

A Stylometric Inquiry into Hyperpartisan and Fake News

Martin Potthast Johannes Kiesel Kevin Reinartz Janek Bevendorff Benno Stein

Leipzig University
martin.potthast@uni-leipzig.de

Bauhaus-Universität Weimar
<first>.<last>@uni-weimar.de

Abstract

We report on a comparative style analysis of hyperpartisan (extremely one-sided) news and fake news. A corpus of 1,627 articles from 9 political publishers, three each from the mainstream, the hyperpartisan left, and the hyperpartisan right, have been fact-checked by professional journalists at BuzzFeed: 97% of the 299 fake news articles identified are also hyperpartisan. We show how a style analysis can distinguish hyperpartisan news from the mainstream ($F_1 = 0.78$), and satire from both ($F_1 = 0.81$). But stylometry is no silver bullet as style-based fake news detection does not work ($F_1 = 0.46$). We further reveal that left-wing and right-wing news share significantly more stylistic similarities than either does with the mainstream. This result is robust: it has been confirmed by three different modeling approaches, one of which employs Unmasking in a novel way. Applications of our results include partisanship detection and pre-screening for semi-automatic fake news detection.

1 Introduction

The media and the public are currently discussing the recent phenomenon of “fake news” and its potential role in swaying elections, how it may affect society, and what can and should be done about it. Prone to misunderstanding and misuse, the term “fake news” arose from the observation that, in social media, a certain kind of ‘news’ spreads much more successfully than others, and this kind of ‘news’ is typically extremely one-sided (hyperpartisan), inflammatory, emotional, and often riddled with untruths. Although traditional yellow press has been spreading ‘news’ of varying de-

grees of truthfulness long before the digital revolution, its amplification over *real* news within social media gives many people pause. The fake news hype caused a widespread disillusionment about social media, and many politicians, news publishers, IT companies, activists, and scientists concur that this is where to draw the line. For all their good intentions, however, it must be drawn very carefully (if at all), since nothing less than free speech is at stake—a fundamental right of every free society.

Many favor a two-step approach where fake news items are detected and then countermeasures are implemented to foreclose rumors and to discourage repetition. While some countermeasures are already tried in practice, such as displaying warnings and withholding ad revenue, fake news detection is still in its infancy. At any rate, a near-real time reaction is crucial: once a fake news item begins to spread virally, the damage is done and undoing it becomes arduous. Since knowledge-based and context-based approaches to fake news detection can only be applied after publication, i.e., as news events unfold and as social interactions occur, they may not be fast enough.

We have identified style-based approaches as a viable alternative, allowing for instantaneous reactions, albeit not to fake news, but to hyperpartisanship. In this regard we contribute (1) a large news corpus annotated by experts with respect to veracity and hyperpartisanship, (2) extensive experiments on discriminating fake news, hyperpartisan news, and satire based solely on writing style, and (3) validation experiments to verify our finding that the writing style of the left and the right have more in common than any of the two have with the mainstream, applying Unmasking in a novel way.

After a review of related work, Section 3 details the corpus and its construction, Section 4 introduces our methodology, and Section 5 reports the results of the aforementioned experiments.

2 Related Work

Approaches to fake news detection divide into three categories (Figure 1): they can be knowledge-based (by relating to known facts), context-based (by analyzing news spread in social media), and style-based (by analyzing writing style).

Knowledge-based fake news detection. Methods from information retrieval have been proposed early on to determine the veracity of web documents. For example, [Etzioni et al. \(2008\)](#) propose to identify inconsistencies by matching claims extracted from the web with those of a document in question. Similarly, [Magdy and Wanas \(2010\)](#) measure the frequency of documents that support a claim. Both approaches face the challenges of web data credibility, namely expertise, trustworthiness, quality, and reliability ([Ginsca et al., 2015](#)).

Other approaches rely on knowledge bases, including the semantic web and linked open data. [Wu et al. \(2014\)](#) “perturb” a claim in question to query knowledge bases, using the result variations as indicator of the support a knowledge base offers for the claim. [Ciampaglia et al. \(2015\)](#) use the shortest path between concepts in a knowledge graph, whereas [Shi and Weninger \(2016\)](#) use a link prediction algorithm. However, these approaches are unsuited for new claims without corresponding entries in a knowledge base, whereas knowledge bases can be manipulated ([Heindorf et al., 2016](#)).

