Text classification can distinguish mainstream and fringe scientific papers

(anonymized in compliance with journal submission requirements)

Abstract

In this work, I explore the use of supervised learning in distinguishing mainstream and fringe scientific papers. This work has two goals. The first is to determine whether mainstream and fringe scientific papers can be reliably distinguished through automated means. The second is to determine whether classifiers trained using stylometric features, such as word count, average word and sentences lengths, and frequencies of part-of-speech sequences, can outperform conventional *n*-gram document models in classifying papers across scientific topics. I conduct a systematic study of the ability of classifiers to distinguish mainstream and fringe scientific papers across topics, for example by training a classifier on biophysics papers and testing it against cosmology papers. The term-based and style-based approaches both perform significantly better than chance, with neither approach consistently outperforming the other. Classifiers trained using the combined feature set (i.e., *n*-gram frequencies *and* stylometric features) perform little better than those trained only on one or the other feature set, suggesting that the two feature sets are, in aggregate, highly correlated. Overall, the results of this work suggest that mainstream and fringe scientific papers are readily distinguishable by conventional text classification methods.

1 Introduction

The detection of fringe and poor-quality scientific papers remains an outstanding problem in science and scientific publishing (Bohannon, 2013; Labbé and Labbé, 2013; Moher et al., 2017). Top researchers may occasionally fall prey to predatory journals (Seethapathy et al., 2016; Moher et al., 2017), suggesting that the proliferation of predatory journals and poor-quality research online may be especially troublesome for those researchers who are not well connected to the mainstream scientific community in their field (Seethapathy et al., 2016). To the extent that scientists intentionally publish in so-called "pay-for-play" journals, an automated, scalable means of identifying regular publishers of poor-quality scientific writing could help to discourage such behavior.

Automated text classification methods have been applied toward classifying a variety of document types (Sebastiani, 2002), including spam email (Guzella and Caminhas, 2009; Ren and Ji, 2017), fake news (Shu et al., 2017; Tacchini et al., 2017), hate speech (Saleem et al., 2017), and sensitive government documents (McDonald et al., 2015). Text classification has also been used extensively for authorship attribution and characterization (Abbasi and Chen, 2005; Kucukyilmaz et al., 2008; Cheng et al., 2011; Brocardo et al., 2013). With some exception, this often involves the use of document models derived from *n*-gram frequencies (Tamboli and Prasad, 2013).

So far, attempts at automated identification of unreliable information such as poor quality scientific papers often rely on compiling lists of authors or publishers designated as suspicious[1,2]. These require that human judgments be made on a paper-by-paper or journal-by-journal basis. Kelk and Devine (2012) find average differences in several author and paper attributes between mainstream and fringe scientific papers. This result, along with the success of authorship characterization and other text classification efforts in other fields, suggest that it may be possible to automate the process of identifying fringe or unreliable scientific work. This in turn could aid in the identification of predatory journals at scale, as well as facilitate the curation of papers on pre-print servers that lack peer review but also wish to maintain a

[1] https://beallslist.weebly.com/

[2] http://bsdetector.tech/

certain minimum of scientific standards.

In the current work, I explore the feasibility of automating the identification of fringe scientific papers. Specifically, I use supervised learning methods to determine whether fringe scientific writing can be red-flagged based on either or both of two types of paper attributes: one based on the frequencies of words and short phrases (n-grams), the other based on stylometry (i.e., attributes related to document structure, sentence structure, usage patterns of common words, etc.). I compare the two kinds of attributes in part to test the hypothesis that classifiers trained on stylometric features will perform better when tested across topics (e.g., trained on biophysics papers and tested against cosmology papers). In Section 2, I discuss the dataset, feature extraction, and classification and validation scheme. Results are discussed in Section 3, and conclusions, including possible avenues for future work, are discussed in Section 4.

