A Topic-Agnostic Approach for Identifying Fake News Pages

Sonia Castelo
New York University
s.castelo@nyu.edu

Thais Almeida
Federal University of Amazonas
tga@icomp.ufam.edu.br

Anas Elghafari
New York University
anas.elghafari@nyu.edu

Aécio Santos
New York University
aecio.santos@nyu.edu

Kien Pham
New York University
kien.pham@nyu.edu

Eduardo Nakamura
Federal University of Amazonas
nakamura@icomp.ufam.edu.br

Juliana Freire
New York University
juliana.freire@nyu.edu

ABSTRACT

Fake news and misinformation have been increasingly used to manipulate popular opinion and influence political processes. To better understand fake news, how they are propagated, and how to counter their effect, it is necessary to first identify them. Recently, approaches have been proposed to automatically classify articles as fake based on their content. An important challenge for these approaches comes from the dynamic nature of news: as new political events are covered, topics and discourse constantly change and thus, a classifier trained using content from articles published at a given time is likely to become ineffective in the future. To address this challenge, we propose a topic-agnostic (TAG) classification strategy that uses linguistic and web-markup features to identify fake news pages. We report experimental results using multiple data sets which show that our approach attains high accuracy in the identification of fake news, even as topics evolve over time.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**; Feature selection.

KEYWORDS

Misinformation; Fake News Detection; Classification; Online News

ACM Reference Format:
Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316739

1 INTRODUCTION

Fake news have been increasingly used to manipulate public opinion and influence political processes. This has been made possible both by the existing Web infrastructure and online media platforms (e.g., Facebook, Twitter), which make it easy for information to be propagated. Unlike traditional print media, information can be published on Web sites and shared among users in social media platforms with no third party filtering or fact-checking [1]. Given that 62% of adults in the US consume news from social media [8] and many who see fake stories report that they believe them [21], these platforms have become a target for propaganda campaigns [2, 11].

While fake news have attracted substantial attention, the problem is not well understood [11]. It is challenging to discover and cross-reference conflicting sources and claims. This problem is compounded due to the large number of news sites and high volume of content. Automated methods that identify potential fake news and unreliable news sources can aid manual fact checking by providing contextual information and limiting the volume of content that the human fact-checker needs to consider. Such methods can also help us better understand the ecosystem of fake news: where they start, how they propagate, and how to counter their effects.

However, the automatic identification of fake news is a hard problem, given that news cover a wide variety of topics and linguistic styles, and can be shared on many different platforms [18].

In this paper, we study the problem of detecting fake news published on the Web. Given a web page P, our goal is to determine whether P is likely to contain fake news. Since some sites publish a mix of fake and real news [1], we consider pages published by suspicious sites as unreliable, and pages published by legitimate media outlets as reliable. We note that there is no widely-accepted definition for fake news. Here, we focus on all types of active political misinformation that go beyond old-fashioned partisan bias, and consider unreliable not only sites that publish fabricated stories, but also sites that have a pattern of misleading headlines, thinly-sourced claims, and that promote conspiracy theories. We exclude from our definition satire sites as well as opinion sites – even if extreme – if they do not display a pattern of promoting misinformation.

Previous approaches to fake news identification have largely focused on using the content of news web pages to determine their veracity [14, 17]. However, using the content has important limitations. Notably, given the dynamic nature of news, as new events are covered, topics and discourse constantly change and thus, a classifier trained using content from articles published at a given time is likely to become ineffective in the future. In addition,

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316739

studies have found that page content alone is not sufficient to accurately classify the veracity of news [4, 6, 22].

Our Approach. To address this limitation, we *propose a new classification strategy that is topic-agnostic*. Instead of using the bag of words in a page, we explore topic-agnostic features, including web-markup and linguistic features that are common in fake news.

We perform a *detailed experimental evaluation* using publicly-available datasets [1, 14] and a new dataset we created – the PoliticalNews dataset. Since existing datasets contain a small number of articles or cover a narrow time span, they are insufficient to assess the topic-agnosticism aspect of our approach. The PoliticalNews dataset contains a mix of political topics and spans several years. We report results which show that our approach is effective, obtaining accuracies that are between 8% and 24% higher than the baseline, and it is robust as topics change over time as well when applied to different domains.