Context-based fake news detection. Here, fake news items are identified via meta information and spread patterns. For example, [Long et al. \(2017\)](#) show that author information can be a useful feature for fake news detection, and [Derczynski et al. \(2017\)](#) attempt to determine the veracity of a claim based on the conversation it sparks on Twitter as one of the RumourEval tasks. The Facebook analysis of [Mocanu et al. \(2015\)](#) shows that unsubstantiated claims spread as widely as well-established ones, and that user groups predisposed to conspiracy theories are more open to sharing the former. Similarly, [Acemoglu et al. \(2010\)](#), [Kwon et al. \(2013\)](#), [Ma et al. \(2017\)](#), and [Volkova et al. \(2017\)](#) model the spread of (mis-)information, while [Budak et al. \(2011\)](#) and [Nguyen et al. \(2012\)](#) propose algorithms to limit its spread. The efficacy of countermeasures like debunking sites is studied by [Tambuscio et al. \(2015\)](#). While achieving good results, context-based approaches suffer from working only a posteriori, requiring large amounts of data, and disregarding the actual news content.

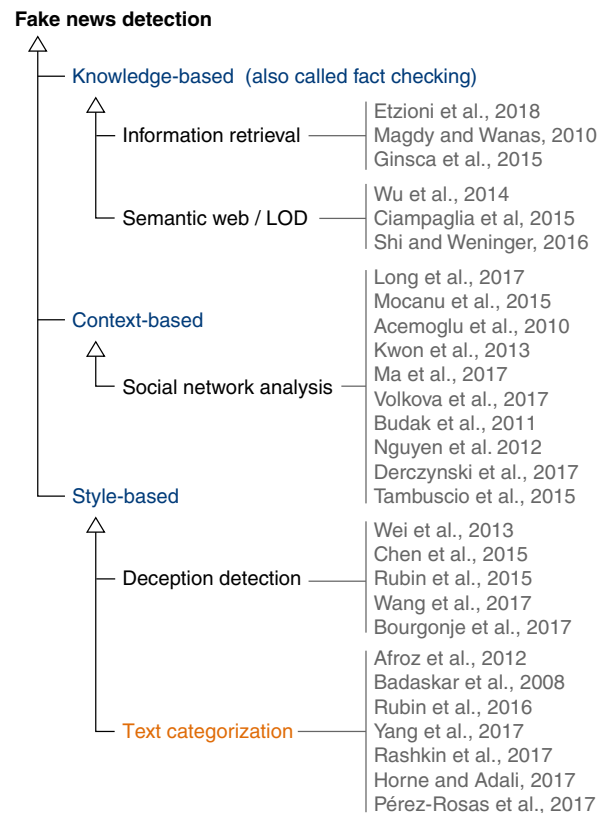


Figure 1: Taxonomy of paradigms for fake news detection alongside a selection of related work.

Style-based fake news detection. Deception detection originates from forensic linguistics and builds on the Undeutsch hypothesis—a result from forensic psychology which asserts that memories of real-life, self-experienced events differ in content and quality from imagined events ([Undeutsch, 1967](#)). The hypothesis led to the development of forensic tools to assess testimonies at the statement level. Some approaches operationalize deception detection at scale to detect uncertainty in social media posts, for example [Wei et al. \(2013\)](#) and [Chen et al. \(2015\)](#). In this regard, [Rubin et al. \(2015\)](#) use rhetorical structure theory as a measure of story coherence and as an indicator for fake news. Recently, [Wang \(2017\)](#) collected a large dataset consisting of sentence-length statements along their veracity from the fact-checking site PolitiFact.com, and then used style features to detect false statements. A related task is stance detection, where the goal is to detect the relation between a claim about an article, and the article itself ([Bourgonje et al., 2017](#)). Most prominently, stance detection was the task of the Fake News Challenge¹ which ran in 2017 and received 50 submissions, albeit hardly any participants published their approach.

¹<http://www.fakenewschallenge.org/>


Where deception detection focuses on single statements, style-based text categorization as proposed by Argamon-Engelson et al. (1998) assesses entire texts. Common applications are author profiling (age, gender, etc.) and genre classification. Though susceptible to authors who can modify their writing style, such obfuscations may be detectable (e.g., Afroz et al. (2012)). As an early precursor to fake news detection, Badaskar et al. (2008) train models to identify news items that were automatically generated. Currently, text categorization methods for fake news detection focus mostly on satire detection (e.g., Rubin et al. (2016), Yang et al. (2017)). Rashkin et al. (2017) perform a statistical analysis of the stylistic differences between real, satire, hoax, and propaganda news. We make use of their results by incorporating the best-performing style features identified.