2 Methods

2.1 Dataset

Papers studied in this analysis come from two websites, arXiv[3] and viXra[4]. The arXiv site is a preprint server used by professional scientists in several fields including physics, astronomy, and mathematics. The viXra site is formatted as a spin-off site set up by authors rejected by the arXiv's moderators [5]. Authors who publish on viXra are less likely to be university-affiliated scientists, and 65% of arXiv papers as opposed to 15% of viXra papers are published in journals (Kelk and Devine, 2012). Titles found on viXra include

- "Michelson and Morley Experiment Does not Validate Length Contraction" (refuting Einstein is a common theme among viXra's physical science papers);
- "Long Term Stability and the Meaning of Life"; and
- "The Purpose of Nature is Super - Intelligence".

Although the terminology of "mainstream" versus "fringe" scientific writing used in the current work entails some presumptions about the typical scientific quality and authorship intent of arXiv and viXra papers, it must be emphasized that the analysis carried out in the current work does not and cannot determine the scientific validity of a given paper. Rather, the goal of this study is only to test the hypothesis that arXiv and viXra papers indeed reflect two or more different populations of writing style.

As viXra is a spin-off of arXiv, its categorization scheme is similar, though not identical, to that of arXiv, enabling apples-to-apples comparisons of mainstream and fringe scientific papers for several topics. To construct the dataset, I take papers listed under the following categories, where n denotes the number of papers selected from *each* site:

- biophysics ("bio-ph" on arXiv; "Physics of Biology" on viXra; $n \approx 300$),
- cosmology ("gr-qc" on arXiv; "Relativity and Cosmology" on viXra; $n \approx 2{,}100$),
- high-energy physics ("hep-ex/hep-lat/hep-ph/hep-th" on arXiv; "High Energy Particle Physics" on viXra; $n \approx 1{,}000$), and
- mathematics ("math" on arXiv; "General Mathematics" on viXra; $n \approx 200$).

The number of papers in each category is chosen such that the two sites are evenly represented. In all four categories, there are fewer papers on viXra than on arXiv, therefore I retrieved all available viXra papers for these four categories as of the time that this dataset was aggregated, and I selected the same number of papers from arXiv that were published around the same time. A small handful of papers was excluded from the analysis due to unextractable PDFs or other problematic features that precluded automated analysis.

The incorporation of multiple scientific topics in this analysis helps to accommodate the possibility that the arXiv–viXra category pairs given above are not entirely substantively equivalent. It could be argued for instance that viXra "Physics of Biology" papers address different topics than arXiv "bio-ph" papers; any automated detection of a difference in writing style could then be attributed to differences in topic. However, if classifiers trained on cosmology papers can effectively distinguish arXiv and viXra papers in biophysics, high-energy physics, and mathematics, then that would be strong evidence that the differences between arXiv and viXra papers detected in this

[3] http://arxiv.org

[4] http://vixra.org

[5] http://vixra.org/why

analysis are truly a reflection of different broad modes of writing style between mainstream and fringe scientific papers.

2.2 Feature extraction

I calculate two sets of attributes for the papers under consideration and conduct separate classification analyses, as well as an analysis in which the two feature sets are combined. The first feature set is generated from word *n*-grams. Specifically, the top 1,000 *n*-grams where $n \in \{1, 2, 3\}$ are selected by term frequency–inverse document frequency (TFIDF) value, with the arXiv and viXra paper sets each constituting a single "master" document. That is, an *n*-gram achieves a high TFIDF value by being found frequently in arXiv papers but infrequently in viXra papers, or vice-versa. Once those features are selected, term frequencies (TF) are tabulated for each document; preliminary experimentation revealed poorer classification performance when using TFIDF values compared with TF.

The second feature set is based on a number of stylometric attributes. The stylometric attributes calculated for each paper come in four broad categories, the derivation of which are further described below. These categories are

- "general" paper attributes, including number of authors, number of references to other papers, average sentence length, average word size, and average rate of word repetition;
- frequencies of parts of speech and short sequences of parts of speech;
- graph metrics based on word co-occurrence networks; and
- frequencies of character unigrams and bigrams, punctuation marks, and common words.

