Bigfoot in Big Tech: Detecting Out of Domain Conspiracy Theories

Matthew Fort, Zuoyu Tian, Elizabeth Gabel, Nina Georgiades, Noah Sauer,
Daniel Dakota, Sandra Kübler
Indiana University
{mattfort,zuoytian,eligabel,ngeorgia,sauerno,ddakota,skuebler}
@iu.edu

Abstract

We investigate approaches to classifying texts into either conspiracy theory or mainstream using the Language Of Conspiracy (LOCO) corpus. Since conspiracy theories are not monolithic constructs, we need to identify approaches that robustly work in an out-of-domain setting (i.e., across conspiracy topics). We investigate whether optimal in-domain settings can be transferred to out-of-domain settings, and we investigate different methods for bleaching to steer classifiers away from words typical for an individual conspiracy theory. We find that BART works better than an SVM, that we can successfully classify out-of-domain, but there are no clear trends in how to choose the best source training domains. Additionally, bleaching only topic words works better than bleaching all content words or completely delexicalizing texts.

1 Introduction

With the rise of social media over the last 10 years, there has also been a rise in the uses of the internet to spread different types of information, some of it of a more questionable nature. We are interested in the spread of conspiracy theories, which have morphed from a fringe phenomenon to a more widely visible, mainstream phenomenon. Along with the increasing spread of misinformation, conspiracy theories have been shown to polarize opinions to extremes and to incite violence (Douglas and Sutton, 2018; Enders et al., 2022).

While conspiracy theories are often seen as monolithic belief systems, the truth is more complex: People who admit to believing a specific conspiracy theory tend to also believe in other conspiracy theories, but they may only believe different subsets of factoids associated with a specific conspiracy theory (Enders et al., 2021). For any computational approach to detecting conspiracy theories, this means that we cannot expect to have access to accurate training data. Instead, we will face novel mixes of factoids and conspiracy theories, which deviate from existing training data. For this reason, we investigate here whether it is possible to find out-of-domain conspiratorial texts. We use the Language Of Conspiracy (LOCO) corpus (Miani et al., 2021) to develop classifiers that label a text as either conspiratorial or mainstream, and we investigate under which conditions such classifiers work robustly out-of-domain. More specifically, we investigate bleaching methods to steer the classifiers away from words that are typical for a single conspiracy theory (e.g., 'global warming' for conspiracy theories revolving around climate change).

The remainder of this paper is structured as follows: Section 2 explains our research questions, section 3 describes related work, and section 4 describes our data and methodology. Section 5 describes our results for the in-domain setting (section 5.1), for the out-of-domain setting (section 5.2), and for the bleaching experiments (section 5.3). We conclude in section 6.

2 Research Questions

In this paper, we investigate the following research questions:

1. Which machine learning architectures are well suited for classifying texts into conspiracy theory and mainstream? Which feature types do we need? Does feature selection improve results for SVMs?

2. Can we classify out-of-domain texts? In other words, do we need training data from a specific conspiracy theory, or is it possible to reuse existing training data to detect novel conspiracy theories?

Proceedings of Recent Advances in Natural Language Processing, pages 353–363
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_040

3. Does bleaching specific words improve out-of-domain results? I.e., can we identify sets of words that are too specific for a single conspiracy theory but do not work well for classifying texts from another conspiracy theory?

3 Related Work

We restrict our review to work on conspiracy theories and their detection. We acknowledge work on propaganda detection and persuasive technology detection (e.g., Barrón-Cedeño et al., 2019; Da San Martino et al., 2019; Martino et al., 2019). There is overlap between these areas of research and the detection of conspiracy theories, given that both approaches work on the document level and examine how information is manipulated. However, propaganda detection primarily focuses on politically related events, whereas conspiracy beliefs tend to span a wide array of topics.

Although exact markers have proven difficult to identify for conspiracy theories, Wood et al. (2012) showed that conspiracy theory proponents often subscribe to multiple conspiracies, some contradictory, which led them to conclude that conspiracy theories are not stand-alone phenomena from individuals. Instead, conspiracies might come in clusters caused by general conspiratorial thinking.

