Detecting Hate Speech Spreaders on Twitter using LSTM and BERT in English and Spanish

Notebook for PAN at CLEF 2021

Moshe Uzan[1], Yaakov HaCohen-Kerner[2]

[1] *Computer Science Department, Bar Ilan University, Ramat-Gan 5290002, Israel*
[2] *Computer Science Department, Jerusalem College of Technology (Lev Academic Center), Jerusalem 9116001, Israel*

Abstract

In this paper, we describe our submissions for PAN at CLEF 2021 contest. We tackled the subtask "Profiling Hate Speech Spreaders on Twitter". We developed different models for English and Spanish languages, using classic machine learning methods like Support Vector Classifier, Multi-Layer Perceptron, Logistic Regression, Random Forest, Ada-Boost Classifier and K-Neighbors Classifier to more recent deep learning methods like BERT and Bidirectional LSTM.

Keywords

Author Profiling · Hate Speech · Twitter · Spanish · English · BERT · LSTM · Logistic Regression · SVM · MLP · Random Forest · Ada-Boost Classifier

1. Introduction

In recent years, with the increasing use of social media, we have seen an increase in the spread of hateful content. Indeed the anonymity given by these social media allow any user to post what he or she wants without having to fear about consequences. This bullying, trolling, and harassment content can be very serious, in several cases might lead to suicide of the victim[1]. Following various pressures, the companies concerned are looking for more and more efficient solutions to deal with this problem. Considering the huge quantity of text posted every day, the need of an automatic and scalable detection system become a priority. The use of machine learning (ML) and natural language processing (NLP) solutions to find this offensive content has been surprisingly useful. Still, the detection of offensive language from social media is not an easy research problem due to the different levels of ambiguities present in natural language and the noisy nature of social media language. In addition, social media subscribers come from linguistically diverse communities. PAN at CLEF 2021 with "Profiling Hate Speech Spreaders on Twitter" [1, 2] deals with the detection of hate speech spreaders in two languages English and Spanish meaning that classification need to be done at the user level and not at the post level. The submission was done using TIRA automates software submission [3].

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

uzanmacho@gmail.com (M. Uzan); kerner@jct.ac.il (Y. HaCohen-Kerner)

[1] https://en.wikipedia.org/wiki/List_of_suicides_that_have_been_attributed_to_bullying

2. Background

Early works [4, 5] referred to hate speech as *abusive* and *hostile* messages or *flames.* Recent authors [6, 7, 8] preferred to employ the term *cyberbullying*. However, more terms related to hate speech are often used in the NLP community, such as: *discrimination, flaming, abusive language, profanity, toxic language* or *comment* [9]. But, in defining this phenomenon, the words *hate speech* tends to be used the most [10].

Identifying if a text contains hateful language is not an easy task, even not for humans. However, there is not one formal definition of hate speech, a common definition is given by [11] as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic [9, 10, 12, 13, 14, 15]. Some examples are given by Biere et al. [10] and de Gilbert et al. [13]:

1. *God bless them all, to hell with the black.*
2. *Wipe out the Jews.*
3. *Women are like grass, they need to be beaten/cut regularly.*

Fortuna and Nunes [16] noted in their survey paper that for hate speech detection the most used approach is the supervised one with a focus on support vector machines (SVM) ([17], [18], [19]) followed by Random Forests [20], and Decision Trees [21]. Schmidt and Wiegand [9] found that recurrent neural networks (RNN) are also very common [22].
Badjatiya et al. [23] proposed a deep learning approach and obtained very good results using word embeddings. Zampieri et al. [24] showed that n-grams can perform well for hate speech detection using SVMs with different surface-level features, such as surface n-grams, word skip-grams, and word representation n-grams induced with Brown clustering. They also noticed that these features reached their limits for more complex tasks, e.g., distinguishing profanity and hate speech. In such tasks, more in-depth linguistic characteristics may be required. But with the recent arrival of attention mechanism [25] and Transfomers [26] in NLP and especially with the development of language representation like BERT [27].

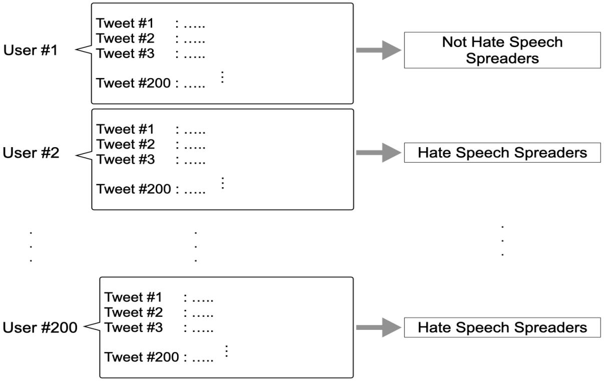
Schmidt and Wiegand [9] noted that in addition to the absence of conventional terminology issue, mentioned above, the lack of common datasets, to conduct research on it, is a challenging obstacle to progress in this area. Indeed making judgements about the general effectiveness or non-effectiveness of research conducted on various datasets can be inconsistent. For better consistency and comparability of different features and developed methods, they argue for a benchmark datasets for hate speech detection. This is the approach suggested by competition such as PAN at CLEF 2021 [2] which provide the same dataset to all the participants and publish the method and the results of each participant method of detection according to this benchmark dataset.

