Automatic Detection of Fake News

Article · August 2017

CITATIONS
6

READS
1,224

4 authors:

Verónica Pérez-Rosas
University of Michigan
30 PUBLICATIONS **261** CITATIONS
SEE PROFILE

Bennett Kleinberg
University of Amsterdam
20 PUBLICATIONS **1,158** CITATIONS
SEE PROFILE

Alexandra Lefevre
University of Michigan
1 PUBLICATION **6** CITATIONS
SEE PROFILE

Rada Mihalcea
University of Michigan
255 PUBLICATIONS **12,027** CITATIONS
SEE PROFILE

Automatic Detection of Fake News

Verónica Pérez-Rosas[1], **Bennett Kleinberg**[2], **Alexandra Lefevre**[1]
Rada Mihalcea[1]
[1]Computer Science and Engineering, University of Michigan
[2]Department of Psychology, University of Amsterdam
vrncapr@umich.edu,b.a.r.kleinberg@uva.nl,mihalcea@umich.edu

Abstract

The proliferation of misleading information in everyday access media outlets such as social media feeds, news blogs, and online newspapers have made it challenging to identify trustworthy news sources, thus increasing the need for computational tools able to provide insights into the reliability of online content. In this paper, we focus on the automatic identification of fake content in online news. Our contribution is twofold. First, we introduce two novel datasets for the task of fake news detection, covering seven different news domains. We describe the collection, annotation, and validation process in detail and present several exploratory analysis on the identification of linguistic differences in fake and legitimate news content. Second, we conduct a set of learning experiments to build accurate fake news detectors. In addition, we provide comparative analyses of the automatic and manual identification of fake news.

1 Introduction

Fake news detection has recently attracted a growing interest from the general public and researchers as the circulation of missinformation online increases, particularly in media outlets such as social media feeds, news blogs, and online newspapers. For instance, a recent report by the Jumpshot Tech Blog[1] found that Facebook referrals accounted for 50% of the total traffic to fake news sites and 20% total traffic to reputable websites. Since the majority of U.S. adults –62%– gets news on social media (Jeffrey and Elisa, 2016), being able to identify fake content in online sources is a pressing need.

To date, computational approaches for fake news detection have relied on satirical news sources such as "The Onion" and fact-checking websites such as "politiFact" and "Snopes". However, the use of these sources poses several challenges and potential drawbacks. For instance, using satirical content as a source for fake content can bring underlying confounding factors into the analysis, such as humor and absurdity. This is particularly the case for satirical news from "The Onion", which has been used in the past to explore other text properties such as humor (Mihalcea and Strapparava, 2005) and irony (Wallace, 2015). On the other hand, fact-checking websites are usually constrained to a particular domain of interest, such as politics, and require human expertise, thus making it difficult to obtain datasets that provide some degree of generalization over several domains.

In this paper, we develop computational resources and models for the task of fake news detection. We present the construction of two novel datasets covering seven different domains. One of the datasets is collected using a combination of manual and crowdsourced annotation efforts, while the second is collected directly from the web. Using these datasets, we conduct several exploratory analyses to identify linguistic properties that are predominantly present in fake content, and we build fake news detectors relying on linguistic features that achieve accuracies of up to 78%. To place our results in perspective, we also compare the accuracy of our fake news detection models with an empirical human baseline accuracy.

2 Related Work

To date, there are three important lines of research into the automated classification of genuine and

[1]https://www.jumpshot.com/data-facebooks-fake-news-problem/

fake news items. First, on a conceptual level, a distinction has been made between 'three types of fake news' (Rubin et al., 2015): serious fabrications (i.e. news items about false and non-existing events or information such as celebrity gossip), hoaxes (i.e. providing false information via, for example, social media with the intention to be picked up by traditional news websites) and satire (i.e. humorous news items that mimic genuine news but contain irony and absurdity). Here, we focus on the first category, serious fabrication, in the two domains of general news (in six different categories), as well as on celebrity gossip.

Second, attempts to differentiate satire from real news yielded promising results (Rubin et al., 2016). The authors built a corpus of satire news (from *The Onion* and *The Beaverton*) and real news (*The Toronto Star* and *The New York Times*) in four domains (civics, science, business, soft news), resulting in a total of 240 news articles. The best classification performances were achieved with feature sets representing absurdity, punctuation, and grammar (each with an F1 score of 0.87).