Work by Klein and Hendler (2022) found that certain lexical items can be used to differentiate between some conspiratorial and non-conspiratorial texts in Reddit posts and a forum popular among anti-vaccine proponents. Examples of conspiracy-indicative lexical items include so-called thought-terminating cliches, such as 'agree to disagree', 'do [your/your own/the] research', and dysphemisms such as 'fraudulent', 'deceptive', and 'deceive' rather than 'lie.'

Attempts to identify linguistic characteristics used in conspiracy theories were explored by Klein et al. (2019). They used the Linguistic Inquiry and Word Count (LIWC) to analyze the conspiracy subreddit, in order to identify lexical categories based on a semantic knowledge base. In a majority of instances, conspiracy users exhibited a statistically relevant usage of words used to induce 'negative emotion' and 'anger' among others, making conspiracy texts more distinguishable.

Similar findings were noted within the Language Of Conspiracy (LOCO) corpus (Miani et al., 2021). The corpus was seeded using phrases related to conspiracy theories to collect close to 100 000 text documents taken from 150 websites, dividing texts into those containing conspiratorial content and mainstream documents. A lexical analysis of conspiracy based on LIWC categories and using Empath, a tool that generates new lexical associations in texts, showed that conspiracy theories contain more emotionally charged language, particularly language indicating negative emotions such as anger.

Mompelat et al. (2022) analyzed two conspiracy theories, Sandy Hook and Coronavirus, in the LOCO corpus, to establish a set of unique features (e.g., linguistic) by which mainstream and conspiracy documents could be differentiated. They noted that a significant portion of conspiracy documents did not contain unique identifiable features, suggesting automatic classification would be difficult. They also found that mainstream documents were frequently irrelevant regarding the topic of the conspiracy theory for which they were retrieved.

As new conspiracy theory corpora have been assembled, the capabilities of models to detect novel conspiracy theories have been explored. Phillips et al. (2022) created a Twitter data set covering four conspiracy topics: climate change, COVID-19 origin, COVID-19 vaccine, and the Epstein-Maxwell trial. They used several BERT variants to classify tweets as conspiracy theory vs. non-CT, to identify the tweets' stance towards a conspiracy theory, and to detect the topic of the conspiracy theory. While they suggest that successful models can be built with relatively small data sets, they also note that annotator disagreement and class imbalance can contribute to difficulties in reliable classification.

4 Methodology

4.1 Data Set

We use the Language Of Conspiracy (LOCO) corpus (Miani et al., 2021) and select five conspiracies that fall across a spectrum of political and social associations: vaccines, climate change, pizzagate, flat earth, and bigfoot. Given the uneven distribution of these conspiracies in the LOCO corpus, ranging from approx. 1 300 to 7 000, we randomly select a subsample of 1 330 texts from each conspiracy, while maintaining a relative balance between the mainstream and conspiracy labels across the conspiracy theories. We then randomize the data and create an 80/10/10 split of training/development/test data. The final numbers of documents per set are shown in Table 1.

Topic	Train		Develpment		Test	
	Mainstream	Conspiracy	Mainstream	Conspiracy	Mainstream	Conspiracy
vaccine	796	268	104	29	100	33
climate change	799	265	99	34	102	31
pizza gate	808	256	95	38	97	36
flat earth	802	262	100	33	98	35
bigfoot	816	248	93	40	91	42

Table 1: Data split per conspiracy theory.

4.2 Classifiers

SVM We train a model using an SVM (Cortes and Vapnik, 1995) with a linear kernel using different feature sets including word n-grams, character n-grams, and POS tags. We set the minimum frequency to 1; word n-grams include unigrams, bigrams, and trigrams while character n-grams are between 3-7 in length. All experiments are performed using scikit-learn (Pedregosa et al., 2011). We perform a grid search to find the best parameters of our SVM models by evaluating on the development set on in-domain experiments and then use these parameters for all other experiments.

Feature selection For the feature selection experiments, we use the built-in χ^2 metric in scikit-learn.