The analysis of Kelk and Devine (2012) suggests that certain attributes of scientific papers may be useful in distinguishing arXiv and viXra papers, such as number of authors, number of references to other works, and so on. The number of authors for each paper is determined using metadata from each website; all other attributes are derived from the text of each paper. Papers from both websites are downloaded in PDF form and converted to ASCII text using version 3.03 of the "pdftotext" tool in Linux (written by Glyph & Cog, LLC.), which handles the text extraction very well. Papers for which PDF extraction fails (e.g., due to rasterized PDFs or problematic character encoding) are identified based on tell-tale features of the outputs (e.g., by unusually short document lengths, or by space characters that separate every letter in a given document) and are discarded from the dataset. Prior to feature extraction, the watermark found on the first page of every arXiv paper is automatically removed. I also calculate the frequencies of character unigrams and bigrams for each paper, several punctuation marks, and common words such as "the," "and," and so on, as well as words common in scientific papers across many subjects such as "equation," "table," and "figure."

Penn Treebank part-of-speech tags are tagged for each paper using the Natural Language Toolkit in Python (Taylor et al., 2003; Loper and Bird, 2002; Bird, 2006). Frequencies of parts of speech and short sequences of parts of speech up to four long are calculated for each paper. In order to limit the dimensionality of the reduced dataset, part-of-speech sequences two to four in length are ranked by their tendency to be found either in arXiv or viXra papers. Specifically, each sequence is scored by the difference in its overall frequency between arXiv and viXra papers (across all four categories). A strongly positive score reflects a part of speech sequence commonly used in arXiv papers but not in viXra papers, while a strongly negative score reflects a sequence used commonly in viXra but not arXiv papers. The sequences with the hundred most positive and the hundred most negative scores are included in the reduced dataset. All individual parts of speech are included.

The frequencies of named entities are tabulated for each paper, separately for people, places, and organizations, tagged using the spaCy toolkit[6]. The TextBlob package[7] is used to derive polarization and subjectivity scores for each sentence; the mean and standard deviation of sentence polarities and subjectivities are then calculated for each paper.

For each paper, I generate two graphs where the nodes represent unique decapitalized words. In one graph, edges are drawn between words located next to each other in a sentence. In the other graph, edges are drawn between words found

[6] https://spacy.io/

[7] https://textblob.readthedocs.io/en/dev/

in the same sentences. In both graphs, edges are weighted according to the frequency of co-occurrence. A number of connectivity and clustering metrics are calculated for these word co-occurrence graphs using the NetworkX package (Hagberg et al., 2008).

2.3 Classification and validation

I train classifiers on papers from each of the four categories under consideration and test them against papers from each of the other categories. For comparison, I also train and test classifiers within each category. This yields $4 \times 4 = 16$ comparisons total. For each comparison, I use 50–50 train–test splits validation, conducted with randomized splits six times, both to optimize hyperparameters when training within each category and to estimate the consistency of classification when testing against another category. Results are discussed in Section 3.1. I also calculate average feature importances based on random forest classifiers (Chen and Guestrin, 2016) trained in this way. Results are discussed in Section 3.2.1.

In order to explore the relative importances of the various subsets of stylometric features, I produce a dataset wherein all four scientific categories are equally represented, and train and test classifiers within that dataset using the aforementioned repeat-stratified validation scheme. Results of this analysis are described in Section 3.2.2.

The classifiers tested in the current work are XGBoost (Chen and Guestrin, 2016), decision tree (Breiman et al., 1984), naive Bayes (Manning et al., 2008), logistic regression (M.D., 1944; Liu and Nocedal, 1989), and random forest (Breiman, 2001). All of the classifiers used come from the scikit-learn package[8], with the exception of XGBoost, which was implemented through its built-in scikit-learn wrapper.