Finally, two preprint papers have been recently shared. Horne and Adali (2017) use style features for fake news detection. However, the relatively high accuracies reported must be taken with a grain of salt: their two datasets comprise only 70 news articles each, whose ground-truth is based on where an article came from, instead of resulting from a per-article expert review as in our case; their final classifier uses only 4 features (number of nouns, type-token ratio, word count, number of quotes), which can be easily manipulated; and based on their experimental setup, it cannot be ruled out that the classifier simply differentiates news portals rather than fake and real articles. We avoid this problem by testing our classifiers on articles from portals which were not represented in the training data. Similarly, Pérez-Rosas et al. (2017) also report on constructing two datasets comprising around 240 and 200 news article *excerpts* (i.e., the 5-sentence lead) with a balanced distribution of fake vs. real. The former was collected via crowdsourcing, asking workers to write a fake news item based on a real news item, the latter was collected from the web. For style analysis, the former dataset may not be suitable, since the authors note themselves that “workers succeeded in mimicking the reporting style from the original news”. The latter dataset encompasses only celebrity news (i.e., yellow press), which introduces a bias. Their feature selection follows that of Rubin et al. (2016), which is covered by our experiments, but also incorporates topic features, rendering the resulting classifier not generalizable.

3 The BuzzFeed-Webis Fake News Corpus

This section introduces the BuzzFeed-Webis Fake News Corpus 2016, detailing its construction and annotation by professional journalists employed at BuzzFeed, as well as key figures and statistics.²

3.1 Corpus Construction

The corpus encompasses the output of 9 publishers on 7 workdays close to the US presidential elections 2016, namely September 19 to 23, 26, and 27. Table 1 gives an overview. Among the selected publishers are six prolific hyperpartisan ones (three left-wing and three right-wing), and three mainstream ones. All publishers earned Facebook’s blue checkmark , indicating authenticity and an elevated status within the network. Every post and linked news article has been fact-checked by 4 BuzzFeed journalists, including about 19% of posts forwarded from third parties. Having checked a total of 2,282 posts, 1,145 mainstream, 471 left-wing, and 666 right-wing, Silverman et al. (2016) reported key insights as a data journalism article. The annotations were published alongside the article.³ However, this data only comprises URLs to the original Facebook posts. To construct our corpus, we archived the posts, the linked articles, and attached media as well as relevant meta data to ensure long-term availability. Due to the rapid pace at which the publishers change their websites, we were able to recover only 1,627 articles, 826 mainstream, 256 left-wing, and 545 right-wing.

Manual fact-checking. A binary distinction between fake and real news turned out to be infeasible, since hardly any piece of fake news is entirely false, and pieces of real news may not be flawless. Therefore, posts were rated “mostly true,” “mixture of true and false,” “mostly false,” or, if the post was opinion-driven or otherwise lacked a factual claim, “no factual content.” Four BuzzFeed journalists worked on the manual fact-checks of the news articles: to minimize costs, each article was reviewed only once and articles were assigned round robin. The ratings “mixture of true and false” and “mostly false” had to be justified, and, when in doubt about a rating, a second opinion was collected, whereas disagreements were resolved by a third one. Finally, all news rated “mostly false” underwent a final check to ensure the rating was justified, lest the respective publishers would contest it.

²Corpus download: <https://doi.org/10.5281/zenodo.1239675>

³<http://github.com/BuzzFeedNews/2016-10-facebook-fact-check>

The journalists were given the following guidance:

Mostly true: The post and any related link or image are based on factual information and portray it accurately. The authors may interpret the event/info in their own way, so long as they do not misrepresent events, numbers, quotes, reactions, etc., or make information up. This rating does not allow for unsupported speculation or claims.

Mixture of true and false (mix, for short): Some elements of the information are factually accurate, but some elements or claims are not. This rating should be used when speculation or unfounded claims are mixed with real events, numbers, quotes, etc., or when the headline of the link being shared makes a false claim but the text of the story is largely accurate. It should also only be used when the unsupported or false information is roughly equal to the accurate information in the post or link. Finally, use this rating for news articles that are based on unconfirmed information.

Mostly false: Most or all of the information in the post or in the link being shared is inaccurate. This should also be used when the central claim being made is false.

No factual content (n/a, for short): This rating is used for posts that are pure opinion, comics, satire, or any other posts that do not make a factual claim. This is also the category to use for posts that are of the “Like this if you think...” variety.

3.2 Limitations

Given the significant workload (i.e., costs) required to carry out the aforementioned annotations, the corpus is restricted to the given temporal period and biased toward the US culture and political landscape, comprising only English news articles from a limited number of publishers. Annotations were recorded at the article level, not at statement level. For text categorization, this is sufficient. At the time of writing, our corpus is the largest of its kind that has been annotated by professional journalists.