Contributions. Our main contributions are: (i) we propose a classification strategy for identifying fake news which, to the best of our knowledge, is the first to rely solely on topic-agnostic features; and (ii) we present the results of an experimental evaluation, using multiple datasets and considering various baselines, which show that our approach is robust and attains high accuracy.

2 RELATED WORK

In this section, we discuss techniques that have been proposed to detect fake news published on Web sites. We also discuss publicly-available datasets that have been used to evaluate these techniques.

Detecting Fake News on the Web. Potthast et al. [16] investigated the use of writing style which included features such as n-grams and readability scores. They found that while style-based and topic-based classifiers are effective at differentiating hyper-partisan news from mainstream news (0.75 accuracy), they are not effective at differentiating fake from real news (0.55 accuracy). Horne and Adali [9] used linguistic features including readability scores, sentence structure, and part of speech of the words used. These features were very effective for the task of differentiating satire from real news (0.91 accuracy), but somewhat less so for differentiating fake news from real (0.78 accuracy). Pérez-Rosas et al. [14] also considered writing style and proposed the use of features that capture content-based aspects of web pages, such as n-grams, punctuation, psycho-linguistic, readability and syntax. Their model attains accuracy up to 0.76. Fairbanks et al. [6] investigated whether credibility and bias can be assessed using content-based and structure-based methods. The structure-based method constructs a reputation graph where each node represents a site, and the edges represent mutually linked sites, as well as shared files. This work shares some elements with our work in its usage of features derived from the HTML source of the pages (they use a subset of the feature in our classifier), but their focus is on the network between sites.

The picture that emerges from these approaches is that content-based features, while effective for detection of bias and satire, often fall short for detecting fake news. Some of works that have achieved good results on fake news detection have done so by including additional information about the sites. In our approach, we follow a similar direction and examine the combination of linguistic style with sites appearance. Because these features do not rely on the actual content of pages, our approach is topic-agnostic and robust; this is in contrast to models that use content (e.g., n-grams), which must be retrained as the topics in the news shift.

(a) Fake News: Clash Daily (b) Fake News: CDP (c) Real News: Reuters

Figure 1: Web pages from unreliable and reliable new sites.

Fake News Corpora. Despite the recent research efforts focusing on fake news, there is a dearth of publicly-available datasets focusing on *web content*. Some of the public datasets relevant to fake news detection are: BS Detector[1], BuzzFeed [15], FakeNewsNet [19], NewsMediaSources [3], US-Election2016 [1] and Celebrity [14]. These datasets are limited with respect to size, the time period they cover, and variety of topics covered. Since we aim to determine the time-invariant and topic-invariant features of political fake news on the Web, these datasets are insufficient to evaluate our work.

3 TOPIC-AGNOSTIC CLASSIFICATION

In this section, we present our approach to fake news classification.

3.1 Topic-Agnostic Features

While exploring web pages containing fake news, we observed some salient topic-agnostic features. For example, the pages contain a large number of ads – this is not surprising given that providers can gain significant advertising revenue by attracting users to their web site with appealing fake news headlines [1]. Recent works have also argued that fake news articles are designed to induce inflammatory emotions in readers, and contain text patterns related to understandability that differ from mainstream news [2, 9, 14].

Figure 1 shows some examples of fake and real news pages. Fake news pages not only have a larger number of ads and polluted layouts but also have a distinctive style to their headlines, often in the form of a sensationalist slant. Additionally, besides attempting to describe the article, these headlines often contain terms designed to catch the readers' attention (e.g., "Just in", "Read this", "Breaking News"). These observations motivated us to investigate two broad classes of features: web-markup and linguistic-based (morphological, psychological and readability-related). The features are listed in Table 1 and we summarize them below.