Transformer We use BART (Lewis et al., 2020), a pre-trained transformer-based seq2seq model with a bidirectional encoder, but a left-to-right autoregressive decoder. Rather than optimizing on next sentence prediction, the model is trained by restoring corrupted documents to their original form. One advantage of this is that the model is thus learning larger structures and context within a document rather than a more localized neighboring sentence. We view this as preferential given the longer length of documents and irregular information ordering. Additionally, the maximum tokenized input is 1024, which is double the maximum input to standard BERT models (Devlin et al., 2019). Both aspects should benefit our use-case given the relatively long length of individual documents within the corpus (see section 5.4). Despite this, most documents are still too long to be embedded. We choose to embed the first and last 512 subtokens in order to attempt to capture more information on a document level[1]. We experiment with one, three, and five epochs on the dev set for in-domain experiments and select the epoch (5) with the highest average across all conspiracy theories for all additional experiments.

The best hyperparameters for both models are listed in Table 8 in the Appendix A.

4.3 POS Tagging and Topic Modeling

POS tagging We use Stanza (Qi et al., 2020) and extract POS unigrams, bigrams, and trigrams; using a minimum frequency of 1 and absolute counts across our datasets.

Topic modeling To determine the most important words for a conspiracy theory, topics were extracted via topic modeling. We use LDA (Blei et al., 2003), set $N = 5$ (to represent the five conspiracies), and exclude stopwords[2] since a first run including stopwords showed a high number of stopwords in the topics word, most of them repeated among different topics.

We then extract the 20 highest ranked words (see Table 2). We can see that some of the conspiracies are clearly represented in a certain cluster, such as cluster one heavily containing words associated with vaccines while clusters three and five represent climate change. We assume that these highly associated words can hinder the ability to identify more in-domain conspiracies and use these words as a basis for bleaching experiments (see Section 5.3).

4.4 Evaluation

We report the F1 score on the test sets.

5 Results

5.1 In-Domain Experiments

We first experiment with an in-domain setting, i.e., we train and test on the same domain. This provides us with an upper bound in terms of how difficult the problem is and how much variation we can expect across the five conspiracy theories. We also use

[1] Prior experiments with BERT or using the first 1024 subtokens in BART resulted in lower scores.

[2] We use NLTK (Bird et al., 2009) stopwords and an additional set of common words not present in that base list.

cluster	topic words
1	vaccine vaccines health people children may virus also disease said one autism coronavirus 19 covid medical study vaccination cases flu
2	it people like that one think going re know we get said would you time there ve want go they
3	earth climate change years one warming global water could scientists also would world ice like planet sea time science new
4	trump one said news people us media it also world conspiracy would new like president time many clinton the state
5	climate change said world global new countries emissions also would health government year states energy china people economic public united

Table 2: Words associated with each LDA topic.

classifier	features	vaccine	climate change	pizzagate	flat earth	bigfoot
SVM	word	82.33	85.84	91.19	84.57	81.41
	char	88.30	84.05	92.38	85.28	83.67
	word+char	88.30	84.05	92.38	85.28	83.67
BART	word	**96.88**	**93.39**	**95.20**	**93.02**	**95.56**

Table 3: Results (F1) of in-domain experiments across 5 conspiracy theories.

these experiments to determine which classifiers work well for the problem and which features are useful, results of which are in Table 3.

For the SVM, word n-grams provide strong baselines, but most domains benefit from character embeddings, with vaccine seeing an almost 6% absolute increase, and only climate change showing a decrease about 1.8%. Interestingly, we see that character only and word+character features yield the same results. We assume that this indicates that character n-grams are more useful, as they are higher in frequency and capture many words at the subword level. BART has the highest overall performance, with bigfoot increasing almost 12% absolute over the word+char SVM experiment, and the variation across domains is reduced.

It is also obvious that different conspiracy theories provide various levels of difficulty, with vaccine generally being the easiest and climate change and flat earth being the most difficult ones for BART. However, we also see differences between the different classifiers and features. For the word-based SVM, for example, bigfoot seems to be the most difficult and pizzagate the easiest.

We experiment with feature selection for the word model as we assume that many n-grams will be of little use or misleading. We chose the word setting since this is the most explainable setting, and the setting that has the highest potential of improvement. Table 4 presents results for the feature selection experiments, with the 'all' setting containing all word features from Table 3 (approximately one million).

Results for feature selection do not show any clear tendencies, as three different trends emerge as the number of features are reduced: a trend towards a slight increase in performance (vaccine), a general decrease in performance (climate change and flat earth) and then a slight buoy effect with an increase then decrease (bigfoot). This suggests the optimal number of features for each domain is unique and we cannot generalize feature thresholds effectively.