Third, recently, a stylometric (i.e. writing-style) approach has been proposed for the identification of fake and genuine news articles (Potthast et al., 2017). The investigation used the Buzzfeed dataset[2] of mainstream and hyperpartisan news articles of which the veracity was manually annotated. Stylometric features were, among others, character and stop word n-grams, readability indices, as well as features such as external links and the average number of words per paragraph. As a comparison, a topic-based feature set of a non-domain specific bag-of-words approach was used. The dataset used by (Potthast et al., 2017) consisted of 1,627 news articles that were obtainable from the original Buzzfeed dataset, including 299 fake news articles. Although the stylometric approach was promising for the classification of hyperpartisan versus mainstream articles (accuracy: 0.75, compared to 0.71 for the topic-based feature set), both approaches were not able to differentiate fake from real news (accuracy: 0.55 and 0.52 for stylometric and topic-based feature sets, respectively).

Also related to our research is work done on the automatic identification of deceptive content, which has explored domains such as forums, consumer reviews websites, online advertising, online dating, and crowdfounding platforms (Warkentin et al., 2010; Ott et al., 2011a; Zhang and Guan, 2008; Toma and Hancock, 2010; Shafqat et al., 2016). Linguistic clues such as self references or positive and negative words have been used to profile true tellers from liars (Newman et al., 2003). Other work has focused on analyzing the number of words, sentences, self references, affect, spatial and temporal information associated with deceptive content (Qin et al., 2005). Expressivity, informality, diversity and non-immediacy have also been explored to identify deceitful behaviors (Shafqat et al., 2016).

3 Fake News Datasets

As highlighted earlier, the datasets used in previous work have either relied on satirical news (e.g., "The Onion"), which also have confounds such as humor or irony; or used fact-checking websites (e.g., "politiFact" or "Snopes"), which are typically focused on only one domain (generally politics). We thus decided to construct two new datasets of fake news that cover several news domains and specifically model the deceptive property of fake news without major confounds. One dataset is collected via crowdsourcing, and covers six news domains; the second dataset is obtained directly from the web, and covers celebrity fake news.

Guidelines for a Fake News Corpus. In building a fake news dataset, we adhered to the nine requirements of a fake news corpus proposed by (Rubin et al., 2016). Specifically, the authors suggested that such a corpus should (1) include both fake and real news items, (2) contain text-only news items, (3) have a verifiable ground-truth, (4) be homogeneous in length and (5) writing style, (6) contain news from a predefined time frame, (7) be delivered in the same manner and for the same purpose (e.g. humor, breaking news) for fake and real cases, (8) be made publicly available, and (9) should take language and cultural differences into account. In our work, to the extent possible, we aimed to address all of the above guidelines.[3] As outlined in the following, the ground-truth remains challenging since we cannot verify with absolute certainty whether all the content of real news items is in fact true.

[2] http://github.com/BuzzFeedNews/2016-10-facebook-fact-check

[3] We did not explicitly account for cultural differences since the primary aim was to build a fake news dataset that met criterion 1 to 8.

3.1 Building a Crowdsourced Dataset

Collecting Legitimate News. We started by collecting a dataset of legitimate news belonging to six different domains (sports, business, entertainment, politics, technology, and education). The news were obtained from a variety of mainstream news websites (predominantly in the US) such as the ABCNews, CNN, USAToday, NewYorkTimes, FoxNews, Bloomberg, and CNET among others.

To ensure the veracity of the news, we conducted manual fact-checking on the news content, which included verifying the news source and cross-referencing information among several sources. Using this approach, we collected 40 news in each of the six domains, for a total of 240 legitimate news.

Collecting Fake News using Crowdsourcing. To generate fake versions of the news in the legitimate news dataset, we make use of crowdsourcing via Amazon Mechanical Turk, which has been successfully used in the past for collecting deception data on several domains, including opinion reviews (Ott et al., 2011b), and controversial topics such as abortion and death penalty (Pérez-Rosas and Mihalcea, 2015).

However, collecting deceptive data via AMT poses additional challenges on the news domain. First, the reporting language used by journalists might differ from AMT workers language (e.g., journalistic vs. informal style). Second, journalistic articles are usually lengthier than consumer reviews and opinions, thus increasing the difficulty of the task for AMT workers as they would be required to read a full news article and create a fake version from it.

