On the Performance of Different Text Classification Strategies on Conspiracy Classification in Social Media

Manfred Moosleitner
Universität Innsbruck, Austria
manfred.moosleitner@uibk.ac.at

Benjamin Murauer
Universität Innsbruck, Austria
b.murauer@posteo.de

ABSTRACT

This paper summarizes the contribution of our team UIBK-DBIS-FAKENEWS to the shared task “FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task” as part of MediaEval 2021, the goal of which is to classify tweets based on their textual content. The task features the three sub-tasks (i) Text-Based Misinformation Detection, (ii) Text-Based Conspiracy Theories Recognition, and (iii) Text-Based Combined Misinformation and Conspiracies Detection. We achieved our best results for all three sub-tasks using the pre-trained language model *BERT Base*[1], with extremely randomized trees and support vector machines as runner ups. We further show that syntactic features using dependency grammar are ineffective, resulting in prediction scores close to a random baseline.

1 INTRODUCTION

This task consists of the three sub-tasks, in which properties of short social media posts must be predicted. The sub-tasks’ goals are similar but distinctively different: In sub-task 1 (Text-Based Misinformation Detection), each document belongs to one of three classes ("Promotes/Supports Conspiracy", "Discusses Conspiracy", "Non-Conspiracy"), making it a multi-class classification problem. In sub-task 2 (Text-Based Conspiracy Theories Recognition), a list of conspiracies is provided. For each conspiracy, the goal is to predict whether that conspiracy is mentioned in a document, whereas more than one conspiracy can be mentioned in one document. This makes it a multi-label classification problem. Finally, in sub-task 3 (Text-Based Combined Misinformation and Conspiracies Detection), the above sub-tasks are combined and for each evaluation document, the model must predict for each of the provided conspiracies the way that conspiracy is mentioned according to sub-task 1. The development data provided for this task consists of about 1,500 tweets collected by Schroeder et al. [5], and a detailed description of the individual sub-tasks can be found in the task overview paper [4]. The code of our solution is available online[1].

For each task, each team is allowed to submit five runs with the following restrictions: For run 1, only features extracted from the provided texts without additional external data or pre-trained models were allowed (concretely, this also disallows using any BERT-related model). For run 2, the usage of pre-trained models was additionally allowed, and for runs 3 and 4, also the use of external data for training any model was permitted.

Table 1: Hyper-Parameters Used in Grid Search

Parameter	Tested Values
n-gram size	1, 2, ..., 10
n-grams max. features	unlimited, 1000
lowercase text	true, false
DT-gram word repr.	universal POS tag, English POS tag
BERT model	RoBERTa, DistilBERT, BERT base
num. trees	100, 250, 500, 750, 1000, 2000, ..., 5000

In our approach to this task, we perform a large-scale grid search experiment testing a variety of different feature extraction methods and classification models and hyper-parameter configurations thereof. We show that the pre-trained language model BERT outperforms the other presented approaches for all the sub-tasks and that the syntax-based features are not able to detect the conspiracy-related classes in the documents.

2 TEXT FEATURES

We include multiple types of features in our grid experiments. As a widely used and simple to calculate baseline, we include word and character *n*-grams. We thereby test different configurations of the extraction, including different sizes of *n* and pre-processing the texts to lowercase. The full list of parameters is shown in Table 1. We normalize the frequency of the resulting *n*-grams by using tf-idf.

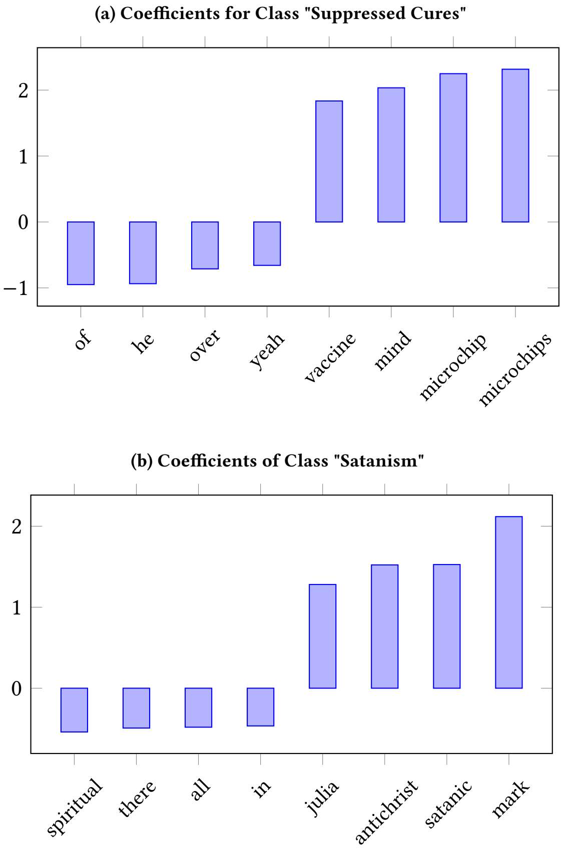
As the second type of text features, we calculate Dependency-Tree-grams (DT-grams) [3] to determine if texts within one class or label have similar grammatical structures. DT-grams are substructures of the dependency graph of sentences, and can be interpreted as a way to enhance *n*-grams of part-of-speech tags by redefining which tokens are “close” to one another and therefore form an *n*-gram. Thereby, each word is represented either by its English-specific, part-of-speech tag, or a combination thereof. This choice is a hyper-parameter and is tuned by the grid-search (cf. Table 1). Like their character- and word-based counterparts, we calculate the frequency of the resulting *n*-grams as feature and use tf-idf normalization. We include this feature to check whether some of the conspiracy classes exhibit stylistic markers that are typical for that category.