5.2 Out of Domain: Comparing Source Domains

Table 5 shows the results when we train on one domain and classify out-of-domain texts. For ease of comparison, we repeat the in-domain results (underlined). In this setting, we either use a single conspiracy theory as training set, or we use a mix of the four conspiracy theories and test on the fifth. We assume that a mix of conspiracy theories may provide a more general basis in an out-of-domain setting. In order to avoid effects of training set size, we use quarter of the texts per conspiracy theory so that the mixed training set is similar in size to the individual sets.

no. features	vaccine	climate change	pizzagate	flat earth	bigfoot
all	82.33	**85.84**	91.19	**84.57**	81.41
3000	79.79	82.78	**91.50**	83.54	82.64
2000	83.18	80.16	87.73	76.33	**82.86**
1000	83.18	82.32	89.62	75.93	81.58
500	**83.54**	79.79	90.31	76.33	79.74

Table 4: Results (F1) of feature selection experiments using SVMs and word n-grams.

For most conspiracies, out-of-domain detection yields poorer performance compared to in-domain results, with some pairs exhibiting extreme drops of performance. For example, training on climate change and testing on pizzagate using word-based features in the SVM results in an F score of 53.17, as compared to 91.19 when testing on pizzagate in-domain. In general, the decrease is less pronounced for BART, with some exceptions. For example, when training on bigfoot and testing on pizzagate, the F score only reaches 69.00 while we reach 95.20 in-domain[3].

The best results overall are reached by BART. However, for climate change, flat earth, and bigfoot, we reach the best results when training on a single conspiracy theory. For vaccine, using a mix of conspiracy theories for training works better, and for pizzagate, both settings work equally well.

Overall, there is no clear trend concerning which conspiracy theory is best suited as training set in an out-of-domain setting. Even for a specific target domain, the best training domain varies based on the choice of classifier and features. For example, when testing on vaccine, the word-based SVM and BART prefer a mixed training set, while the character-based and char+word SVM prefer bigfoot.

For out-of-domain feature selection results, we see the same general trend as in Table 4 as performance not only drops across domains, but, in the majority of cases, a reduction of features yields even worse performance (for details see Table 9 in the Appendix B). Single out-of-domain conspiracy detection may simply not be highly detectable with small subsets of features due to the specific lexical co-occurrences within a specific domain. The mixed setting mostly gives the best results, either with all features (vaccine, pizzagate) or with 2000 features (climate change, flat earth); for bigfoot, the mixed results using all features are very close to the results using all features when training on vaccine. However, even in the mixed setting, we see a degradation in performance, even though this set should include a higher degree of lexical variation. This vocabulary seems to be specific to the source conspiracies, not a potentially evolving conspiracy.

5.3 Bleaching Features for Domain Adaptation

A classifier's generalizing ability in an out-of-domain setting can be affected by words that are good predictors for individual conspiracy theories. For example, the word 'Sasquatch' will be especially useful in identifying bigfoot conspiracy theory texts, but it will not be useful for pizzagate. For this reason, we need to create more abstract feature representations abstracting away from lexical information. One approach is bleaching, which aims to abstract meaning away from specific word features and to create more robust abstract features that may capture more meta or abstract characteristics of a text. While some bleaching techniques are focused on generating meta characteristics of words (e.g., how many alphanumeric characters) and have helped in cross-lingual gender prediction (van der Goot et al., 2018), we are more interested in lexical bleaching, similar to work by Tian and Kübler (2021), who bleached proper nouns for period classification of Chinese texts, by replacing them by their POS tags.