In order to prevent overfitting, key hyperparameters for each algorithm are optimized via grid search over a nominal range of values. Models are selected by $F_{1,\text{macro}}$, the harmonic mean of the F_1 scores for arXiv and viXra papers. One parameter tuned for all five algorithms is the percentage of features to be used, where features are selected by ANOVA F-value. Preliminary experiments with dimensionality reduction by ANOVA F-value filtering and by principal component analysis showed that dimensionality reduction makes at most a marginal difference to overall classifier performance.

3 Results

3.1 Training and testing across topics

Tables 1–3 show $F_{1,\text{macro}}$ and macro-average AUROC score, $\text{AUROC}_{\text{macro}}$, for the best-performing (highest $F_{1,\text{macro}}$) classifier trained on papers in each scientific topic when tested against papers in each other topic. Each of the two tables also shows the results of testing within each category as an estimate of the maximum achievable performance. Table 1 shows these results for classifiers trained on the *n*-gram feature set, Table 2 for the stylometric feature set, and Table 3 for the combined feature set. For all of the best-performing classifiers, the accuracy scores (not shown) are very close to the $F_{1,\text{macro}}$ scores, indicating even levels of precision and recall both for arXiv and viXra papers.

The best-performing classifiers are a mix of random forest, XGBoost, and logistic regression. It is interesting to note that classifiers trained on the combined feature set (Table 3) do not significantly outperform those trained only on *n*-gram frequencies (Table 1) or only on stylometric attributes (Table 2). Ostensibly, there are two plausible hypotheses to explain this. One is that some papers on each site are simply so reminiscent of papers on the other site that there is a fundamental limit on the extent to which machine learning techniques can distinguish them.

The second possibility is that *n*-gram frequencies and stylometric attributes are overall highly correlated, and therefore ultimately provide the same information, which is consistent with their similar performance. The first hypothesis is unlikely given that classifiers trained and tested within categories typically achieve upward of 95% $F_{1,\text{macro}}$ scores, showing that the information is in principle present for distinguishing arXiv and viXra papers. It is therefore likely that *n*-gram frequencies and stylometric features are, in aggregate, highly correlated. Based on the dataset under consideration, it seems that stylometric features, some of which are moderately computationally intensive to calculate, do not provide substantially greater insight into distinguishing arXiv and viXra papers compared with *n*-gram frequencies alone.

The less sophisticated classifiers tested in this work, naive Bayes and decision tree, showed little

[8] http://scikit-learn.org

Train set	Test set	Best classifier	$F_{1,\text{macro}}$	AUROC$_{\text{macro}}$
Biophysics	Biophysics	Logistic regression	0.959 (0.018)	0.988 (0.009)
Biophysics	Cosmology	Random forest	0.890 (0.013)	0.954 (0.008)
Biophysics	High-energy	Logistic regression	0.894 (0.009)	0.955 (0.007)
Biophysics	Math	Logistic regression	0.801 (0.045)	0.901 (0.034)
Cosmology	Biophysics	XGBoost	0.953 (0.009)	0.991 (0.002)
Cosmology	Cosmology	XGBoost	0.965 (0.007)	0.994 (0.002)
Cosmology	High-energy	XGBoost	0.914 (0.013)	0.984 (0.005)
Cosmology	Math	Logistic regression	0.855 (0.048)	0.937 (0.038)
High-energy	Biophysics	XGBoost	0.935 (0.021)	0.988 (0.005)
High-energy	Cosmology	XGBoost	0.935 (0.007)	0.982 (0.003)
High-energy	High-energy	XGBoost	0.965 (0.007)	0.994 (0.002)
High-energy	Math	XGBoost	0.852 (0.047)	0.927 (0.032)
Math	Biophysics	XGBoost	0.871 (0.032)	0.918 (0.030)
Math	Cosmology	XGBoost	0.794 (0.010)	0.900 (0.009)
Math	High-energy	XGBoost	0.759 (0.010)	0.846 (0.012)
Math	Math	Logistic regression	0.951 (0.025)	0.986 (0.016)

Table 1: Best $F_{1,\text{macro}}$ and AUROC$_{\text{macro}}$ scores among classifiers trained across scientific subjects, using the n-gram feature set. Numbers in parentheses represent 2σ values based on repeated 50–50 train–test splits.