3.3 Corpus Statistics

Table 1 shows the fact-checking results and some key statistics per article. Unsurprisingly, none of the mainstream articles are mostly false, whereas 8 across all three publishers are a mixture of true and false. Disregarding non-factual articles, a little more than a quarter of all hyperpartisan left-wing articles were found faulty: 15 articles mostly false, and 51 a mixture of true and false. Publisher “The Other 98%” sticks out by achieving an almost per-

Orientation Publisher	Fact-checking results					Key statistics per article				
	true	mix	false	n/a	Σ	Paras.		Links		Words
						extern	all	quoted	all	
<i>Mainstream</i>	806	8	0	12	826	20.1	2.2	3.7	18.1	692.0
ABC News	90	2	0	3	95	21.1	1.0	4.8	21.0	551.9
CNN	295	4	0	8	307	19.3	2.4	2.5	15.3	588.3
Politico	421	2	0	1	424	20.5	2.3	4.3	19.9	798.5
<i>Left-wing</i>	182	51	15	8	256	14.6	4.5	4.9	28.6	423.2
Addicting Info	95	25	8	7	135	15.9	4.4	4.5	30.5	430.5
Occupy Democrats	55	23	6	0	91	10.9	4.1	4.7	29.0	421.7
The Other 98%	32	3	1	1	30	20.2	6.4	7.2	21.2	394.5
<i>Right-wing</i>	276	153	72	44	545	14.1	2.5	3.1	24.6	397.4
Eagle Rising	107	47	25	36	214	12.9	2.6	2.8	17.3	388.3
Freedom Daily	48	24	22	4	99	14.6	2.2	2.3	23.5	419.3
Right Wing News	121	82	25	4	232	15.0	2.5	3.6	33.6	396.6
Σ	1264	212	87	64	1627	17.2	2.7	3.7	20.6	551.0

Table 1: The BuzzFeed-Webis Fake News Corpus 2016 at a glance (“Paras.” short for “paragraphs”).

fect score. By contrast, almost 45% of the right-wing articles are a mixture of true and false (153) or mostly false (72). Here, publisher “Right Wing News” sticks out by supplying more than half of mixtures of true and false alone, whereas mostly false articles are equally distributed.

Regarding key statistics per article, it is interesting that the articles from all mainstream publishers are on average about 20 paragraphs long with word counts ranging from 550 words on average at ABC News to 800 at Politico. Except for one publisher, left-wing articles and right-wing articles are shorter on average in terms of paragraphs as well as word count, averaging at about 420 words and 400 words, respectively. Left-wing articles quote on average about 10 words more than the mainstream, and right-wing articles 6 words more. When articles comprise links, they are usually external ones, whereas ABC News rather uses internal links, and only half of the links found at Politico articles are external. Left-wing news articles stick out by containing almost double the amount of links across publishers than mainstream and right-wing ones.

3.4 Operationalizing Fake News

In our experiments, we operationalize the category of fake news by joining the articles that were rated mostly false with those rated a mixture of true and false. Arguably, the latter may not be exactly what is deemed “fake news” (as in: a complete fabrication), however, practice shows fake news are hardly ever devoid of truth. More often, true facts are misconstrued or framed badly. In our experiments, we hence call mostly true articles real news, mostly false plus mixtures of true and false—except for satire—fake news, and disregard all articles rated non-factual.

4 Methodology

This section covers our methodology, including our feature set to capture writing style, and a brief recap of Unmasking by Koppel et al. (2007), which we employ for the first time to distinguish genre styles as opposed to author styles. For sake of reproducibility, all our code has been published.⁴

4.1 Style Features and Feature Selection

Our writing style model incorporates common features as well as ones specific to the news domain. The former are n-grams, n in [1, 3], of characters, stop words, and parts-of-speech. Further, we employ 10 readability scores⁵ and dictionary features, each indicating the frequency of words from a tailor-made dictionary in a document, using the General Inquirer Dictionaries as a basis (Stone et al., 1966). The domain-specific features include ratios of quoted words and external links, the number of paragraphs, and their average length.

In each of our experiments, we carefully select from the aforementioned features the ones worthwhile using: all features are discarded that are hardly represented in our corpus, namely word tokens that occur in less than 2.5% of the documents, and n-gram features that occur in less than 10% of the documents. Discarding these features prevents overfitting and improves the chances that our model will generalize.

If not stated otherwise, our experiments share a common setup. In order to avoid biases from the respective training sets, we balance them using oversampling. Furthermore, we perform 3-fold cross-validation where each fold comprises one publisher from each orientation, so that the classifier does not learn a publisher’s style. For non-Unmasking experiments we use WEKA’s random forest implementation with default settings.

4.2 Unmasking Genre Styles

Unmasking, as proposed by Koppel et al. (2007), is a meta learning approach for authorship verification. We study for the first time whether it can be used to assess the similarity of more broadly defined style categories, such as left-wing vs. right-wing vs. mainstream news. This way, we uncover relations between the writing styles that people may involuntarily adopt as per their political orientation.

