Don't Just Drop Them: Function Words as Features in COVID-19 Related Fake News Classification on Twitter

Pascal Schröder
Radboud University, Netherlands
pascal.schroeder@ru.nl

ABSTRACT

This research shows that function words can be useful as features for machine learning models tasked with detecting conspiratorial content in COVID-19 related Twitter posts. A significance test exposes that the distribution of function words between fake and legitimate content varies greatly. Further, a support vector machine classifier is demonstrated to perform above chance when using function word-only features, achieving a Matthews correlation coefficient of 0.139 on unseen test data.

1 INTRODUCTION

Previous research into detecting conspiratorial online content indicates that it can be distinguished by the author's writing style [2, 10, 11]. Simultaneously, function words provide a meaningful proxy to an author's writing style in authorship attribution [1, 12]. Therefore, function words could be valuable in aiding machine learning models tasked with detecting conspiratorial content. However, many approaches in fake news classification still rely on a purely content-based approach, in which function words are excluded as part of the preprocessing [2]. While these approaches oftentimes offer impressive performances, it is paramount that potentially relevant features are not excluded in the process.

Fortunately, recent years have seen a growing body of research on the importance of stylistic features in misinforming and conspiratorial content. For instance, Posadas-Durán et al. [9] showed that better performance levels can be reached for classifiers when function words are incorporated in the training data, versus when they are not. However, like many approaches, they have used online news articles as their data [2]. Arguably, the style entertained by authors of social media posts will be different, and it is to be expected that the results do not generalise.

For social media and especially Twitter data, the literature is rather sparse. Del Tredici and Fernández [13] have classified articles shared on Twitter as fake or real, and enhanced their data with the user's post history and profile description, and found more function words in the latter. However, they have not classified posts directly, but rather articles linked in posts. Niven et al. [6] have used function words as a proxy for 'thoughtfulness', arguing that the latter correlates with the fakeness of a post's content, but have not found a significant difference in distribution between fake and legit content. But since they only had available posts from 300 different users, individual authors could have possibly skewed the distribution. Im et al. [4] have found an above chance performance for function word-only features when predicting whether a post stems from a Russian troll account, which while related, is still not focused on fake news detection specifically.

This research thus aims to expand on the available literature by analysing how function word usage is distributed between authors of conspiratorial versus legitimate content, and quantifying whether function words alone can act as sufficient features in the classification of fake news content.

Table 1: Top 10 function words sorted by p-values of the χ^2 test, with absolute frequencies per class. Italic words are pronouns. Bold denotes the class with most occurrences.

Word	P-Value	# No Consp.	# Consp.
my	1×10^{-8}	**77**	18
is	8×10^{-8}	**415**	340
the	1×10^{-7}	**596**	443
they	6×10^{-6}	**166**	140
used	0.00001	8	**27**
am	0.00003	**53**	11
during	0.00007	**33**	4
their	0.00052	57	**62**
he	0.00071	**89**	40
she	0.00074	**16**	1

2 APPROACH

The data were provided through the MediaEval 2021 Conference, for the task 'FakeNews: Corona Virus and Conspiracies Multimedia Analysis' [7, 8]. It consists of 1554 Twitter posts related to COVID-19 and different conspiracy theories.[1] Three different class labels are provided: tweets that *do not mention conspiracies* (1), tweets that *discuss conspiracies* without actively supporting them (2), and tweets that *promote or support conspiracies* (3). Note that the classes are imbalanced, with 767 examples for class 1, 271 examples for class 2, and 516 examples for class 3.

To investigate a possible difference in distribution between conspiratorial and non-conspiratorial content, most of this research thus focused on classes 1 and 3. First, function words were extracted from all data using spaCy.[2] To test for significance, a χ^2 test was performed on the distribution of function words between the classes 1 and 3.

Next, the usefulness of the extracted function words was tested. As a representative model, a support vector machine (SVM) with non-linear kernels was chosen, and evaluated using classification

MediaEval'21, December 13-15 2021, Online

[1] All posts are written in English and were collected between January 17, 2020 and June 30, 2021.
[2] For the full list of function words, see https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py

accuracy and Matthews correlation coefficient (MCC). Two scenarios were investigated, one with data containing all three classes, and one for data containing the classes 1 and 3 only. To account for random effects, 100 runs were computed for both scenarios, each with a random 10% validation split. As a final evaluation, a single SVM model was fitted on all available data, and evaluated on an unseen test set. Due to organisational means, this evaluation is only available for the 3-class case using MCC. Preprocessing of function words was done using the `TfidfVectorizer` of SciKit Learn.[3]

3 RESULTS AND ANALYSIS

The results of the χ^2 test can be seen in Table 1, showing the 10 function words with lowest p-value, all of which are below 1%. 5 are pronouns (50%), which is higher than the overall frequency of pronouns in the function words (≈11.48%). Only 2 out of the 10 words occur more often in class 3 (conspiracy), while the remaining 8 occur more often in class 1 (no conspiracy).

The following are all sub 5% function words, 40 in total, sorted from lowest to highest p-value:

> *my*, is, the, *they*, **used**, am, during, ***their***, *he*, *she*, by, *this*, *him*, **serious**, doing, might, *his*, if, *us*, but, be, ***these***, all, seem, about, **part**, *her*, **along**, could, *your*, due, have, are, here, **using**, at, per, when, would, now

Of these words, 12 are pronouns (30%), marked in italic. 7 belong to class 3, marked in bold, while 33 belong to class 1.

Table 2 shows the results of the SVM classifiers.