Train set	Test set	Best classifier	$F_{1,\text{macro}}$	AUROC$_{\text{macro}}$
Biophysics	Biophysics	XGBoost	0.955 (0.020)	0.991 (0.006)
Biophysics	Cosmology	Random forest	0.898 (0.006)	0.966 (0.002)
Biophysics	High-energy	XGBoost	0.872 (0.015)	0.953 (0.007)
Biophysics	Math	Random forest	0.743 (0.058)	0.859 (0.050)
Cosmology	Biophysics	XGBoost	0.951 (0.010)	0.992 (0.004)
Cosmology	Cosmology	XGBoost	0.960 (0.006)	0.992 (0.003)
Cosmology	High-energy	XGBoost	0.912 (0.016)	0.981 (0.006)
Cosmology	Math	Random forest	0.833 (0.019)	0.925 (0.017)
High-energy	Biophysics	XGBoost	0.953 (0.025)	0.992 (0.006)
High-energy	Cosmology	XGBoost	0.948 (0.005)	0.987 (0.002)
High-energy	High-energy	XGBoost	0.949 (0.012)	0.989 (0.005)
High-energy	Math	Random forest	0.819 (0.041)	0.898 (0.040)
Math	Biophysics	Random forest	0.806 (0.025)	0.879 (0.028)
Math	Cosmology	Random forest	0.816 (0.010)	0.900 (0.006)
Math	High-energy	Random forest	0.749 (0.016)	0.810 (0.019)
Math	Math	Logistic regression	0.936 (0.040)	0.983 (0.015)

Table 2: Best $F_{1,\text{macro}}$ and AUROC$_{\text{macro}}$ scores among classifiers trained across scientific subjects, using the stylometric feature set. Numbers in parentheses represent 2σ values based on repeated 50–50 train–test splits.

Train set	Test set	Best classifier	$F_{1,\text{macro}}$	$\text{AUROC}_{\text{macro}}$
Biophysics	Biophysics	Logistic regression	0.959 (0.023)	0.987 (0.008)
Biophysics	Cosmology	XGBoost	0.898 (0.009)	0.965 (0.005)
Biophysics	High-energy	XGBoost	0.896 (0.016)	0.970 (0.006)
Biophysics	Math	Logistic regression	0.801 (0.044)	0.901 (0.032)
Cosmology	Biophysics	XGBoost	0.960 (0.015)	0.995 (0.004)
Cosmology	Cosmology	XGBoost	0.969 (0.005)	0.995 (0.002)
Cosmology	High-energy	XGBoost	0.929 (0.011)	0.989 (0.003)
Cosmology	Math	Logistic regression	0.861 (0.029)	0.940 (0.018)
High-energy	Biophysics	XGBoost	0.958 (0.019)	0.995 (0.003)
High-energy	Cosmology	XGBoost	0.953 (0.006)	0.990 (0.002)
High-energy	High-energy	XGBoost	0.967 (0.011)	0.995 (0.002)
High-energy	Math	XGBoost	0.843 (0.039)	0.928 (0.025)
Math	Biophysics	XGBoost	0.829 (0.024)	0.889 (0.015)
Math	Cosmology	XGBoost	0.798 (0.011)	0.901 (0.010)
Math	High-energy	XGBoost	0.749 (0.028)	0.824 (0.031)
Math	Math	Logistic regression	0.948 (0.037)	0.982 (0.010)

Table 3: Best $F_{1,\text{macro}}$ and $\text{AUROC}_{\text{macro}}$ scores among classifiers trained across scientific subjects, using the combined feature set. Numbers in parentheses represent 2σ values based on repeated 50–50 train–test splits.

difference in performance between the two feature sets. The only exception is that, when training and testing against mathematics papers, naive Bayes classifiers perform moderately better when using the *n*-grams feature set compared with the stylometric feature set. This indicates that the stylometric features, in addition to not providing any overall advantage in terms of maximum achievable performance, do not appear to provide any computational advantage in terms of allowing for less computationally intensive algorithms to be used.