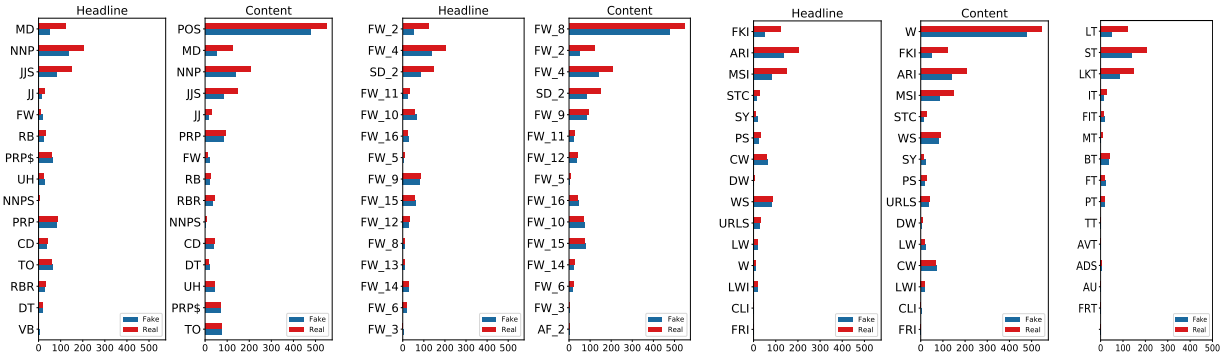
Morphological Features. This set corresponds to the frequency (word count) of morphological patterns in texts. We obtain these patterns through part-of-speech tagging, which assigns each word in a document to a category based on both its definition and context.

Psychological Features. Psychological features capture the percentage of total semantic words in texts[2]. We obtain the words'

[1] https://www.kaggle.com/mrisdal/fake-news/home
[2] http://liwc.wpengine.com/interpreting-liwc-output/

Table 1: Linguistic and web-markup features used to represent news articles.

	Abbr.	Description	Abbr.	Description	Abbr.	Description	Abbr.	Description	Abbr.	Description
Morphological Features	WDT	WH-determiner	PDT	Pre-determiner	JJ	Adjective or numeral, ordinal	VB	Verb, base form	MD	Modal auxiliary
	CD	Numeral, cardinal	VBD	Verb, past tense	VBG	Verb, present participle or gerund	VBN	Verb, past participle	RP	Particle
	DT	Determiner	NNPS	Noun, proper, plural	NN	Noun, common, singular or mass	CC	Conjunction, coordinating	WRB	Wh-adverb
	FOW	Foreign word	NNS	Noun, common, plural	TO	"To" as preposition/infinitive, superlative	WP$	WH-pronoun, possessive	JJS	Adjective, superlative
	WP	WH-pronoun	POS	Genitive marker	VBP	Verb, present tense, not 3rd singular	RBR	Adverb, comparative	NNP	Noun, proper, singular
	UH	Interjection	PRP	Pronoun, personal	VBZ	Verb, present tense, 3rd singular	RBS	Adverb superlative	PRP$	Pronoun, possessive
	RB	Adverb	JJR	Adjective, comparative	IN	Preposition or conjunction, subordinating				
Psychological Features	SO	Social (family, friend)	SD	Summary Dimensions	BP	Biological Processes (ingest, health, sexual)	AF	Affect (anger, sad, anxiety)	PC	Personal Concerns
	RL	Relativity (space, time)	FW	Function Words	TR	Time Orientation (focuspast, focuspresent)	PP	Perceptual Process	IL	Informal Language
	DR	Drives (power, risk)	PM	Punctuation Marks	OG	Other Grammar (quantifiers, interrogatives)	CP	Cognitive Processes		
Readability Features	FRI	Flesch Reading Ease	WS	Words per sentence	LW	Long words	LWI	Linsear Write	CLI	Coleman-Liau
	FKI	Flesch Kincaid Grade	CW	Capitalized words	SY	Syllables	CW	Complex words	DW	Difficult words
	MSI	McLaughlinâĂŹs SMOG	LX	Lexicon	PS	Percentage of stop words	ARI	Automated Readability	W	Words
	GFI	Gunning Fog	URL	URLs	STC	Sentences	CH	Characters		
Web-markup features	AU	Author	IT	Images (e.g., img, canvas)	ST	Semantics (e.g., article, section)	FRT	Frames (e.g., frame, frameset)	LT	Lists (e.g., ul, ol, li)
	BT	Basic (e.g., title,h1, p)	FT	Formatting (e.g., acronym)	FIT	Forms and inputs (e.g., textarea, button)	MT	Metainfo (e.g., head, meta)	TT	Tables (e.g., tbody, tfoot)
	AVT	Audio and video	LKT	Links (e.g., a, nav, link)	PT	Programming (e.g., script, object)	ADS	Advertisements		