To address the former, we asked the workers to the extent possible to emulate a journalistic style in their writing. This decision was motivated by the 5th point of the fake news corpus guidelines described in section 3, which suggests to obtain news with homogeneous writing style. To address the latter, we opted to working with smaller information units. Our approach consists of manually selecting a news excerpt that briefly describes the news article.[4] Thus, from the legitimate news dataset collected earlier, we manually extracted 240 news excerpts. The final dataset consists of 33,378 words. Each news excerpt has on average 139 words and approximately 5 sentences.

[4]In many cases, this corresponded to the first 2-3 paragraphs in the document.

We set up an AMT task that asked workers to generate a fake version of the provided news. Each hit included the legitimate news headline and its corresponding body. We instructed workers to produce both a fake headline and a fake news body within the same topic and length as the original news. Workers were also requested to avoid unrealistic content and to keep the names mentioned in the news. The fake news were produced by unique authors, as we allowed only a single submission per worker. We restricted the submission to workers located in the US as they might be more familiar with news published in the US media. In addition, we restricted participation to workers who maintained an approval rate of at least 95% to reduce potential spam contributions.

It took approximately five days to collect 240 fake news. Each hit was manually checked for spam and to make sure workers followed the provided guidelines. In general, we received few spam responses and most of the workers followed instructions satisfactorily; the only exceptions were a few cases where they provided only the headline or included unrealistic content.

Interestingly, we observed that AMT workers succeeded in mimicking the reporting style from the original news, which may be partly explained by typical verbal mirroring behaviors with drive individuals to produce utterances that match the grammatical structure of sentences they have recently read (Ireland and Pennebaker, 2010). This partially addresses our initial concern of authors reporting style being a source of noise while analyzing news generated by journalists and AMT workers.

The final set of fake news consists of 31,990 words. Each fake news has on average 132 words and approximately 5 sentences. Table 1 shows a sample fake news, along with its legitimate version, in the technology domain.

Throughout the rest of the paper, we refer to this crowdsourced dataset as FakeNewsAMT.

3.2 Building a Web Dataset

We collected a second dataset of fake news from web sources following similar guidelines as in the previous dataset. However, this time, we aimed to identify fake content that naturally occurs on the web. We opted for collecting news from public figures as they are frequently targeted by rumors, hoaxes, and fake reports. We focused mainly on celebrities (actors, singers, socialites,

LEGITIMATE	FAKE
Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display. Nintendo officials demonstrated features such as using the detachable remote controllers, called "Joy-Con," to play a gun-duel game. Motion sensors enable players to feel virtual water being poured into a virtual cup.	**New Nintendo Switch game console to launch in March for $99** Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming. However, initial quantities will be limited to 250,000 units available at the sales price. So rush out and get yours today while the promotional offer is running.

Table 1: Sample legitimate and crowdsourced fake news in the Technology domain

LEGITIMATE	FAKE
Kim And Kanye Silence Divorce Rumors With Family Photo. Kanye took to Twitter on Tuesday to share a photo of his family, simply writing, "Happy Holidays." In the picture, seemingly taken at Kris Jenner's annual Christmas Eve party, Kim and a newly blond Kanye pose with their children, North, 3, and Saint, 1. After Kanyes hospitalization, reports that there was trouble in paradise with Kim started brewing. But E! News shut down the speculation with a family source denying the rumors and telling the site, "It's been a very hard couple of months." Kim remains out of the spotlight while Kanye is reportedly seeking outpatient treatment. Though Kim has yet to make a real return to social media herself, she's been spotted on Kanyes page, as well as Khloe Kardashian's and Kylie Jenner's Instagrams and Snapchats. Kim and Ye were also photographed on a dinner date last week for the first time in a while, so things are looking up.	**Kim Kardashian Reportedly Cheating With Marquette King as She Gears up for Divorce From Kanye West.** Kim Kardashian is ready to file for divorce from Kanye West but has she REALLY been cheating on him with Oakland Raiders punter Marquette King? The NFL star seemingly took to Twitter to address rumors that they've been getting close amid Kanye's mental breakdown, which were originally started by sports blogger Terez Owens. While he doesn't appear to confirm or deny an affair, her reps said there is "no truth whatsoever" to the reports and labeled the situation "fabricated." As In Touch previously reported, Kim has been speaking with famed divorce attorney Laura Wasser and asked for documents to be drawn up. It has yet to be confirmed if Laura, who is also a friend of the reality star, will represent Kim during the proceedings. An insider blames the rapper's paranoia as a reason for the demise of their marriage. "Kim is miserable and wants this marriage to be over," says the source.