3.2 Stylometric features

3.2.1 Feature importances

It is worth examining which stylometric features are most informative as to the differences between arXiv and viXra papers discerned by the trained classifiers discussed in Section 3.1. Using the random forest classifiers trained within each scientific topic, I calculate the importance of each feature, the value of which is derived as follows. In each tree generated by each instance of the algorithm, the importance of a feature is equal to the Gini impurity at a given node that splits on that feature, weighted by the number of data points handled by that node. For each feature, this value is averaged over all of the trees, then normalized so that the total importance of all of the features is equal to unity. The overall importance of each feature is finally derived by averaging its importance within the multiple train–test splits done for each topic, then taking the average of its importance across the four topics that is weighted by the number of papers in each topic. The top 25 features are shown in Table 4, along with the website having the higher average value of each of those features.

Interestingly, the most informative stylometric features for distinguishing arXiv and viXra papers come from different categories: part-of-speech sequence frequencies, frequencies of certain common words, and miscellaneous attributes like number of authors and number of references. Summing over the importance values listed, these 25 features account for ~40% of the total feature importance.

Perhaps not surprisingly, arXiv papers have more authors, references to other papers, and words overall than viXra papers. Papers on viXra, on the other hand, are for instance more likely to use apostrophes and question marks. This may reflect a less formal style where contractions are more frequently used, and where scientific or guiding questions are more frequently posed as literal questions. A sampling of sentences in viXra papers that end with question marks includes, “How is DNA searched to arrive at a transcription pattern?” “When I was in my twenties I worked on the problem: What is Life?” and “Does this Gaian perspective represent a return to paganism, the worship of natural rather than spiri-

Feature	Importance	Site with higher avg. value
Number of authors	0.0540	arXiv
Apostrophe frequency	0.0242	viXra
Number of references	0.0235	arXiv
“s” word frequency	0.0211	arXiv
POS frequency: CC-NNP-NNP	0.0206	arXiv
“d” word frequency	0.0201	arXiv
POS frequency: NNP-CC-NNP-NNP	0.0199	arXiv
POS frequency: POS	0.0183	viXra
POS frequency: NNP-NNP-CC-NNP	0.0167	arXiv
“J” character frequency	0.0141	arXiv
Question mark frequency	0.0131	viXra
POS frequency: CC-NNP-NNP-NNP	0.0123	arXiv
Personal entity frequency	0.0122	arXiv
“http” word frequency	0.0118	viXra
“we” word frequency	0.0118	arXiv
POS frequency: NNP-NNP-NNP-CC	0.0116	arXiv
POS frequency: NNP-POS	0.0113	viXra
POS frequency: CD-JJ-NNP-NNP	0.0113	arXiv
“ht” character frequency	0.0100	viXra
POS frequency: CC-NNP	0.0096	arXiv
Number of words	0.0094	arXiv
Sentence co-occurrence graph transitivity	0.0092	viXra
“m” word frequency	0.0088	arXiv
“xr” character frequency	0.0087	viXra
POS frequency: SYM	0.0085	arXiv

Table 4: Top 25 features by feature importance, averaged over random forest models trained within each scientific topic.

tual forces?" Within viXra papers, there is a moderately bimodal distribution in the frequency of question mark usage, indicating that certain papers, and perhaps certain authors, tend to pose many questions, whereas other viXra papers are more conventional in their technical writing style in this regard, asking few or no questions.

