

# University of Regensburg @ PAN: Profiling Hate Speech Spreaders on Twitter

Notebook for PAN at CLEF 2021

Kwabena Odam Akomeah, Udo Kruschwitz and Bernd Ludwig

University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

## Abstract

This paper reports on our approach to addressing the Shared Task *Profiling hate speech spreaders on Twitter* for both English and Spanish, organised as part of the PAN@CLEF 2021 Challenge. We submitted one run for each language based on pre-trained language models. For English we fine-tuned a BERT-model while for Spanish we used a language-agnostic BERT-based sentence embedding model without fine-tuning. The second approach appears to have been a lot more effective than the first one. Given the simplicity of the approaches there is plenty of room for future directions based on the architectures adopted here.

## Keywords

Hate Speech, BERT, Embeddings, Sentence Encoder

## 1. Introduction

Hate speech is not a new phenomenon but it has become more and more of a problem in recent years and has consequently attracted a lot of attention in the research community making hate speech detection a very active research field, e.g. [1]. In particular the growing impact of social media on the way people share and access information has demonstrated the need to tackle the problem systematically as issues such as cyber-bullying and other hurtful and anti-social behaviours [2, 1, 3] have become a growing cancer that needs to be tackled broadly across many different platforms and applications. We should note that the task of removing hate speech is not as simple as it seems as there is a fine balance between filtering hate speech and the possible restriction to the freedom of speech if a perfectly reasonable opinion is incorrectly flagged as hate speech and subsequently removed [4].

The motivation of this task [5] is to move from a purely reactive to a more pro-active approach that does not simply identify messages as hate speech but instead identifies social media users as hate speech spreaders thereby allowing the problem to be addressed more effectively (e.g. by suggesting to the owner of the social media platform to ban such users).

Transformer-based methods have been demonstrated to be highly effective for a wide range of NLP tasks, e.g. [6]. This is the reason we adopt state-of-the-art pre-trained transformer-based

---


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

✉ kaodamie77@gmail.com (K. O. Akomeah); udo.kruschwitz@ur.de (U. Kruschwitz); bernd.ludwig@ur.de (B. Ludwig)

ORCID 0000-0002-5503-0341 (U. Kruschwitz)



© 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)

deep neural text embeddings for tackling the PAN sub-task on profiling hate speech spreaders on Twitter. In this report, we will provide an overview of the steps taken and the models used in our experiment. We will start by briefly describing the dataset, looking at the pre-processing steps and models used before we report our results obtained for the two submissions as compared to the baselines [7] submitted by the organizers.

## 2. PAN Task 3: Profiling Hate Speech Spreaders on Twitter

The third task [5] of the PAN Challenge at CLEF 2021 [8] involves the profiling of hate speech spreaders on Twitter towards, for instance, immigrants and women using sampled data from the individual user's timeline.

The training dataset consists of 40,000 tweets constituting a set of 200 tweets sampled per each of 200 anonymized users in XML format for two languages, English and Spanish. The test set contains tweets from 100 anonymized users per language. The tasks were treated separately for each language and therefore two different models were used for both English and Spanish.

An small snapshot of the raw training English data is reproduced in Figure 1.

Note that the 200 tweets of hate speech spreaders may not all contain hate speech. The aim of the experiment is to discover if frequent hate spreaders can be identified based on their timeline history.

The systems are ranked using the average of *Accuracy* achieved on the English and Spanish test sets. Submission and evaluation of this year's tasks were done on TIRA [9] or sent to the organizers through mail. All codes used in this experiment can be accessed via GitHub.<sup>1</sup>

### 2.1. Preparing the Data using Contextual Embeddings

In recent years, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) have emerged as the dominant paradigm in a broad range of NLP applications ranging from translation to classification, e.g. [6, 10]. Part of the success story is the fact that the expensive task of pre-training is only done once and this pre-trained model can be subsequently fine-tuned to each NLP task at hand with just one additional output layer to create state-of-the-art models. In effect there has been a large number of different BERT-based models that have emerged from this, e.g. [11, 6, 12, 13, 14]. Specifically, we turn the textual representation of the input documents (tweets) into contextual embeddings as follows. The input data (in XML format) has to be turned initially into tensors for input in the Keras Model by first extracting all text for each user using an XML parser and pandas in Python. The dataframes indexed with user-ids are then formatted into tensors ready to be used as input for the transformer model. The training dataset was split for train-test reasons; 160 for training, 32 for validation, 8 for testing and shuffled for every other time it was run.