⁴Code download: <http://www.github.com/webis-de/ACL-18>

⁵Automated Readability Index, Coleman Liau Index, Flesh Kincaid Grade Level and Reading Ease, Gunning Fog Index, LIX, McAlpine EFLAW Score, RIX, SMOG Grade, Strain Index

Originally, Unmasking takes two documents as input and outputs its confidence whether they have been written by the same author. Three steps are taken to accomplish this: first, each document is chunked into a set of at least 500-word long chunks; second, classification errors are measured while iteratively removing the most discriminative features of a style model consisting of the 250 most frequent words, separating the two chunk sets with a linear classifier; and third, the resulting classification accuracy curves are analyzed with regard to their slope. A steep decrease is more likely than a shallow decrease if the two documents have been written by the same author, since there are presumably less discriminating features between documents written by the same author than between documents written by different authors. Training a classifier on many examples of error curves obtained from same-author document pairs and different-author document pairs yields an effective authorship verifier—at least for long documents that can be split up into a sufficient number of chunks.

It turns out that what applies to the style of authors also applies to genre styles. We adapt Unmasking by skipping its first step and using two sets of documents (e.g., left-wing articles and right-wing articles) as input. When plotting classification error curves for visual inspection, steeper decreases in these plots, too, indicate higher style similarity of the two input document sets, just as with chunk sets of two documents written by the same author.

4.3 Baselines

We employ four baseline models: a topic-based bag of words model, often used in the literature, but less practical since news topics change frequently and drastically; a model using only the domain-specific news style features to check whether the differences between categories measured as corpus statistics play a significant role; and naive baselines that classify all items into one of the categories in question, relating our results to the class distributions.

4.4 Performance Measures

Classification performance is measured as accuracy, and class-wise precision, recall, and F_1 . We favor these measures over, e.g., areas under the ROC curve or the precision recall curve for simplicity sake. Also, the tasks we are tackling are new, so that little is known to date about user preferences. This is also why we chose the evenly-balanced F_1 .

5 Experiments

We report on the results of two series of experiments that investigate style differences and similarities between hyperpartisan and mainstream news, and between fake, real, and satire news, shedding light on the following questions:

1. Can (left/right) hyperpartisanship be distinguished from the mainstream?
2. Is style-based fake news detection feasible?
3. Can fake news be distinguished from satire?

Our first experiment addressing the first question uncovered an odd behavior of our classifier: it would often misjudge left-wing for right-wing news, while being much better at distinguishing both combined from the mainstream. To explain this behavior, we hypothesized that *maybe* the writing style of the hyperpartisan left and right are more similar to one another than to the mainstream. To investigate this hypothesis, we devised two additional validation experiments, yielding three sources of evidence instead of just one.

5.1 Hyperpartisanship vs. Mainstream

A. Predicting orientation. Table 2 shows the classification performance of a ternary classifier trained to discriminate left, right, and mainstream—an obvious first experiment for our dataset. Separating the left and right orientation from the mainstream does not work too well: the topic baseline outperforms the style-based models with regard to accuracy, whereas the results for class-wise precision and recall are a mixed bag. The left-wing articles are apparently significantly more difficult to be identified compared to articles from the other two orientations. When we inspected the confusion matrix (not shown), it turned out that 66% of misclassifications of left-wing articles are falsely classified as right-wing articles, whereas 60% of all misclassified right-wing articles are classified as mainstream articles. Misclassified mainstream articles spread almost evenly across the other classes.

The poor performance of the domain-specific news style features by themselves demonstrate that orientation cannot be discriminated based on the basic corpus characteristics observed with respect to paragraphs, quotations, and hyperlinks. This holds for all subsequent experiments.

B. Predicting hyperpartisanship. Given the apparent difficulty of telling apart individual orientations, we did not frantically add features or switch classifiers to make it work. Rather, we trained a binary

Features	Accuracy	Precision		Recall			F ₁			
	all	left	right main.	left	right main.	left	right main.	left	right main.	
Style	0.60	0.21	0.56	0.75	0.20	0.59	0.74	0.20	0.57	0.75
Topic	0.64	0.24	0.62	0.72	0.15	0.54	0.86	0.19	0.58	0.79
News style	0.39	0.09	0.35	0.59	0.14	0.36	0.49	0.11	0.36	0.53
All-left	0.16	0.16	-	-	1.00	0.0	0.0	0.27	-	-
All-right	0.33	-	0.33	-	0.0	1.00	0.0	-	0.50	-
All-main.	0.51	-	-	0.51	0.0	0.0	1.00	-	-	0.68

Table 2: Performance of predicting orientation.