We chose to apply various levels of word bleaching: complete delexicalization (POS), content word bleaching, and topic word bleaching. In the delexicalization process, we utilized POS unigrams, bigrams, and trigrams instead of word n-grams. However, we assume that this form of bleaching will be too extensive, and that the POS features will not retain enough information for our task. Thus, for content word bleaching, we substituted nouns, verbs, adjectives, adverbs, and foreign words by their re-

[3] We acknowledge that overfitting may play a role in performance drops in out-of-domain settings. This is due to our experimental setting where we optimize the parameters in-domain, assuming it is infeasible to optimize for every test domain.

classifier	features	source	vaccine	climate change	pizzagate	flat earth	bigfoot
SVM	word	vaccine	82.33	73.07	82.12	74.32	80.34
		climate change	79.41	85.84	53.17	66.74	76.32
		pizzagate	74.26	61.91	91.19	64.06	66.96
		flat earth	70.92	73.64	82.32	84.57	70.48
		bigfoot	74.17	66.19	72.44	72.37	81.41
	char	vaccine	88.30	74.06	73.41	75.69	77.20
		climate change	73.52	84.05	51.88	63.93	74.74
		pizzagate	70.11	64.46	92.38	65.02	73.08
		flat earth	70.22	73.64	80.18	85.28	76.40
		bigfoot	74.51	69.84	63.84	72.49	83.67
	char+word	vaccine	88.30	74.80	73.41	75.66	77.20
		climate change	73.52	84.05	51.88	65.27	74.04
		pizzagate	69.07	61.93	92.38	62.96	73.08
		flat earth	71.43	76.76	81.75	85.28	75.66
		bigfoot	76.40	68.01	65.21	72.49	83.67
	word	mix	81.48	69.79	82.89	77.37	80.18
	char	mix	72.49	69.50	79.29	77.06	78.98
BART	word	vaccine	96.88	91.18	**90.79**	**93.02**	90.24
		climate change	95.01	93.39	80.26	88.57	90.36
		pizzagate	91.94	87.04	95.20	90.84	**93.70**
		flat earth	92.87	89.40	89.27	93.01	90.37
		bigfoot	94.91	**92.19**	69.00	86.50	95.56
	word	mix	**95.49**	88.05	**90.79**	89.71	91.90

Table 5: Results (F1) for out-of-domain experiments across 5 test CTs. In-domain results are underlined; best out-of-domain results are bolded.

spective POS tags. Again, this form of bleaching is less extreme than complete delexicalization, but it may still delete too many important lexical items. Thus, we investigate a third form of bleaching where we identify words that are typical for a conspiracy theory, and then only substitute those. For topic word bleaching, we use topic modeling to identify these CT specific words and substitute the words from Table 2.

Table 6 presents results for all bleaching experiments. For delexicalization, SVM results for in-domain experiments are substantially lower than the baseline word n-grams seen in Table 3. BART experiments show a more severe degradation, with F-scores ranging from 42.42 (flat earth) to 68.44 (climate change). This may be anticipated as an input of POS tags instead of words leads to a misalignment with the training words used to train the contextual embeddings. Then, the generated embeddings from the POS representations are most likely lower in quality and information. Our results suggest that the model cannot be fine-tuned on a more coarse-grained representation, which contradicts findings for cross-lingual zero-shot parsing using a multilingual language model (Zhou and Kübler, 2021).

For the out-of-domain experiments, in most cases, the POS setting still yields worse performance than the equivalent baseline experiments (Table 5), there is one exception: When training on pizzagate and testing on climate change, abstracting away from the lexical level can potentially help. Thus, overall, we conclude that POS tagging removes too much lexical content and leaves the classifier unable to distinguish conspiracy and mainstream texts.

For content word bleaching, we also see mixed results across settings in comparison to POS bleaching. For some domains, there is an increased performance across all settings (e.g., vaccine) while for others, there are mostly negative trends (e.g., pizzagate), and other domains show volatility in both directions (e.g., climate change). For BART, almost all settings show increased performance compared to POS representations, but they are all still substantially lower than their word experiment coun-