Table 2: Average performance of SVM classifiers trained on function word features for the binary case between classes 1 and 3, as well as for all three classes, with random 10% validation splits. Test result is for a single SVM trained on all available data.

	Train		Val		Test
	Mean	Std.	Mean	Std.	
Accuracy (binary)	0.911	0.005	0.637	0.039	-
MCC (binary)	0.818	0.009	0.203	0.088	-
Accuracy (all)	0.819	0.006	0.533	0.034	-
MCC (all)	0.710	0.010	0.165	0.056	0.139

4 DISCUSSION AND OUTLOOK

The results of the χ^2 test show that a significant difference in the distribution of function words can be observed between conspiratorial and non-conspiratorial content, confirming the idea that such words are indeed important for distinguishing fake from legitimate content. This is further supported by the classification results, where an above chance performance on unseen test data was achieved. Unsurprisingly, the classification performance in the binary case was higher than in the full case, since an overlap in style between non-conspiratorial content as well as content which does not actively support conspiracies, but merely discusses them, is to be expected.

Interestingly, the category of function words most common in the list of sub 5% p-value words were pronouns, for which the relative frequency was greatly increased compared to their frequency in all function words. Further, all of these pronouns except *their* and *these*, occurred more often in the non-conspiracy category. Most prominent are third person singular pronouns (*he*, *she*, *him*, *his*), all featured more in class 1, which indicates that conspiracy authors are less likely to talk about a person at length. This could be because giving extensive detail (e.g. 'She said *X*') rather than implying makes their claims falsifiable, which they might be interested to avoid [3]. Interestingly, this stands in contrast to Rashkin et al. [10], who found a higher frequency of pronouns in conspiratorial content. Of further interest is the fact that the pronoun *my* displays the most significance overall. This could again be because conspiratorial authors want to steer the argument away from their own opinion to a more general claim, thereby avoiding responsibility. This finding is somewhat supported by Newman et al. [5], who found higher usage rates of the pronoun *I* in people who are lying.

Apart from pronouns, the overall majority of function words with p-values below 5% belong to the non-conspiratorial class. This indicates that authors of fake content use a more simplistic style, as the complexity of a text correlates with the number of different function words used.

An earlier analysis on a smaller subset of the data showed different patterns in the function word distributions, most notably the presence of 'hedging' words like *quite*, *rather* and *somehow*. However, these patterns disappeared when the larger data set was released. Therefore, it is important to note that the data set at hand, with only 1554 total posts, is a very limited subset of all COVID-19 related data found on Twitter. Thus, it cannot be ruled out that the patterns found in this research, although powerful in predicting on the chosen dataset, may not generalise.

This limitation is extended by the fact that the data analysed in this report does not contain author information. As stylistic information correlates very strongly with the author of a text, the patterns found could, in theory, be caused by a few authors having a disproportionately high representation in the data. This effect unfortunately could not be accounted for due to the missing authorship information.

In conclusion, this research has shown that function words are a strong proxy for detecting conspiratorial content in the context of COVID-19 related fake news on Twitter. To address the limitations of this research, future work should explore in how far these results generalise to larger corpora of Twitter data, different domains of conspiracy, as well as other social media platforms.

ACKNOWLEDGMENTS

I thank Martha Larson for her critical input during the development of the research question as well as the analysis process, and Lynn de Rijk for her input on function word usage and interesting lexical patterns in fake news content.

REFERENCES

[1] Shlomo Argamon and Shlomo Levitan. 2005. Measuring the Usefulness of Function Words for Authorship Attribution. *Proceeding of the*

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Joint Conference on Association for Literary and Linguistic Computing/Association Computer Humanities.

[2] Nicollas R. de Oliveira, Pedro S. Pisa, Martin Andreoni Lopez, Dianne Scherly V. de Medeiros, and Diogo M.F. Mattos. 2021. Identifying fake news on social networks based on natural language processing: Trends and challenges. *Information (Switzerland)* 12 (Jan. 2021), 1–32. Issue 1. https://doi.org/10.3390/info12010038

[3] Lynn de Rijk. 2020. You Said it? How Mis- and Disinformation Tweets Surrounding the Corona-5G-conspiracy Communicate Through Implying. *Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020* (2020).

[4] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2020. Still out There: Modeling and Identifying Russian Troll Accounts on Twitter. In *12th ACM Conference on Web Science (WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3394231.3397889

[5] Matthew Newman, James Pennebaker, Diane Berry, and Jane Richards. 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality & Social Psychology Bulletin* 29 (June 2003), 665–75. https://doi.org/10.1177/0146167203029005010

[6] Timothy Niven, Hung-Yu Kao, and Hsin-Yang Wang. 2020. Profiling Spreaders of Disinformation on Twitter: IKMLab and Softbank Submission. In *CLEF 2020*.

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

[8] 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. *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks, pp. 21-25* (2021).

[9] Juan Pablo Posadas-Durán, Helena Gomez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the Spanish language. *Journal of Intelligent and Fuzzy Systems* 36 (May 2019), 4868–4876. Issue 5. https://doi.org/10.3233/JIFS-179034

[10] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. https://doi.org/10.18653/v1/D17-1317

[11] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *Special Interest Group on Knowledge Discovery in Data: Explorations Newsletter* 19 (Aug. 2017). https://doi.org/10.1145/3137597.3137600

[12] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the Association for Information Science and Technology* 60 (March 2009), 538–556. https://doi.org/10.1002/asi.21001

[13] Marco Del Tredici and Raquel Fernández. 2020. Words are the Window to the Soul: Language-based User Representations for Fake News Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5467–5479. https://doi.org/10.18653/v1/2020.coling-main.477