(a) Morphological (b) Psychological (c) Readability (d) Web-markup

Figure 2: Mean frequency distribution of features per class in the PoliticalNews dataset.

semantics by using a dictionary that has lists of words that express psychological processes (personal concerns, affection, perception).

Readability Features. This set captures the ease or difficulty of comprehending a text. We obtain these features through readability scores and counting of character, words, and sentences usage.

Previous works have found that fake news often displays a divergence between the news headline and the body content [9, 20]: (i) a headline declares a piece of information to be false and the body text declares it to be true (or vice-versa); and (ii) fake news packs the main claim of the article into its title, allowing the reader to skip reading the body article, which tends to be short, repetitive, and less informative when compared with real news. These divergences between the textual pieces of news articles motivated us to apply linguistic features at different granularities: considering only the headline, only the content, and combining both. Figures 2(a), 2(b), and 2(c) show the ranking of the top-15 linguistic-based features, in PoliticalNews dataset, with the largest differences between classes considering distinct text granularities. In Figure 2, the psychological features are named alongside an index. This is because each of the psychological features represents a category that contains other features. A complete list of these features is available at http://liwc.wpengine.com/compare-dictionaries.

Web-Markup Features. These features capture patterns of the web pages' layout . The web-markup features we use include: frequency (number of occurrences) of advertisements, presence of an author name (binary value), and the frequency of distinct categories of tag groups[3]. Figure 2(d) shows the mean distribution frequency of each web-markup features in PoliticalNews dataset, within fake and real news. Note that the distributions are different, thus supporting the use of these features to distinguish news.

3.2 Feature Selection

We perform feature selection using a combination of four different methods: Shannon Entropy [12], Tree Based Rule [7], L1 Regularization [23] and Mutual Information [10]. We combine the outputs from these methods by normalizing them and applying the geometric mean to obtain a new score $r(f_i)$ which corresponds to the importance of a feature f_i:

$$r(f_i) = \sqrt[4]{SE(f_i)^{-1} \times TB(f_i) \times L1(f_i) \times MI(f_i)} \quad (1)$$

where SE refers to Shannon Entropy, TB to Tree Based Rule, L1 to L1 Regularization and MI to Mutual Information. We compute the feature importance scores and remove features with $r(f_i)$ value equal to 0. When we performed this process for features from news headlines, we found the following patterns turned out to be ineffective: FOW, IN, JJR, PRP$, TO, VBD, VBG, VBZ, WP$, MSI, CW, TT, FW (semicolon), BP (ingest), RL (time) and PC (home). When we consider features from news content, DT, PDT, RBR, RP, OG, and UH are removed. When we extract features from both

[3]https://www.w3schools.com/tags/ref_byfunc.asp

(headline + content), DT, JJS, PDT, POS, RBR, RBS, UH and WRB are eliminated. For the web-markup features, we kept IT, AVT, AU, LKT, ADS, ST and BT. The sizes of the sets of topic-agnostic features are: (1) for headlines, 137 features; (2) for content, 148; and (3) headline+content, 145. We study the effectiveness of these features in Section 4. Note that we performed the feature selection process only once, using the PoliticalNews dataset, and used the resulting features on all experiments.

3.3 Classification

We use supervised learning to classify news based on the selected features. Formally, the resulting classifier corresponds to a function that takes as input the TAG features of a web page $x \in \mathbb{R}^d$ and produces an output $\hat{y} \in C$, where C is the set of all categories. Here we use two categories: fake and real news. We experimented with three different learning methods: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest (RF). We use SVM classifier with linear kernel and cost equal to 0.1. The other methods are setup with default parameters of the scikit-learn[4] library.