class.	features	source	vaccine	climate change	pizzagate	flat earth	bigfoot
SVM	POS	vaccine	71.79	64.06	69.10	66.19	63.43
		climate change	77.49	75.26	58.36	63.14	66.07
		pizzagate	66.41	66.35	81.58	63.16	66.03
		flat earth	67.71	60.93	62.46	73.80	72.79
		bigfoot	71.79	64.44	74.06	69.53	69.16
		mix	70.48	62.82	72.79	62.14	63.28
	content words	vaccine	80.16	70.05	70.24	68.94	65.10
		climate change	71.22	72.86	58.36	65.77	73.86
		pizzagate	66.30	59.47	82.86	60.30	58.87
		flat earth	72.86	66.24	74.80	75.54	71.92
		bigfoot	75.54	67.04	80.50	69.08	79.41
		mix	64.06	73.73	70.29	65.72	71.22
	topic words	vaccine	83.74	74.00	77.08	73.08	73.08
		climate change	82.55	88.69	60.40	71.07	75.55
		pizzagate	72.09	61.30	93.15	63.16	63.82
		flat earth	69.77	72.12	84.05	82.81	70.48
		bigfoot	72.66	66.96	71.07	70.98	79.05
		mix	77.95	72.91	82.36	71.92	77.53
BART	POS	vaccine	54.76	55.86	45.06	53.53	50.74
		climate change	54.23	68.44	45.06	56.50	55.85
		pizzagate	61.10	56.51	60.59	58.87	54.73
		flat earth	42.92	43.40	52.69	42.42	40.63
		bigfoot	55.87	57.00	44.74	50.06	44.80
		mix	56.43	43.40	45.06	48.26	49.61
	content words	vaccine	85.28	80.52	76.08	85.17	40.63
		climate change	73.80	83.25	75.66	80.24	65.74
		pizzagate	70.22	59.29	77.37	73.80	70.98
		flat earth	69.86	68.46	75.93	81.02	70.37
		bigfoot	75.37	81.02	83.37	75.54	74.80
		mix	81.49	82.33	74.32	84.73	74.21
	topic words	vaccine	96.88	**90.20**	69.16	**93.02**	86.87
		climate change	93.02	93.21	71.08	87.51	82.64
		pizzagate	83.67	81.41	96.19	86.77	85.26
		flat earth	86.43	87.72	89.79	91.77	85.90
		bigfoot	**93.69**	89.95	63.44	83.67	91.19
		mix	90.43	89.69	**90.48**	88.56	**87.66**

Table 6: Results (F1) of comparing bleaching methods for out-of-domain experiments. In-domain results are underlined; best out-of-domain results are bolded.

terparts.

Topic word bleaching shows some increased performances for SVM in in-domain settings not only over content words, but over the initial word n-gram SVM models, specifically for vaccine, climate change, and pizzagate. However, the words in Table 2 are heavily representative of these three conspiracies, not seemingly including words more associated with flat earth and bigfoot. It is an open question whether including more words associated with the latter CTs could yield improvements, or whether those CTs are less specific and do not have any clear topic words.

5.4 Text Length Distributions

One factor that may influence both in-domain and out-of-domain results is text length. Table 7 presents the means and standard deviations for both conspiracy and mainstream texts across domains. Some domains show rather large variations. For

	mainstream		conspiracy	
source	mean	stdev	mean	stdev
vaccine	836.89	879.80	1079.87	1112.09
climate change	949.67	1080.65	1085.83	1150.55
pizzagate	1031.49	1421.66	1504.92	1637.73
flat earth	849.93	985.01	1644.65	1622.10
bigfoot	886.80	1095.04	1693.90	1805.82

Table 7: Average length and standard deviation of the number of words per data set.

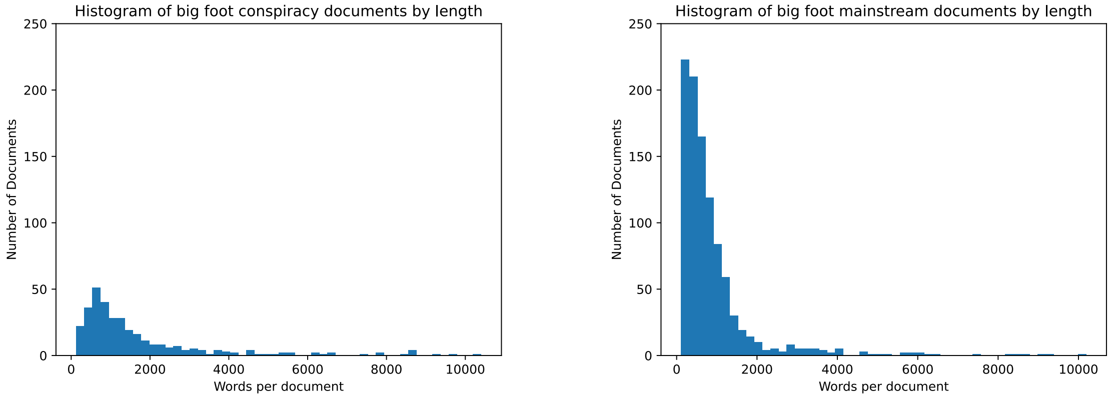

Figure 1: Text length distributions for bigfoot conspiracy (left) and mainstream (right) documents.