### 2.2. Approach for the English Run

The experiment for the English task involved the binary classification test set of 100 users similarly parsed in XML format. The training set is composed of 200 users with 200 tweets each.

---

<sup>1</sup><https://github.com/kaodamie>

```

<author lang="en" class="1">
  <documents>
    <document><![CDATA["Hey Jamal (snickering uncontrollable) You want some (PFFF)
LEMONADE!" What an IDIOT! #URL#]></document>
    <document><![CDATA[RT #USER#: Cotton coming out with a banger #URL#]></document>
    <document><![CDATA[This is meant to be sarcasm but it's a good point considering how
underwhelming the pandemic has been #URL#]></document>
    <document><![CDATA[Nick really just compared homosexuality to people shooting themselves in
the head 🤔🤔]]></document>
    <document><![CDATA[PROTECT AMERICA FIRST! LET'S GO!!!!!! #URL#]></document>
    <document><![CDATA[All these fears about safety but a total refusal to address the problem is
NONWHITE IMMIGRANTS #URL#]></document>
    <document><![CDATA[People will notice. It just takes time.]]></document>
    <document><![CDATA[#USER# I thought people were doing it to prevent getting reported, not for
fun]]></document>

```

**Figure 1:** Small sample of the raw XML of the English Training Data

The model architecture used was a dense artificial neural network with a single output with sigmoid activations. With the success of BERT-based models in NLP, we employed ALBERT, a BERT-based model trained on a large corpus of the English language with reduced parameters without a significant effect on the performance benchmark [11]. Reusing the model only required a fine-tuning of its parameters on the dataset which requires that output from ALBERT is learned as well in the network. The network had a dense layer receiving BERT encoder outputs with a dropout of 0.1 with sigmoid activation on a single output layer. The network run for 10 epochs with 5 steps per epoch and a batch size of 32. The loss function used was binary cross-entropy with an adaptive moment estimation (Adam) optimizer and a learning rate of 1e-6. The metric used for training was binary accuracy in line with accuracy as the specified metric for evaluation of the challenge.

A checkpoint was implemented for the neural network. The epochs had an average runtime of about 250 seconds and therefore a larger number of epochs would be costly. The checkpoint was to monitor the minimum binary validation loss with a patience of 3 epochs. The validation loss was chosen other than training loss to check overfitting of the training dataset.

**Table 1**  
Accuracy Results for English

English(Albert)	Sample Size	Accuracy
Training	160	0.59
Validation	32	0.65
Test	8	0.86
Evaluation Test	100	0.53

### 2.3. Approach for the Spanish Run

A similar model to the English run was used for the Spanish run in that we used a transformer-based model containing a text input layer, preprocessing layer, an encoding layer, and a single densely-connected layer as output also with a dropout of 0.1. The binary cross-entropy loss function, as well as Adam optimizer were also used for the Spanish run. The preprocessing layer was a multilingual universal sentence encoder preprocessor [10]. This preprocessor is a companion to the BERT models for preprocessing plain text inputs into the input format expected by BERT. The model uses a vocabulary for multilingual models extracted from Wikipedia, CommonCrawl, and translation pairs from the Web. It has no trainable parameters and can be used in an input pipeline outside the training loop [10, 15].

The encoder layer used was the language-agnostic BERT sentence embedding model (LaBSE) [16]. LaBSE supports about 109 languages including Spanish. The language-agnostic BERT sentence embedding encodes sentences into high-dimensional vectors. The model is trained and optimized to produce similar representations solely for bilingual sentence pairs that are translations of each other. Because of its usefulness in sentence translations in a larger multilingual corpus, text classification, semantic similarity, clustering and other natural language tasks [16, 15] we applied it for this classification task.

The encoder was not fine-tuned because of the large memory requirement. Running on Google Colabs required a RAM of about 12 gigabyte and even for better performance and speed, a GPU and a RAM greater 32 gigabyte is recommended. The model was trained in 10 epochs with callbacks on the binary validation loss.

## 3. Results Obtained

During training the fine-tuned ALBERT model used for the English task peaked at a best binary validation accuracy of 0.65 as illustrated in Table 1. A possible explanation for this rather low figure is that the length of data used for training was length of 200 and longer sentences for each user. Besides each user having 200 tweets, those users profiled as hate speech spreaders may still be quite similar to non-hate spreaders as not all tweets in the history of hate spreaders may contain hate. Therefore training a model to perform classification on such a dataset can be quite a challenge.