The diversity of attribute types that are useful in distinguishing arXiv and viXra papers is tentative evidence that a number of stylometric analytic approaches may be useful in distinguishing various types of documents in other domains. This is important in domains where certain approaches may not be available or may not be as useful. For instance, authorship attribution or other classification tasks for social media posts where there are built-in word limits will not likely make use of attributes like number of words, but other attributes based on parts of speech or sentiment analysis may still hold useful.

3.2.2 Feature subsets

In order to help further determine which stylometric features are the most informative, I use the combined dataset in which all four categories of scientific paper are equally represented to perform classification analysis within four separate feature subsets: general attributes (such as number of authors, number of words, and number of references to other papers); frequencies of part-of-speech sequences; character, punctuation, and common word frequencies; and word co-occurrence graph metrics. For comparison, I also perform the same analysis using the full attribute set.

Table 5 shows $F_{1,\text{macro}}$ and $\text{AUROC}_{\text{macro}}$ scores for the best-performing classifiers on each subset of stylometric features, along with the full stylometric feature set. Random forest and XGBoost are the strongest performers. The likelihood of misclassification for the various algorithms tested is fairly even for all four categories and both websites, with a slight tendency for math papers to be most commonly misclassified.

Classifiers trained on each of these four feature subsets perform nearly as well as those trained on the full stylometric attribute set. The subset of attributes based on word co-occurrence graphs contains fewer attributes than the other subsets, and provides only general insight into sentence and document structure, which may help to explain the relatively poor performance of classifiers trained on this subset compared with the others. For all of the feature subsets, the classifiers achieve nearly even rates of precision and recall for papers from both sites.

4 Conclusions and future work

The results of this work demonstrate that mainstream and fringe scientific writing are readily distinguishable through a number of types of attributes. While stylometric attributes are effective in this classification task, they prove to be no more effective than more conventional methods (i.e., *n*-gram document modeling). Stylometric attributes also show no indication of enabling greater computational efficiency, in that they do not perform any better than *n*-gram models for simpler algorithms like naive Bayes and decision tree.

It should be reiterated that to the extent that the methods used in this work successfully distinguish mainstream and fringe scientific writing, it is not because these methods have provided any direct assessment of the reliability of claims being made in any given paper. The attributes used in the above analysis are likely relevant to distinguishing arXiv and viXra papers for a variety of reasons. Certain jargon or aspects of writing style may reflect ideological attitudes or goals, while others might reflect an author's experience level with technical writing, education level, or depth of familiarity with some given subject matter.

One limitation on the extensibility of the current results is that the examples of fringe scientific papers are drawn only from a single website. While the similarity in organization between arXiv and viXra makes for convenient comparisons, and while viXra's papers are written across a range of topics by many authors, it is possible that viXra's papers are only limitedly representative of fringe scientific writing in general. The relatively narrow range of scientific subject matter of papers analyzed in the current work is another limitation. Future work should explore the utility of stylometric analysis in classifying papers across topics more disparate than, say, cosmology and high-energy physics, where term frequencies may become less reliable for distinguishing mainstream and fringe papers.

In addition to the classification of individual papers, the methods used in this study may prove useful for identifying predatory and other poor-quality journals, as well as automatically flagging unwanted submissions on pre-print servers like

Feature subset	Best classifier	$F_{1,\text{macro}}$	$\text{AUROC}_{\text{macro}}$
All	XGBoost	0.970 (0.014)	0.995 (0.004)
General	Random forest	0.925 (0.054)	0.978 (0.028)
Parts of speech	Random forest	0.955 (0.008)	0.991 (0.003)
Characters	XGBoost	0.967 (0.015)	0.994 (0.005)
Graph	Random forest	0.872 (0.060)	0.951 (0.039)

Table 5: $F_{1,\text{macro}}$ and $\text{AUROC}_{\text{macro}}$ scores for best-performing classifiers (by $F_{1,\text{macro}}$) trained on subsets of stylometric features in a dataset where all four categories are equally represented. Numbers in parentheses represent 2σ values based on repeated 50–50 train–test splits.

arXiv. In the future, automated tools may help researchers to identify journals that regularly publish papers with writing styles corresponding with those published on websites like viXra. One interesting area of future research would be to apply automated algorithms such as those discussed in this work to papers published by journals and publishers identified as predatory, such as those included in Beall's List[9].