4 EXPERIMENTAL EVALUATION

In this section, we describe the evaluation that we applied to assess how effective our approach is at identifying fake news.

Corpus. As discussed in Section 2, existing public datasets are limited. So, we built a new dataset that we call ***PoliticalNews***[5], by compiling a list of known reliable and unreliable sites and crawling them. As seeds for the crawl, we used web sites coming from Politifact[6], BuzzFeed [15], and OpenSources.co[7] as unreliable news sites and 242 sites from the most visited news sites *Alexa's top 500 news sites*[8] as reliable sites. Since the last one is based just on popularity, we manually selected the web sites focused on political news. One challenge in constructing such a large dataset is the impracticality of individually labeling all the articles. Our approach was to project the domain-level ground truth onto the content collected from those domains. We collected over 1.6 million web pages published between 2013 and 2018. Then, we post-processed the data removing non-political pages using a Naïve Bayes model combined with TF-IDF feature representation, and as a training set, we use a publicly-available corpus[9]. Furthermore, to ensure a balanced distribution of web pages sites for each year, we sampled 32 pages from each site for each of the years we have crawl data for. The result of this balancing is a dataset of 14,240 news pages with 7,136 pages coming from 79 unreliable sites, and another 7,104 coming from 58 reliable sites. The motivation for this balancing is to avoid the problem of overfitting (since real news is orders of magnitude more prevalent than fake news). Additionally, ensuring balance across the years covered allows us to evaluate the effectiveness of different approaches as news topics change over time.

To demonstrate the robustness of the TAG model, we also evaluate our approach over two additional datasets that have been used in previous works: ***Celebrity*** [14] and ***US-Election2016*** [1]. We note that these datasets contain only the text of the web pages. Since we need to extract web-markup features, we fetched the original web page source from the Web. For sites and pages that were no longer available, we retrieved versions from the Web Archive. In addition, duplicated articles were removed. As a result, the versions of the Celebrity and US-Election2016 datasets[10] used in our experiments contain 479 and 691 news articles, respectively.

Experimental Setup. We use the NLTK [5] library part-of-speech tagger to compute *Morphological Features*. To extract the *Psychological Features*, we use the Linguistic Inquiry and Word Count software (LIWC, Version 1.3.1 2015) [13]. To compute *Readability Features*, we use a Textstat[11] library, and finally, to extract the *Web-Markup Features* from web pages, we use BeautifulSoup[12] and Newspaper.[13]

We compare the TAG classifier with the Fake News Detector (FNDetector) presented in [14]. This model represents documents by using four sets of linguistic features: n-grams (unigrams + bigrams encoded by TF-IDF values), psychological, readability and syntactical features (production rules of context-free grammars encoded by TF-IDF values). The psychological and readability sets are the same as we use (see Table 1). We selected the FNDetector for two main reasons: like our approach, they focus on the automatic identification of fake content in online news; and because their classifier attained high accuracy using content-based features, it is a suitable baseline for our TAG classifier. To better understand the contribution of individual features, we created multiple baseline classifiers using different feature combinations. We used a linear SVM classifier and conducted our evaluations using five-fold cross-validation with accuracy as the performance metric.

4.1 Effectiveness of Different Features

We evaluate the performance of models trained with different combinations of feature sets (separately and jointly). In addition, to assess the performance of the classifiers using different representations for a news article, each experiment was performed for the headline (H), content (C) and the combination of both (HC).

Table 2 shows the accuracy obtained for the different TAG feature sets over the three datasets. Note that combining features often leads to the highest accuracies. For example, for political news, the highest accuracies are obtained by the configuration that combines **LIWC (L) - NLTK (N) - readability (R) - webmarkup (W)**, which attains 0.86 accuracy for US-Election2016 and 0.83 for PoliticalNews. For the Celebrity data, the best configuration is **LIWC (L) - readability (R) - webmarkup (W)** which attains 0.78 accuracy. If we further examine the results from LIWC, we can see that combining LIWC with other features often leads to higher accuracies: for all datasets, the accuracy gains vary between 0.012 and 0.21. This reinforces previous findings [2, 14] that psychological factors play an important role in the disinformation ecosystem. The web-markup features also lead to improvements, in particular when compared against readability for the PoliticalNews corpus.