example, Figure 1 shows the text distributions for bigfoot between conspiracy and mainstream texts: The distribution of mainstream texts is heavily right skewed and of shorter lengths, while the bigfoot conspiracy texts are not as heavily right skewed and reflect a high average of text lengths: The average bigfoot conspiracy texts are almost twice as long as their mainstream counterparts.

Across domains, both mainstream and conspiracy texts also vary substantially, with vaccine texts having the shortest average length, pizzagate exhibiting longer mainstream texts, and bigfoot longer conspiracy texts. Similar trends are seen across the domains. One side effect of such distributions is that more information is contained in the conspiracy texts that may be relevant for identification than their mainstream counterparts, which, while shorter, are more frequent. This means we have data imbalance in both directions, both in the number of texts labeled conspiracy, and in the length of the texts, with conspiracy texts presumably containing more relevant information but spanning over longer contexts.

6 Conclusion

We presented a systematic set of experiments into how successfully we can classify conspiracy theories in both an in-domain and out-of-domain settings using different features and classifiers. Results showed, unsurprisingly, that while an SVM model presents strong baselines, a transformer-based model yields superior performance in both in-domain and out-of-domain settings. Of more interest though is that determining good source topics for detecting out-of-domain conspiracy theories is extremely difficult and not intuitive. It remains unclear what exactly the core semantic and structural relationships between conspiracies and mainstream texts are. While bleaching too much content (replacing all words or content words by POS tags) yields poor performance, bleaching typical words per conspiracy theory is promising.

One inherent difficulty that makes further in-depth analysis difficult is data quality of the automatically retrieved LOCO documents (Mompelat et al., 2022), which may hinder the efficacy of the resulting models. However, it is also clear that conspiracy theories are not as monolithic as assumed here. Research into the spread of conspiracy theories shows that people who believe in one conspiracy theory are also likely to believe in others, but not everybody believing in a CT will believe the same subset of factoids (Enders et al., 2021). This may also mean that the texts collected per CT are less homogeneous than necessary for classification.

Further research will need to investigate in more detail the inter-relatedness between different conspiracy theories. A better understanding of how they relate content-wise may allow us a better understanding of how to create a robust training set that can be used to detect conspiracy theories out of domain. Additionally, we are planning to investigate better bleaching methods, along with having a closer look at the SVM features that show the highest correlation with conspiracy theories, to determine defining characteristics of conspiratorial language across different domains.

7 Ethics Statement

Creating automated methods for detecting conspiratorial content in texts is always associated with the risk that the machine learner will learn and potentially amplify biases present in the training data. The LOCO corpus, which serves as the basis for our investigation, was collected automatically, using seed phrases. For this reason, it is unknown how well the data collection worked, and which biases the corpus contains. Mompelat et al. (2022) have shown that for at least one conspiracy theory, the mainstream collection of texts contains a non-trivial number of irrelevant texts. This can lead to a classifier that is more topics-based than focused on separating conspiracy theories from factual texts concerning similar topics. However, at this point of time, this corpus is the most extensive collection of texts that contains a range of conspiracy theories along with mainstream documents covering the same topics.

Acknowledgments

This work is based on research supported by US National Science Foundation (NSF) Grant #2123618. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *AAAI Conference on Artificial Intelligence*, Honolulu, HI.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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)*, pages 5636–5646, Hong Kong, China.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186, Minneapolis, MN.

Karen M. Douglas and Robbie M. Sutton. 2018. Why conspiracy theories matter: A social psychological analysis. *European Review of Social Psychology*, 29(1):256–298.

Adam Enders, Joseph Uscinski, Casey Klofstad, Michelle Seelig, Stefan Wuchty, Manohar Murthi, Kamal Premaratne, and John Funchion. 2021. Do conspiracy beliefs form a belief system? Examining the structure and organization of conspiracy beliefs. *Journal of Social and Political Psychology*, 9(1):255–271.

