English dataset and 75% for the Spanish dataset. The accuracy dropped to 59% for the English dataset using Bi-LSTM in combination with SBERT, while it increased by 4% achieving 79% accuracy on the Spanish dataset. Therefore, we submitted the SVM model as our final hate speech detection system as it achieved an overall average accuracy of 69.5% performing slightly better than the Bi-LSTM model which achieved an average accuracy of 69% for both languages. The final accuracy scores of both systems are listed in Table 4. To make our Bi-LSTM model reproducible, we have listed all hyperparameters used to train the Bi-LSTM model in Table 3.

6. Discussion and Conclusion

In this paper, we described our participation in the PAN 2021 author profiling task. The goal was to develop a system that can detect Twitter users who spread hate speech on a regular basis. First, we observed the distribution of specific tokens in the tweets like the usage of emojis or user mentions to see whether we could use these for the feature engineering process. Unfortunately, we could not spot any significant differences between the two classes. Furthermore, we experimented with emotional signals and dictionaries listing hate words as

handcrafted features in addition to automatically learned features. In relation to this, we could not detect any difference in emotions between the two classes and have shown that insults and profanities are not a discriminative features of hate speech spreaders and other users.

Our final submitted system uses an SVM with TF-IDF weighted character n -grams. This model performed best for the English language. To detect hate speech spreaders in Spanish tweets, a bidirectional LSTM (Bi-LSTM) trained with Sentence-BERT achieved better classification results. The SVM model achieved an average accuracy of 69.5% for both languages which is slightly better than the Bi-LSTM model (69%).

The experiments show that it is challenging to detect hate speech spreaders on Twitter. It is challenging in different ways. First, we have shown that insults and profanities are not only used by hate speech spreaders, but also by users who do not offend other individuals or groups. Additionally, Twitter posts contain plenty of poorly written text (spelling mistakes, abbreviations, etc.) and paralinguistic signals such as emoticons, @-mentions, and hashtags. In the future, we want to make the classification results interpretable to analyse how hate words and the context in which they are expressed contribute to the classification.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and under grant agreement "Lernlabor Cybersicherheit" (LLCS) for cyber security research and training.

References

- [1] M. Mohiyaddeen, S. Siddiqui, Automatic hate speech detection: A literature review, *International Journal of Engineering and Management Research* 11 (2021) 116–121. URL: <https://www.ijemr.net/ojs/index.php/ojs/article/view/766>. doi:10.31033/ijemr.11.2.17.
- [2] S. Agarwal, A. Sureka, Using knn and svm based one-class classifier for detecting online radicalization on twitter, in: R. Natarajan, G. Barua, M. R. Patra (Eds.), *Distributed Computing and Internet Technology*, Springer International Publishing, Cham, 2015, pp. 431–442.
- [3] G. Kovács, P. Alonso, R. Saini, Challenges of hate speech detection in social media, *SN Computer Science* 2 (2021). doi:10.1007/s42979-021-00457-3.
- [4] J. Bevendorff, B. Chulvi, G. L. 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.
- [5] F. Rangel, G. L. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling hate speech spreaders on twitter task at pan 2021, in: A. J. M. M. F. P. Guglielmo Faggioli, Nicola Ferro (Ed.), *CLEF 2021 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2021.

- [6] P. Rosso, F. Rangel Pardo, Author profiling tracks at fire, *SN Computer Science* 1 (2020). doi:[10.1007/s42979-020-0073-1](https://doi.org/10.1007/s42979-020-0073-1).
- [7] C. A. Russell, B. H. Miller, Profile of a terrorist, *Studies in conflict & terrorism* 1 (1977) 17–34. URL: <https://doi.org/10.1080/10576107708435394>.
arXiv:<https://doi.org/10.1080/10576107708435394>.
- [8] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://www.aclweb.org/anthology/D19-1410>.
doi:[10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410).
- [9] F. Rangel, G. L. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling hate speech spreaders on twitter, 2021. URL: <https://doi.org/10.5281/zenodo.4603578>.
doi:[10.5281/zenodo.4603578](https://doi.org/10.5281/zenodo.4603578).
- [10] R. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, *International Journal of Advanced Computer Science and Applications* 11 (2020). doi:[10.14569/IJACSA.2020.0110972](https://doi.org/10.14569/IJACSA.2020.0110972).
- [11] T. Davidson, D. Warmusley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.
- [12] 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](https://doi.org/10.1007/978-3-030-22948-1_5).
- [13] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *ArXiv abs/1910.01108* (2019).