[4] http://scikit-learn.org/
[5] https://osf.io/ez5q4/
[6] https://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/
[7] http://www.opensources.co
[8] https://www.alexa.com/topsites/category/News
[9] https://www.kaggle.com/rmisra/news-category-dataset/home
[10] https://osf.io/qj86g/
[11] http://pypi.python.org/pypi/textstat/
[12] https://www.crummy.com/software/BeautifulSoup
[13] https://newspaper.readthedocs.io/en/latest

Table 2: Accuracy results for models that use different set of topic-agnostic features – where L is LIWC, N is NLTK, R is readability, and W is webmarkup features – over three different datasets: Celebrity, US-Election2016, and PoliticalNews. The best accuracies for each feature set are bold; the best accuracies for each news article's representations (H, C and HC) are underlined.

Dataset	Celebrity			US-Election2016			PoliticalNews		
Features	H	C	HC	H	C	HC	H	C	HC
W	**0.68**	**0.68**	**0.68**	**0.65**	**0.65**	**0.65**	**0.71**	**0.71**	**0.71**
L	0.69	**0.73**	**0.73**	0.77	0.81	**0.83**	0.71	0.75	**0.76**
N	0.58	**0.68**	0.66	**0.81**	0.75	0.76	**0.77**	0.66	0.67
R	0.57	**0.62**	0.57	**0.75**	0.73	0.73	**0.69**	0.62	0.64
L-R	0.65	**0.76**	0.74	0.79	0.82	**0.83**	0.74	0.75	**0.76**
N-R	0.65	**0.68**	0.67	**0.83**	0.78	0.78	**0.78**	0.71	0.72
N-W	0.68	**0.72**	**0.72**	0.79	0.79	**0.80**	**0.81**	0.79	0.79
L-W	0.70	**0.77**	0.75	0.79	0.82	**0.83**	0.78	**0.81**	**0.81**
R-W	0.67	**0.72**	0.67	**0.80**	0.76	0.76	**0.77**	0.76	0.76
N-R-W	0.67	**0.72**	0.71	**0.83**	0.79	0.80	**0.81**	0.78	0.79
L-R-W	0.71	**0.78**	0.71	0.79	0.83	**0.85**	0.80	0.80	**0.81**
L-N-R-W	**0.73**	**0.73**	0.71	0.83	0.82	**0.86**	**0.83**	0.81	0.82

Table 3: Classification results (accuracies) for three datasets.

Dataset	Celebrity	US-Election2016	PoliticalNews
FNDetector	0.73	0.81	0.76
TAG Model	**0.78**	**0.86**	**0.83**

When we consider the features from different article parts (H, C, and HC), the distribution of accuracies for US-Election2016 and PoliticalNews are similar. This suggests that the headlines contain useful information that allows the identification of political fake news. Thus, either headlines or content or both can be used for classification. On the other hand, for the Celebrity dataset, we can clearly see that the classifier achieves slightly better results using the content of the articles. Consider the following two examples from this dataset: (1) a real news, where the headline is "*Stop Right Now! The Spice Girls Might Be Planning A Reunion*" and the content is "*There are certain things that people just shouldn't joke about. A potential Spice Girls reunion with all five members is one of them. Earlier this morning...*", and (2) a fake news, where the headline is "*Taylor Swift Goes Naked in '...Ready for It?' Watch!*" and the content is "*Nope, definitely not ready for this! Taylor Swift released a 15-second teaser for the music video for her new song '... Ready for It?' on Monday...*". The previous example in the Celebrity dataset shows a concrete example where headlines (H) for fake and real news have similar linguistic features (e.g., use of capitalized words, punctuation), but content (C) is more discriminative in terms of linguistic features and readability. Furthermore, the web-markup features are effective in this scenario, since they make it possible to detect a clear difference between fake and real news based on web page characteristics (e.g., number of ads, links, and images).