Table 2: Sample legitimate and web fake news in the Celebrity domain

and politicians) and our sources include online magazines such as Entertainment Weekly, People Magazine, RadarOnline, among other tabloid and entertainment-oriented publications. The data were collected in pairs, with one article being legitimate and the other fake. In order to determine if a given celebrity news was legitimate or not, the claims made in the article were evaluated using gossip-checking sites such as "GossipCop.com", and were cross-referenced with information from other sources.

During the initial stages of the data collection, we noticed that celebrity news tend to center on sensational topics that sources believe readers want to read about, such as divorces, pregnancies, and fights. Consequently, celebrity news tends to follow certain celebrities more than others further leading to an inherent lack in topic diversity in celebrity news. To address this issue, we evaluated several sources to make sure we obtain a diversified pool of celebrities and topics. Upon beginning the data collection procedure using these guidelines, another characteristic surfaced: several pairs contained nearly the same information with similar lexicon and reporting style, with differences being as simple as just negating the false news. For example, the following headlines correspond to a news pair where the legitimate version only negates the fake version: "Aniston gets into fight with husband" (fake) and "Aniston did NOT get into fight with husband" (legitimate). To address this issue, we sought to identify related news that still followed the fake-legitimate pair property while being sufficiently diverse in lexicon and tone. In the former example, the fake news was paired with an article titled "Aniston and Husband enjoy dinner" that was published on the date of the alleged fight.

Using this approach, we collected 100 fake news articles and 100 legitimate news articles in the celebrity domain. The final fake news set has an average of 399 words and 17 sentences per article, for a total of 39,940 words. The corresponding legitimate news set has an average of 709 words

and 33 sentences per article, for a total of 70,975 words. 2 shows an example of an article pairing in the dataset.

Throughout the rest of the paper, we refer to this web dataset as Celebrity.

4 Linguistic Features

To build the fake news detection models, we start by extracting several sets of linguistic features:

Ngrams. We extract unigrams and bigrams derived from the bag of words representation of each news article. To account for occasional differences in content length, these features are encoded as tf-idf values.

Punctuation. Previous work on fake news detection (Rubin et al., 2016) as well as on opinion spam (Ott et al., 2011b) suggests that the use of punctuation might be useful to differentiate deceptive from truthful texts. We construct a punctuation feature set consisting of eleven types of punctuation derived from the Linguistic Inquiry and Word Count software (LIWC, Version 1.3.1 2015) (Pennebaker et al., 2015). This includes punctuation characters such as periods, commas, dashes, question marks and exclamation marks.

Psycholinguistic features. We use the LIWC lexicon to extract the proportions of words that fall into psycholinguistic categories. LIWC is based on large lexicons of word categories that represent psycholinguistic processes (e.g., positive emotions, perceptual processes), summary categories (e.g., words per sentence), as well as part-of-speech categories (e.g., articles, verbs). Previous work on verbal deception detection showed that LIWC is a valuable tool for the deception detection in various contexts (e.g., genuine and fake hotel reviews, (Ott et al., 2011b, 2013); prisoners' lies (Bond and Lee, 2005)). In our work, we cluster the single LIWC categories into the following feature sets: summary categories (e.g., analytical thinking, emotional tone), linguistic processes (e.g., function words, pronouns), and psychological processes (e.g., affective processes, social processes).

We also test a combined feature set of all the LIWC categories (including punctuation).[5]

Readability. We also extract features that indicate text understandability. These include content features such as the number of characters, complex words, long words, number of syllables, word types, and number of paragraphs, among others content features. We also calculate several readability metrics, including the Flesch-Kincaid, Flesch Reading Ease, Gunning Fog, and the Automatic Readability Index (ARI).

Syntax. Finally, we extract a set of features derived production rules based on context free grammars (CFG) trees using the Stanford Parser (Klein and Manning, 2003). The CFG derived features consist of all the lexicalized production rules (rules including child nodes) combined with their parent and grandparent node, e.g., *NNˆNP→commission (in this example NN –a noun– is the grandparent node, NP –personal pronoun– the parent node, and "commissions" the child node. Features in this set are also encoded as tf-idf values.

5 Computational Models for Fake News Detection

We conduct several experiments with different (combinations of) feature sets. We use a linear SVM classifier and five-fold cross-validation, with accuracy, precision, recall, and F1 measures averaged over the five iterations.