Adam Enders, Joseph Uscinski, Casey Klofstad, Stefan Wuchty, Michelle Seelig, John Funchion, Manohar N. Murthi, Kamal Premaratne, and Justin Stoler. 2022. Who supports QAnon? A case study in political extremism. *The Journal of Politics*, 84(3):1844–1849.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 383–389, Melbourne, Australia.

Colin Klein, Peter Clutton, and Adam G Dunn. 2019. Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum. *PloS one*, 14(11):e0225098.

Emily Klein and James Hendler. 2022. Loaded language and conspiracy theorizing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, pages 2671–2679, Toronto, Canada.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020.

BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. *ArXiv*, abs/1910.09982.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2021. LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*.

Ludovic Mompelat, Zuoyu Tian, Amanda Kessler, Matthew Luettgen, Aaryana Rajanala, Sandra Kübler, and Michelle Seelig. 2022. How “loco” is the LOCO corpus? Annotating the language of conspiracy theories. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI)*, pages 111–119, Marseille, France.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Samantha C. Phillips, Lynnette Hui Xian Ng, and Kathleen M. Carley. 2022. Hoaxes and hidden agendas: A Twitter conspiracy theory dataset. In *Companion Proceedings of the Web Conference (WWW’22)*, page 876–880.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Zuoyu Tian and Sandra Kübler. 2021. Period classification in Chinese historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Punta Cana, Dominican Republic.

Michael J Wood, Karen M Douglas, and Robbie M Sutton. 2012. Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science*, 3(6):767–773.

He Zhou and Sandra Kübler. 2021. Delexicalized cross-lingual dependency parsing for Xibe. In *Proceedings of the Conference on Recent Advances in NLP (RANLP)*, Online.

Appendix A

SVM	kernel	linear
	loss	squared hinge
	C	0.01
BART	model	facebook/bart-base
	batch size	2
	optimizer	adam
	lr	$1 * 10^{-5}$
	epochs	5

Table 8: Fine-tuned parameters for the SVM and BART.

Appendix B

source	no. features	vaccine	climate change	pizzagate	flat earth	bigfoot
vaccine	all	82.33	73.07	82.12	74.32	**80.34**
	3000	79.79	64.37	76.76	70.22	78.39
	2000	83.18	66.93	77.89	75.31	76.40
	1000	83.18	67.98	76.04	71.19	78.29
	500	83.54	67.62	77.08	71.79	76.42
climate change	all	79.41	85.84	53.17	66.74	76.32
	3000	73.79	82.78	55.48	64.75	74.15
	2000	69.77	80.16	49.61	64.14	75.06
	1000	75.23	82.32	54.76	68.43	71.38
	500	71.75	79.79	52.24	65.76	76.23
pizzagate	all	74.26	61.91	91.19	64.06	66.96
	3000	69.16	57.09	91.50	64.37	70.03
	2000	69.16	55.93	87.73	65.02	69.61
	1000	69.53	58.23	89.62	70.56	68.68
	500	69.36	56.77	90.31	65.83	60.01
flat earth	all	70.92	73.64	82.32	84.57	70.48
	3000	60.51	67.62	74.60	83.54	71.92
	2000	63.28	70.55	79.41	76.33	74.06
	1000	61.72	69.06	76.76	75.93	77.53
	500	59.49	68.33	75.32	76.33	72.62
bigfoot	all	74.17	66.19	72.44	72.37	81.41
	3000	72.77	67.71	62.47	59.86	82.64
	2000	73.64	68.17	62.12	61.10	82.86
	1000	72.77	67.82	61.09	63.81	81.58
	500	71.79	72.86	58.31	57.70	79.74
mix	all	**81.48**	69.79	**82.89**	77.37	80.18
	3000	73.50	72.87	72.83	74.76	71.22
	2000	75.66	**74.58**	71.07	**76.08**	74.06
	1000	73.79	69.86	72.44	69.87	76.29
	500	72.34	66.95	72.83	66.41	73.79

Table 9: Results (F1) of feature selection out-of-domain using SVMs and word n-grams.