However, with the Spanish model which applied a BERT-based language-agnostic sentence encoder that was not fine-tuned performance peaked at a binary validation accuracy of 0.71 (see Table 2) but (unlike the English system) performed better on the test set. The model attained

**Table 2**

Accuracy results for Spanish

Spanish(LaBSE)	Sample Size	Accuracy
Training	160	0.56
Validation	32	0.71
Test	8	0.75
Evaluation Test	100	0.77

**Table 3**

Baselines comparison of Accuracy results for Spanish test set

Model	Accuracy
Word nGram+SVM	0.83
LSDE	0.82
USE+LSTM	0.79
LaBSE(ours)	0.77
MBERT LSTM	0.76
XLMR-LSTM	0.73
TFIDF-LSTM	0.51

an accuracy of 0.53 for English and 0.77 for Spanish after evaluation on the test set for the challenge.

It is not easy to put these numbers in context – other than observing that the performance of the English run was surprisingly low. The language-agnostic BERT-based sentence encoder on the other hand performed better as our Spanish run outperformed three of the baselines [7] submitted (see Table 3). Understanding why different approaches perform better or worse on a particular dataset is not easy anyway, in particular when it comes to the explainability and interpretability of neural network-based approaches, e.g. [17][18] as performance can be affected by many parameters including a particular sample, learning rate, initialized weights among others used in training. What we do however see is a huge performance variation across training, validation and test data as well as across different submissions for this task. We attribute this in part to the small sample making it difficult to draw generalizable conclusions from a single run. We conclude that the approaches need to be tested on a wide range of additional collections to gain a better understanding of strengths and weaknesses as well as variation of results and robustness more generally, something that fits well with the idea of moving away from aiming to train systems that do amazingly well on specific collections but tend to fall over when applying them elsewhere, e.g. [19].

## 4. Conclusions

The use of transformer-based models for NLP tasks has pushed the state of the art in many applications. In this experiment, two different (very simple) BERT-based models were applied in an attempt to classify hate speech spreaders through training on the bulkiness of their timelines.

We extracted contextual embeddings by using pretrained transformer-based models and then run through a single layered output for classification. What we found in our experiments was that the power of transformer-based approaches varied substantially across runs, and some traditional baselines did in fact perform surprisingly well (at a much lower cost overall). We do however not see this as a reflection of the weakness of more sophisticated methodologies but more of an issues arising from the datasets that are used for training and testing. The aim should be to explore a wide range of datasets to find out which of the methods is most robust, something particularly important when thinking about hate speech. A particularly promising approach, which has been shown to work well in many NLP tasks including hate speech detection [4], is to use ensemble classifiers which can tap into the different strengths of individual classifiers, be it transformer-based or traditional ideas.

## Acknowledgements

This work was supported by the project *COURAGE: A Social Media Companion Safeguarding and Educating Students* funded by the Volkswagen Foundation, grant number 95564.

## References

- [1] M.-A. Rizoïu, T. Wang, G. Ferraro, H. Suominen, Transfer learning for hate speech detection in social media, arXiv preprint arXiv:1906.03829 (2019).
- [2] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, 2012, pp. 19–26.
- [3] D. Ognibene, D. Taïbi, U. Kruschwitz, R. S. Wilkens, D. Hernandez-Leo, E. Theophilou, L. Scifo, R. A. Lobo, F. Lomonaco, S. Eimler, H. U. Hoppe, N. Malzahn, Challenging social media threats using collective well-being aware recommendation algorithms and an educational virtual companion, arXiv preprint arXiv:2102.04211 (2021).
- [4] 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.
- [5] 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.
- [6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.
- [7] F. Rangel, M. Franco-Salvador, P. Rosso, A Low Dimensionality Representation for Language Variety Identification, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 156–169.
- [8] 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. Wol-ska, , 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.
- [9] 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.
- [10] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).
- [11] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942 (2019).
- [12] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
- [13] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).
- [14] 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).
- [15] Z. Yang, Y. Yang, D. Cer, J. Law, E. Darve, Universal sentence representation learning with conditional masked language model, arXiv preprint arXiv:2012.14388 (2020).
- [16] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).
- [17] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semanova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, arXiv preprint arXiv:2103.11251 (2021).
- [18] Y. Zhang, P. Tiño, A. Leonardis, K. Tang, A survey on neural network interpretability, arXiv preprint arXiv:2012.14261 (2020).
- [19] S. R. Bowman, G. E. Dahl, What will it take to fix benchmarking in natural language understanding?, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 4843–4855. URL: <https://www.aclweb.org/anthology/2021.naacl-main.385/>.