Identifying author attributes and attitudes through writing style may be of interest to consumers and decision makers in domains outside of academic science. It is conceivable for instance that fake news articles reflect different modes of writing style compared with ordinary news articles in similar ways to those observed in viXra papers in the current work. It is possible that fake news articles written with certain political goals in mind are more likely to contain superlative adjectives and exclamation marks. The quite generic nature of the attributes used in this study leaves open the possibility that they can be used to distinguish fake news and other unreliable writing that might be revealed through proxies of author attitudes, subject matter background, and so on. Future work should explore the question of whether stylometric document models have an advantage over more directly topic-based models like n-grams in domains that span a wider range of subject matter than that covered by the four categories of scientific paper examined in the current work.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Steven Bird. 2006. Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Bohannon. 2013. Who's afraid of peer review? *Science*, 342(6154):60–65.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A.

M. L. Brocardo, I. Traore, S. Saad, and I. Woungang. 2013. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Na Cheng, R. Chandramouli, and K.P. Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78 – 88.

Thiago S. Guzella and Walmir M. Caminhas. 2009. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206 – 10222.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Confer-*

[9] https://beallslist.weebly.com/

ence (SciPy2008), pages 11–15, Pasadena, CA USA.

David Kelk and David Devine. 2012. A scienceographic comparison of physics papers from the arxiv and vixra archives. *CoRR*, abs/1211.1036.

Tayfun Kucukyilmaz, B. Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448 – 1466.

Cyril Labbé and Dominique Labbé. 2013. Duplicate and fake publications in the scientific literature: how many scigen papers in computer science? *Scientometrics*, 94(1):379–396.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2015. Using part-of-speech n-grams for sensitive-text classification. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 381–384, New York, NY, USA. ACM.

Joseph Berkson M.D. 1944. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

David Moher, Larissa Shamseer, Kelly D. Cobey, Manoj M. Lalu, James Galipeau, Marc T. Avey, Nadera Ahmadzai, Mostafa Alabousi, Pauline Barbeau, Andrew Beck, Raymond Daniel, Robert Frank, Mona Ghannad, Candyce Hamel, Mona Hersi, Brian Hutton, Inga Isupov, Trevor A. McGrath, Matthew D. F. McInnes, Matthew J. Page, Misty Pratt, Kusala Pussegoda, Beverley Shea, Anubhav Srivastava, Adrienne Stevens, Kednapa Thavorn, Sasha van Katwyk, Roxanne Ward, Dianna Wolfe, Fatemeh Yazdi, Ashley M. Yu, and Hedyeh Ziai. 2017. Stop this waste of people, animals and money. *Nature*, 549(7670):23–25.

Yafeng Ren and Donghong Ji. 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385:213 – 224.

H. M. Saleem, K. P Dillon, S. Benesch, and D. Ruths. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *ArXiv e-prints*.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Gopalakrishnan Saroja Seethapathy, J. U. Santhosh Kumar, and A. S. Hareesha. 2016. India's scientific publication in predatory journals: need for regulating quality of indian science and education. *Current Science*, 111(11):1759–1764.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ArXiv e-prints*.

E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. *ArXiv e-prints*.

Mubin Shaukat Tamboli and Rajesh S. Prasad. 2013. Article: Authorship analysis and identification techniques: A review. *International Journal of Computer Applications*, 77(16):11–15. Full text available.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. *The Penn Treebank: An Overview*. Springer Netherlands, Dordrecht.