Features	Accuracy	Precision		Recall		F ₁	
	all	hyp.	main.	hyp.	main.	hyp.	main.
Style	0.75	0.69	0.86	0.89	0.62	0.78	0.72
Topic	0.71	0.66	0.79	0.83	0.60	0.74	0.68
News style	0.56	0.54	0.58	0.65	0.47	0.59	0.52
All-hyp.	0.49	0.49	-	1.00	0.0	0.66	-
All-main.	0.51	-	0.51	0.0	1.00	-	0.68

Table 3: Performance of predicting hyperpartisanship.

Features	Left		Right	
	Trained on: right+main.	all	left+main.	all
Style	0.74	0.90	0.66	0.89
Topic	0.68	0.79	0.48	0.85
News style	0.52	0.61	0.47	0.66

Table 4: Ratio of left articles misclassified right when omitting left articles from training, and vice versa.

classifier to discriminate hyperpartisanship in general from the mainstream. Table 3 shows the performance values. This time, the best classification accuracy of 0.75 at a remarkable 0.89 recall for the hyperpartisan class is achieved by the style-based classifier, outperforming the topic baseline.

Comparing Table 2 and Table 3, we were left with a riddle: all other things being equal, how could it be that hyperpartisanship in general can be much better discriminated from the mainstream than individual orientation? Attempts to answer this question gave rise to our aforementioned hypothesis that, perhaps, the writing style of hyperpartisan left and right are not altogether different, despite their opposing agendas. Or put another way, if style and topic are orthogonal concepts, then being an extremist should not exert a different style dependent on political orientation. Excited, we sought ways to *independently* disprove the hypothesis, and found two: Experiments C and D.

C. Validation using leave-out classification. If left-wing and right-wing articles have a more similar style than either of them compared to mainstream articles, then what class would a binary classifier assign to a left-wing article, if it were trained to distinguish only the right-wing from the mainstream, and vice versa? Table 4 shows the results of this experiment. As indicated by proportions well above 0.50, full style-based classifiers have a tendency of clas-

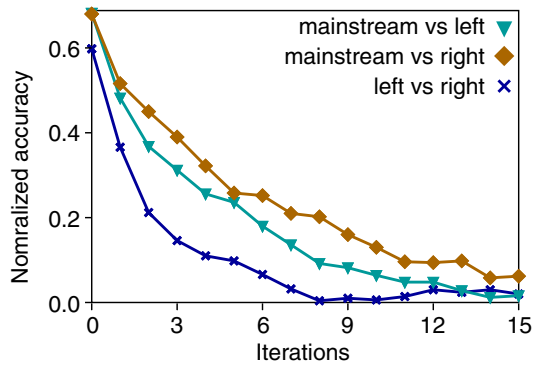


Figure 2: Unmasking applied to pairs of political orientations. The steeper a curve, the more similar the respective styles.

sifying left as right and right as left. The topic baseline, though, gets confused especially when omitting right articles from the training set with performance close to random. The fact that the topic baseline works better when omitting left from the training set may be explainable: leading up to the elections, the hyperpartisan left was often merely reacting to topics prompted by the hyperpartisan right, instead of bringing up their own.

D. Validation using Unmasking. Based on [Koppel et al.](#)'s original approach in the context of authorship verification, for the first time, we generalize Unmasking to assess genre styles: just like author style similarity, genre style similarity will be characterized by the slope of a given Unmasking curve, where a steeper decrease indicates higher similarity. We apply Unmasking as described in Section 4.2 onto pairs of sets of left, right, and mainstream articles. Figure 2 shows the resulting Unmasking curves (Unmasking is symmetrical, hence three curves). The curves are averaged over 5 runs, where each run comprised sets of 100 articles from each orientation. In case of the left-wing orientation, where less than 500 articles are available in our corpus, once all of them had been used, they were shuffled again to select articles for the remainder of the runs. As can be seen, the curve comparing left vs. right has a distinctly steeper slope than either of the others. This result hence matches the findings of the previous experiments.

With caution, we conclude that the evidence gained from our three independent experimental setups supports our hypothesis that the hyperpartisan left and the hyperpartisan right have more in common in terms of writing style than any of the two have with the mainstream. Another more tangible (e.g., practical) outcome of Experiment B is the finding that hyperpartisan news can apparently be

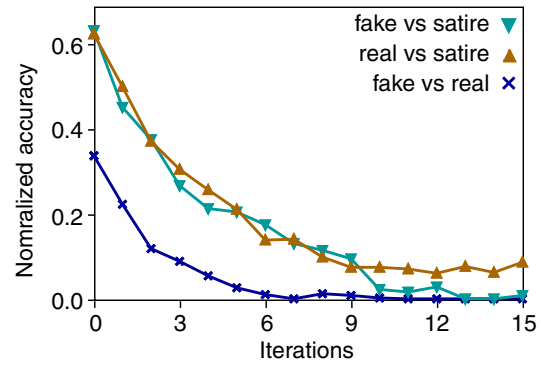


Figure 3: Unmasking applied to pairs of sets of news that are fake, real, and satire.

discriminated well from the mainstream: in particular the high recall of 0.89 at a reasonable precision of 0.69 gives us confidence that, with some further effort, a practical classifier can be built that detects hyperpartisan news at scale and in real time, since an article's style can be assessed immediately without referring to external information.