3. Experimental Results and Submitted Models

3.1. Task dataset

PAN at CLEF 2021 with the subtask “Profiling Hate Speech Spreaders on Twitter” proposed an original task by asking a model that classify a user to hate speach spreader instead of predicting if a post is hateful. For each user we were given 200 tweets and we need to classify it as hate speach spreader or not. The complexity of the task follows from the fact that only 200 users tweets was given as training set meaning we 200 cluster of 200 tweets and a label for each cluster. This task must be performed on 2 languages English and Spanish increasing the difficulty since models giving good results in one language will give less good ones in the other.

Figure 1: PAN at CLEF 2021: “Profiling Hate Speech Spreaders on Twitter” dataset.



3.2. Basic models

First, we split the tweets written 200 users to train and validation set with 20 percent of the given data what give us 160 labeled users for the train and 40 ones for validation (with 200 tweets for each user). Like in our precedent work [28], we began with basic model like Support Vector Classifier (SVC), Multi-Layer Perceptron (MLP) or Logistic Regression but also more sophisticated one like Random Forest (RF), Ada-Boost Classifier (ABF) and K-Neighbors Classifier (KN) using classical feature like char ngram features and word ngram features. Some model gave us very good accuracy but given that the dataset is relatively small this was not representative. So we retry this experiments using 10 cross-fold validation. We get less good result but it seems more representative.

3.3. Deep learning models

We realized that using basic model can lead in a significantly lower accuracy on the test set compared to its cross-validation results so we try going beyond and experiment more deep approaches using Bert as language representation of tweets. We used pretrained Bert model and in English we used [27] and for spanish we used [29]. For each tweet we get the corresponding BERT representation. From there, we tried different method.

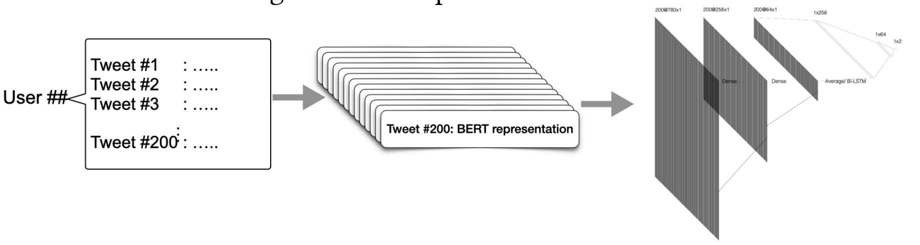
Table 1: Accuracy results of our first models.

Language	**ML Method**	**Features**	**Result**
English	RF	2000 char 5-grams	0.665
	RF	1,000 char 4-grams	0.655
	ABF	20,000 char 4-grams	0.64
	Majority Baseline		**0.50**
Spanish	RF	17,000 word 1-grams	0.81
	LR	1,500 char 4-grams 2,500 word 1-grams	0.78
	ABF	3,500 word 1-grams	0.775
	Majority Baseline		**0.50**

First one was by feding into two successive relu-activated dense layers first with 256 out-features and second with 64 out-features the 200 representated tweets. After that we obtain one 64 vector using a mean operation on 200 vectors. Finally we have a relu-activated dense layers that classify this to hateful or not.

The second model we developped take the 200 Bert representation vectors and fed them into a Bi-LSTM with 2 x 32 features in hidden layer. Finally we have a relu-activated dense layers that classify the 64 feature output to hateful or not. We use Adamw[30] variant of Adam [31] algorithm as optimizer for each model [2]. After our first submission we noticed that there was a rather big gap in English between the result obtained on our development set and the final result on the test set. So we decided to increase dropout rates and use a BERT model that had been trained on tweets [32].

Figure 2: Description of our models.



3.4. Experimental Results

Firstly, we submit two models the averaging one for English and the one using LSTM for Spanish. For the English we get an accuracy of 0.70 in our splitted set and 81 for the Spanish one. We get an accuracy of 0.62 in English and 0.70 in Spanish giving an overall accuracy of 0.66. We then submitted second time two model for the English one we keep the same but for Spanish we switch to the averaging model with different training parameter. Surprisingly, the final results

[2] For more precise details about dropout or batch used we publish the code in github https://github.com/machouz/pan_transformers

[3]. showed that, contrary to our observations, traditional methods give very good results (see Table 2). The best result was obtened by SiinoDiNuovo getting an accuracy of 0.73 in English and 85 in Spanish. We tied for 43rd with this result.

Table 2: Accuracy results of baselines (in bold) and submitted models.