We also compare the best TAG feature set combination for each dataset, identified during our previous experiment, with the baseline. We use a linear SVM and five-fold cross-validation with accuracy as metric. As shown in Table 3, the classifier based on TAG

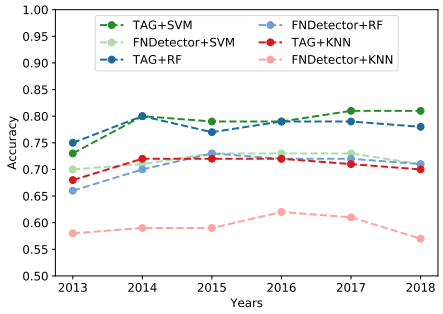
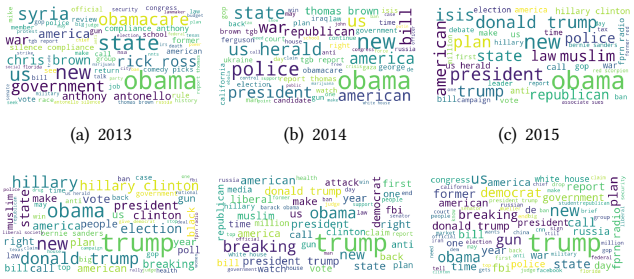


Figure 3: Effectiveness of SVM, KNN and RF using baseline (FNDetector) and TAG model in different time windows.



(a) 2013 (b) 2014 (c) 2015

(d) 2016 (e) 2017 (f) 2018

Figure 4: Tag clouds of the most frequent terms in web pages.

features outperforms the *FNDetector* for all datasets, indicating that the task of fake news classification can be effectively accomplished using topic-agnostic features. Furthermore, it is important to note that *FNDetector* features uses 3 orders of magnitude more features (~**800,000**) than our approach (~**160** features). The number of features has important implications for the processing time, maintaining the model, and explainability.

4.2 Effectiveness over Time

To study the behavior of our approach as news topics change over time, we split PoliticalNews dataset into six time windows (sub-datasets) corresponding to each year from 2013 to 2018. We then experimented with multiple classifier configurations: each configuration uses one sub-dataset for training and the others (one at a time) for testing. For example, we use pages published in 2013 to construct a classifier C_{2013} and use C_{2013} to classify pages in the sub-datasets from 2014 through 2018. Note that each sub-dataset is associated with 5 results. For this experiment, our TAG model uses the set of features NLTK, LIWC, readability and web-markup.

Figure 3 shows the average accuracies for each sub-dataset over the 5 test sets. Our TAG model performs better than *FNDetector* for all the time windows. The relatively lower accuracy values obtained by *FNDetector* can be explained by its dependence on the contents. You can observe the topic changes in Figure 4, which shows tag clouds (built over our training data using n-grams frequencies) summarizing the news in each year. Note that even though the tag clouds share some keywords (e.g., "Obama" and "Trump"), they appear at different frequencies and cover distinct subjects.

The content-based model essentially learns how to detect fake news for a specific time and topics. Note, for example, the difference between the topics in 2013 and in 2015. In 2013, "Trump" was

Table 4: Cross-domain results (accuracies) between models.

Training	Test	Classifier	Accuracy	
			FNDetector	TAG Model
Celebrity	US-Election 2016	SVM	0.59	**0.70**
		KNN	0.59	**0.64**
		RF	0.56	**0.64**
US-Election 2016	Celebrity	SVM	0.59	**0.63**
		KNN	0.56	**0.60**
		RF	0.51	**0.60**

not yet involved in politics (see Figure 4(a)), but he starts to appear in political news in 2015, and more frequently since the 2016 elections as shown in Figure 4(e) and Figure 4(f). Also note the differences between 2016 and 2017: In 2016, political news were centered around the campaign, parties, voters, etc., and we see terms such as "Hillary", "Clinton", "vote", "Trump" and "Obama". But in 2017 (after the elections), the terms "Hillary", "Clinton" decrease and terms like "breaking", "attack", "Russia", "FBI" start to appear in the political discourse. This indicates that, to be effective, content-based models must be constantly retrained. This is both costly and error-prone.