5.2 Fake vs. Real (vs. Satire)

This series of experiments targets research questions (2) and (3). Again, we conduct three experiments, where the first is about predicting veracity, and the last two about discriminating satire.

A. Predicting veracity. When taking into account that the mainstream news publishers in our corpus did not publish any news items that are mostly false, and only very few instances that are mixtures of true and false, we may safely disregard them for the task of fake news detection. A reliable classifier for hyperpartisan news can act as a pre-filter for a subsequent, more in-depth fake news detection approach, which may in turn be tailored to a much more narrowly defined classification task. We hence use only the left-wing articles and the right-wing articles of our corpus for our attempt at a style-based fake news classifier.

Table 5 shows the performance values for a generic classifier that predicts fake news across orientations, and orientation-specific classifiers that have been individually trained on articles from either orientation. Although all classifiers outperform the naive baselines of classifying everything into one of the classes in terms of precision, the slight increase comes at the cost of a large decrease in recall. While the orientation-specific classifiers are slightly better for most metrics, none of them outperform the naive baselines regarding the F -Measure. We conclude that style-based fake news classification simply does not work in general.

Features	Accuracy		Precision		Recall		F ₁	
	all	fake	real	fake	real	fake	real	
<i>Generic classifier</i>								
Style	0.55	0.42	0.62	0.41	0.64	0.41	0.63	
Topic	0.52	0.41	0.62	0.48	0.55	0.44	0.58	
<i>Orientation-specific classifier</i>								
Style	0.55	0.43	0.64	0.49	0.59	0.46	0.61	
Topic	0.58	0.46	0.65	0.45	0.66	0.46	0.66	
All-fake	0.39	0.39	-	1.00	0.0	0.56	-	
All-real	0.61	-	0.61	0.0	1.00	-	0.76	

Table 5: Performance of predicting veracity.

Features	Accuracy		Precision		Recall		F ₁	
	all	sat.	real	sat.	real	sat.	real	
Style	0.82	0.84	0.80	0.78	0.85	0.81	0.82	
Topic	0.77	0.78	0.75	0.74	0.79	0.76	0.77	
All-sat.	0.50	0.50	-	1.00	0.0	0.67	-	
All-real	0.50	-	0.50	0.00	1.00	-	0.67	
Rubin et al.	n/a	0.90	n/a	0.84	n/a	0.87	n/a	

Table 6: Performance of predicting satire (sat.).

B. Predicting satire. Yet, not all fake news are the same. One should distinguish satire from the rest, which takes the form of news but lies more or less obviously to amuse its readers. Regardless the problems that spreading fake news may cause, satire should never be filtered, but be discriminated from other fakes. Table 6 shows the performance values of our classifier in the satire-detection setting used by Rubin et al. (2016) (the S-n-L News DB corpus), distinguishing satire from real news. This setting uses a balanced 3:1 training-to-test set split over 360 articles (180 per class). As can be seen, our style-based model significantly outperforms all baselines across the board, achieving an accuracy of 0.82, and an F score of 0.81. It clearly improves over topic classification, but does not outperform Rubin et al.’s classifier, which includes features based on topic, absurdity, grammar, and punctuation. We argue that incorporating topic into satire detection is not appropriate, since the topics of satire change along the topics of news. A classifier with topic features therefore does not generalize. Apparently, a style-based model is competitive, and we believe that satire can be detected at scale this way, so as to prevent other fake news detection technology from falsely filtering it.

C. Unmasking satire. Given the above results on stylistic similarities between left and right news, the question remains how satire fits into the picture. We assess the style similarity of satire from Rubin et al.’s corpus compared to fake news and real news from ours, again applying Unmasking to compare pairs of the three categories of news as described above. Figure 3 shows the resulting Un-

masking curves. The curve for the pair of fake vs. real news drops faster compared to the other two pairs. Apparently, the style of fake news has more in common with that of real news than either of the two have with satire. These results are encouraging: satire is distinct enough from fake and real news, so that, just like with hyperpartisan news compared to mainstream news, it can be discriminated with reasonable accuracy.