Model	English Accuracy		Spanish Accuracy		Average	
	Dev	Test	Dev	Test	Dev	Test
SiinoDiNuovo		0.73		0.85		0.79
char nGrams+Logistic		**0.69**		**0.83**		**0.76**
AveragingBERT	0.72	0.62	0.87	0.76	0.795	0.69
MBERT-LSTM		**0.59**		**0.75**		**0.67**
Bi-LSTM-BERT	0.62	0.44	0.81	0.74	0.715	0.59
TFIDF-LSTM		**0.61**		**0.51**		**0.56**

4. Conclusions and Future Work

In this paper, we described the submitted models for the Profiling Hate Speech Spreaders on Twitter task at PAN 2021. Originally, we looked at a number of machine learning models using basic features. However, we finally turned to more deep learning models. These deep learning models generally do well in the tasks to which they are submitted and this is what we observed through our research. Our final model consist of using Bert as language representation, and Average or LSTM to make the classification. The difficulty here was to deal with the limited amount of given data. Our overall accuracy in our first submission was 69. Classifying a tweet post still remain a difficult task considering Twitter-style informal written genres.

Many tweets contain acronyms that can be presented in different forms. These acronyms can lead to ambiguity. Future research may look for other ways to lessen this ambiguity. Acronym disambiguation [33], will extend and enrich the tweet's text and might enable better classification. We also suggest examining the usefulness of skip character n-grams because they serve as generalized ngrams that allow us to overcome problems such as noise and sparse data [34]. Other ideas that may lead to better classification are to use stylistic feature sets [35], key phrases [36], and summaries [37].

Final result shows that more traditional methods may turn out more relevant. These methods can be combined with k-fold cross-validation (see [38]), especially when, like in this contest, available data is limited [28].

Acknowledgments

We thank the Bar-Ilan Data Science Institute for kindly providing server for training our models. Without their support, this research would not have been possible. We are also grateful to the organizers and reviewers who gave us the opportunity to do this research.

[3] To see the whole table of results https://pan.webis.de/clef21/pan21-web/author-profiling.html#results

References

[1] 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.

[2] 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.

[3] 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.

[4] E. Spertus, Smokey: Automatic recognition of hostile messages, in: Aaai/iaai, 1997, pp. 1058–1065.

[5] D. Kaufer, Flaming: A white paper, Department of English, Carnegie Mellon University, Retrieved July 20 (2000) 2012.

[6] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, Learning from bullying traces in social media, in: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2012, pp. 656–666.

[7] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, S. Mishra, Detection of cyberbullying incidents on the instagram social network, arXiv preprint arXiv:1503.03909 (2015).

[8] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, C. Caragea, Content-driven detection of cyberbullying on the instagram social network., in: IJCAI, 2016, pp. 3952–3958.

[9] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International workshop on natural language processing for social media, 2017, pp. 1–10.

[10] S. Biere, S. Bhulai, M. B. Analytics, Hate speech detection using natural language processing techniques, Master Business AnalyticsDepartment of Mathematics Faculty of Science (2018).

[11] J. T. Nockleby, Hate speech, Encyclopedia of the American constitution 3 (2000) 1277–1279.

[12] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: European semantic web conference, Springer, 2018, pp. 745–760.

[13] O. de Gibert, N. Perez, A. García-Pablos, M. Cuadros, Hate speech dataset from a white supremacy forum, arXiv preprint arXiv:1809.04444 (2018).

[14] 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.

[15] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long

tail on twitter, Semantic Web 10 (2019) 925–945.

[16] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

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

[18] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Eleventh international aaai conference on web and social media, 2017.

[19] D. Robinson, Z. Zhang, J. Tepper, Hate speech detection on twitter: feature engineering vs feature selection, in: European Semantic Web Conference, Springer, 2018, pp. 46–49.

[20] P. Burnap, M. L. Williams, Us and them: identifying cyber hate on twitter across multiple protected characteristics, EPJ Data science 5 (2016) 11.

[21] P. Burnap, M. L. Williams, Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making (2014).

[22] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deep learning for user comment moderation, arXiv preprint arXiv:1705.09993 (2017).

[23] 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.

[24] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, arXiv preprint arXiv:1902.09666 (2019).

[25] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014. arXiv:1409.0473.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.

[28] M. Uzan, Y. HaCohen-Kerner, Jct at semeval-2020 task 12: Offensive language detection in tweets using preprocessing methods, character and word n-grams, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 2017–2022.

[29] 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.

[30] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[32] 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.

[33] Y. HaCohen-Kerner, H. Beck, E. Yehudai, D. Mughaz, Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin, Applied Artificial Intelligence 24 (2010) 847–862.

[34] Y. HaCohen-Kerner, Z. Ido, R. Ya'akobov, Stance classification of tweets using skip char

ngrams, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 266–278.

[35] Y. HaCohen-Kerner, A. Kass, A. Peretz, Haads: A hebrew aramaic abbreviation disambiguation system, Journal of the American Society for Information Science and Technology 61 (2010) 1923–1932.

[36] Y. HaCohen-Kerner, I. Stern, D. Korkus, E. Fredj, Automatic machine learning of keyphrase extraction from short html documents written in hebrew, Cybernetics and Systems: An International Journal 38 (2007) 1–21.

[37] Y. HaCohen-Kerner, E. Malin, I. Chasson, Summarization of jewish law articles in hebrew., in: CAINE, 2003, pp. 172–177.

[38] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, Journal of machine learning research 5 (2004) 1089–1105.

A. Online Resources

The sources for this work are available via

- GitHub,