Lastly, we use the sequence of tokens in the unmodified text as input for fine-tuning different pre-trained language models to directly perform classification on the evaluation sets. Thereby, no splitting of the training documents was required as the documents are short enough to be included in any of the language models.

[1] https://git.uibk.ac.at/c7031305/mediaeval2021-fakenews

MediaEval’21, December 6-8 2021, Online

Figure 1: Top 4 Positive and Negative Coefficients of the Classes "Suppressed Cures" and "Satanism"



(a) Coefficients for Class "Suppressed Cures"

(b) Coefficients of Class "Satanism"

3 CLASSIFICATION MODELS

We employ three different machine learning models with the tf-idf normalized frequencies of character-, word- and DT-gram-based features: support vector machines (SVM), multinomial naive bayes (MNB), and extremely randomized trees (ET). The hyper-parameters for these models are also listed in Table 1.

We use three BERT-like models: BERT-base [1], RoBERTa [2], and DistilBERT [6]. Thereby, we use a maximum sequence length of 256, three epochs of fine-tuning, and a batch size of 8. All other parameters were left untouched from their default implementation provided by the *huggingface*[2] library.

4 RESULTS AND DISCUSSION

Generally, our results show a strong connection between certain keywords and their corresponding conspiracy theories. Figure 1 shows the top four positive and negative coefficients for two of the classes ("Satanism" and "Mind Control") from sub-task 2, which are an intuitive way to show the relationship of the words with the corresponding classes.

Table 2: Official evaluation results measured with Matthew's correlation coefficient.

	Run 1	Run 2	Run 3	Run 4
Features	char 6-grams tf-idf	plain text sequence	word 1-grams tf-idf	dt-gram tf-idf
Model	ET	BERT	SVM	MNB
Task 1	0.2852	**0.3184**	0.2228	0.1201
Task 2	0.2086	**0.3624**	0.2879	0.0000
Task 3	0.1993	**0.3347**	0.2316	-0.0028

From the grid search experiments, we select the best-performing configurations for each of the allowed runs, which are shown in Table 2. Since for the first run we were not allowed to submit the BERT-based method, we submitted the ET-based solution, which was second in line.

The task organizers released the development dataset in two stages: the first part consisted of 500 documents, and the second part included an additional 1,000 documents. With only the first part available, the ET model outperformed all others and was only outperformed by BERT when the full dataset was available. This indicates that BERT-like models require a minimum amount of text data for pre-training to perform efficiently. Concretely, the addition of the second part of the development dataset increased the number of documents for the smallest class "Discusses Conspiracy" from 76 to 262, giving a rough impression of how many samples are required for BERT to perform better than the ET model.

When comparing the Results of extremely randomized Trees and SVMs, we can see that the performance of ET was better than the performance of SVM for sub-task 1, and vice versa for sub-task 2, where SMV performed better than ET. We think one of the reasons for this is because the use of word uni-grams with SVM, compared to character 6-grams with ET, indicating that whole words better reflect the connection between keywords and labels than only partial words. Also indicating a connection between certain keywords and their labels is that the performance of BERT in sub-task 2 (multi-label) and sub-task 3 (multi-class, multi-label) is a little better than for sub-task 1 (multi-class).

On the other hand, the results produced by our grammar-based approach show a poor performance in all sub-tasks, indicating that the grammatical structure of the texts as feature is not suited to differentiate between the given classes and labels. We attribute this behavior to the short texts, which are not likely to incorporate complex grammatical structures, as well as difficulties in parsing due to the unstructured nature of the text.

ACKNOWLEDGMENTS

We would like to thank our research group for the feedback and Prof. Günther Specht, head of our research group, for providing the necessary infrastructure to perform the research reported in this paper.

[2] https://huggingface.co/

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[3] Benjamin Murauer and Günther Specht. 2021. DT-grams: Structured Dependency Grammar Stylometry for Cross-Language Authorship Attribution. *arXiv preprint arXiv:2106.05677* (2021).

[4] Konstantin Pogorelov, Daniel Thilo Schroeder, Stefan Brenner, and Johannes Langguth. 2021. FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021. In *Proc. of the MediaEval 2021 Workshop, Online, 13-15 December 2021*.

[5] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*. 21–25.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).