6 Conclusion

Fact-checking for fake news detection poses an interdisciplinary challenge: technology is required to extract factual statements from text, to match facts with a knowledge base, to dynamically retrieve and maintain knowledge bases from the web, to reliably assess the overall veracity of an entire article rather than individual statements, to do so in real time as news events unfold, to monitor the spread of fake news within and across social media, to measure the reputation of information sources, and to raise awareness in readers. These are only the most salient things that need be done to tackle the problem, and as our cross-section of related work shows, a large body of work must be covered. Notwithstanding the many attacks on fake news by developing one way or another of fact-checking, we believe it worthwhile to mount our attack from another angle: writing style.

We show that news articles conveying a hyperpartisan world view can be distinguished from more balanced news by writing style alone. Moreover, for the first time, we found quantifiable evidence that the writing styles of news of the two opposing orientations are in fact very similar: there appears to be a common writing style of left and right extremism. We further show that satire can be distinguished well from other news, ensuring that humor will not be outcast by fake news detection technology. All of these results offer new, tangible, short-term avenues of development, lest large-scale fact-checking is still far out of reach. Employed as pre-filtering technologies to separate hyperpartisan news from mainstream news, our approach allows for directing the attention of human fact checkers to the most likely sources of fake news.

Acknowledgements

We thank Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine from BuzzFeed for making their data available, enabling our research.

References

- Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. Spread of (Mis)Information in Social Networks. *Games and Economic Behavior*, 70(2):194–227.
- Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. [Detecting Hoaxes, Frauds, and Deception in Writing Style Online](#). In *2012 IEEE Symposium on Security and Privacy*, pages 461–475.
- Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4.
- Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. [Identifying real or fake articles: Towards better language modeling](#). In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 817–822. The Association for Computer Linguistics.
- Peter Bourgonje, Julián Moreno Schneider, and Georg Rehm. 2017. [From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles](#). In *Proceedings of the 2017 Workshop: Natural Language Processing meets Journalism, NLPmJ@EMNLP, Copenhagen, Denmark, September 7, 2017*, pages 84–89.
- Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. [Limiting the spread of misinformation in social networks](#). In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 665–674, New York, NY, USA. ACM.
- Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. [News in an Online World: The Need for an "Automatic Crap Detector"](#). In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, ASIST '15*, pages 81:1–81:4, Silver Springs, MD, USA. American Society for Information Science.
- Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational Fact Checking from Knowledge Networks. *PLoS one*, 10(6):e0128193.
- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [Semeval-2017 task 8: Rumoureal: Determining rumour veracity and support for rumours](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 69–76.
- Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. [Open Information Extraction from the Web](#). *Commun. ACM*, 51(12):68–74.
- Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. 2015. [Credibility in Information Retrieval](#). *Found. Trends Inf. Retr.*, 9(5):355–475.
- Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. 2016. [Vandalism Detection in Wikidata](#). In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 16)*, pages 327–336. ACM.
- Benjamin D. Horne and Sibel Adali. 2017. [This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news](#). *CoRR*, abs/1703.09398.
- Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *J. Mach. Learn. Res.*, 8:1261–1276.
- Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent Features of Rumor Propagation in Online Social Media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1103–1108. IEEE.
- Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. [Fake news detection through multi-perspective speaker profiles](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 252–256.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. [Detect rumors in microblog posts using propagation structure via kernel learning](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 708–717.
- Amr Magdy and Nayer Wanas. 2010. [Web-based Statistical Fact Checking of Textual Documents](#). In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents, SMUC '10*, pages 103–110, New York, NY, USA. ACM.
- Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. 2015. [Collective Attention in the Age of \(Mis\)Information](#). *Comput. Hum. Behav.*, 51(PB):1198–1204.
- Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. 2012. [Containment of Misinformation Spread in Online Social Networks](#). In *Proceedings of the 4th Annual ACM Web Science Conference, WebSci '12*, pages 213–222, New York, NY, USA. ACM.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. [Automatic detection of fake news](#). *CoRR*, abs/1708.07104.
- 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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2931–2937.

- Victoria Rubin, Niall Conroy, and Yimin Chen. 2015. Towards News Verification: Deception Detection Methods for News Discourse. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, Kauai, Hawaii, USA.
- Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.
- Baoxu Shi and Tim Weninger. 2016. Fact Checking in Heterogeneous Information Networks. In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*, pages 101–102, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan Facebook Pages are Publishing False and Misleading Information at an Alarming Rate. <https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis>. BuzzFeed.
- Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT press.
- Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. 2015. Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion*, pages 977–982, New York, NY, USA. ACM.
- Udo Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der Psychologie*, 11:26–181.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Oken Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 647–653.
- William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 422–426.
- Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 2013. An empirical study on uncertainty identification in social media context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 58–62, Sofia, Bulgaria. Association for Computational Linguistics.
- You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward Computational Fact-checking. *Proc. VLDB Endow.*, 7(7):589–600.
- Fan Yang, Arjun Mukherjee, and Eduard Constantin Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1979–1989.