4.3 Effectiveness for Different Domains

We also evaluated the generalizability of our approach when the training and testing sets are drawn from entirely different topic domains. We ran two experiments: in the first, we use Celebrity as a training dataset and US-Election2016 as testing dataset; and for the second experiment, the other way around. We use all the TAG feature sets identified during our previous experiments (N, L, R and W). We considered three classifiers – SVM, KNN and RF; and used five-fold cross-validation. The results in Table 4 show that our topic-agnostic approach attains accuracies that are substantially higher than those of FNDetector. While our original goal in this work was to design a classifier that is able to distinguish fake and real political news as they evolve over time, these results show that the approach is promising for topic domains beyond political news.

5 CONCLUSIONS & FUTURE WORK

We presented a new approach to detect fake news web pages which uses topic-agnostic features. Through a detailed experimental evaluation, we showed that our approach accurately classifies not only political news as topics evolves over time but also news from different domains, outperforming content-based approaches while using significantly fewer features and requiring no frequent re-training. Our results suggest that topic-agnostic features are effective for distinguishing between fake and real news, and that robust classifiers can be constructed that enable the timely discovery of fake news articles. We have also created a new corpus of over 14,000 political news pages drawn from 137 sites and spanning 6 years. To the best of our knowledge, this is the first of its kind in terms of size, focus on the Web, and inclusion of HTML markup information.

There are many directions we plan to pursue in future work. We will explore further improvements to the topic-agnostic model by considering additional features, for example, user engagement and network structure. We would also like to experiment with different strategies to expand our fake news corpus, including the use of social media to search for previously unknown sites, and by using the top-agnostic classifier in conjunction with a focused crawler to discover new sites as they are created.

ACKNOWLEDGMENTS

This work was partially supported by the DARPA MEMEX and D3M programs, NSF award CNS-1229185, and the Brazilian National Council for the Improvement of Higher Education (CAPES) under grant 88887.130294/2017-00. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31 (2017), 211–36.
[2] Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism* 6 (2018), 154–175.
[3] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proc. of the Conf. on Empirical Methods in NLP*. 3528–3539.
[4] Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A System to Unmask Propaganda in Online News. In *Proc. of the 33th AAAI Conf. on Artificial Intelligence*.
[5] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proc. of the Association for Computational Linguistics on Interactive poster and demonstration sessions*. 31.
[6] James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility Assessment in the News: Do we need to read?. In *Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining*. 799–800.
[7] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63 (2006), 3–42.
[8] Jeffrey Gottfried and Elisa Shearer. 2016. News Use across Social Media Platforms 2016. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016.
[9] 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. In *Proc. of the 2nd Intn'l Workshop on News and Public Opinion*.
[10] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69 (2004), 066138.
[11] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* (2018).
[12] Annick Lesne. 2014. Shannon Entropy: A Rigorous Notion at The Crossroads Between Probability, Information Theory, Dynamical Systems and Statistical Physics. *Mathematical Structures in Computer Science* 24 (2014), 240–311.
[13] James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. 10.15781/T29G6Z.
[14] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proc. of Int'l Conf. on Computational Linguistics*. 3391–3401.
[15] Scott Pham. [n. d.]. Analysis of fake news sites and viral posts, 2016 vs. 2017. https://github.com/BuzzFeedNews/2017-12-fake-news-top-50.
[16] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A Stylometric Inquiry into Hyperpartisan and Fake News. *CoRR* abs/1702.05638 (2017).
[17] 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 *Proc. of the Conf. on Empirical Methods in NLP*. 2931–2937.
[18] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19 (2017), 22–36.
[19] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting Tri-Relationship for Fake News Detection. *CoRR* abs1712.07709 (2017).
[20] Craig Silverman. 2015. Lies, damn lies, and viral content: How news websites spread (and Debunk) online rumors, unverified claims and misinformation. *Tow Center for Digital Journalism* 168 (2015).
[21] Craig Silverman and Jeremy Singer-Vine. 2016. Most Americans Who See Fake News Believe It, New Survey Says. BuzzFeed News.
[22] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proc. of the 11th Int'l Conf. on Web Search and Data Mining*. 637–645.
[23] Peng Zhao and Bin Yu. 2006. On model selection consistency of Lasso. *Journal of Machine learning research* 7 (2006), 2541–2563.