The machine learning classification was conducted with R (R Core Team, 2016) and the caret (Kuhn et al., 2016) and e1071 packages (Meyer et al., 2015).

Tables 3 and 4 show the results obtained for the different feature sets. As seen in the tables, most of the classifiers obtain performances well above the random baseline of 0.50. The best performing classifier for the FakeNewsAMT dataset is derived from the *Readability* features, followed by the combination of all linguistic feature sets. For the Celebrity dataset, the most accurate model is built using the *Punctuation* features, followed by the *Ngrams*, *Complete LIWC*, and *Syntax* features.

Learning Curves. Next, we investigate whether larger amounts of training data can improve the identification of fake content. We plot the learning curves of the bests sets of features using incremental amounts of data as shown in Figures 1 and 2. Except for the decrease obtained with the *Readability* features on the Celebrity dataset, the learning trend for all the other feature sets on both datasets show steady improvement, thus suggesting that larger quantities of training data could im-

[5]The feature sets linguistic processes and punctuation correspond to the 'grammar' and punctuation feature set, respectively, in (Rubin et al., 2016)

Features (number of features)	Acc.	LEGITIMATE			FAKE		
		P	R	F1	P	R	F1
Punctuation (11)	0.71	0.73	0.66	0.69	0.69	0.76	0.72
LIWC - Summary (7)	0.61	0.63	0.54	0.58	0.60	0.68	0.64
LIWC - Linguistic processes (21)	0.67	0.66	0.67	0.66	0.67	0.66	0.66
LIWC - Psychological processes (40)	0.56	0.56	0.57	0.56	0.56	0.56	0.55
Complete LIWC (79)	0.70	0.70	0.71	0.70	0.71	0.70	0.70
Readability (26)	0.78	0.82	0.72	0.77	0.75	0.84	0.79
Ngrams (651)	0.62	0.63	0.62	0.62	0.62	0.63	0.62
Syntax (1375)	0.65	0.66	0.63	0.64	0.64	0.67	0.65
All Features (2131)	0.74	0.75	0.73	0.74	0.74	0.75	0.74

Table 3: Classification results FakeNews dataset collected via crowdsourcing.

Features (number of features)	Acc.	LEGITIMATE			FAKE		
		P	R	F1	P	R	F1
Punctuation (11)	0.70	0.67	0.77	0.72	0.73	0.63	0.68
LIWC - Summary (7)	0.65	0.66	0.61	0.63	0.64	0.68	0.66
LIWC - Linguistic processes (21)	0.64	0.64	0.63	0.63	0.63	0.64	0.63
LIWC - Psychological processes (40)	0.58	0.58	0.58	0.58	0.58	0.57	0.57
Complete LIWC (79)	0.67	0.68	0.66	0.67	0.67	0.68	0.67
Readability (26)	0.50	0.50	0.48	0.49	0.50	0.51	0.50
Ngrams (1378)	0.67	0.67	0.66	0.66	0.66	0.68	0.67
Syntax (1268)	0.67	0.67	0.68	0.67	0.68	0.66	0.67
All Features (2751)	0.73	0.73	0.72	0.72	0.73	0.74	0.73

Table 4: Classification results for the Celebrity news data set.

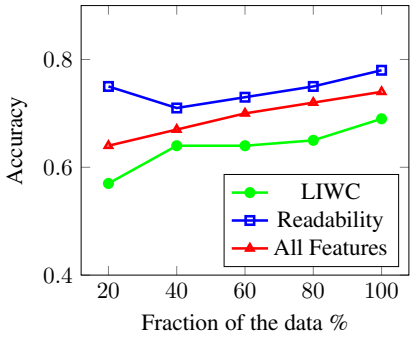

Figure 1: Learning curves on the FakeNewsAMT dataset using three feature sets

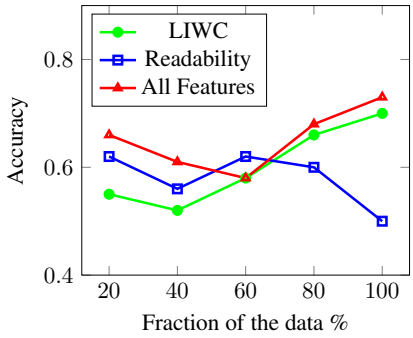

Figure 2: Learning curves on the Celebrity dataset using three feature sets

prove the classification performance.

Cross-domain Analyses. We also explore the applicability of our methods across domains, using the two best feature sets (*Readability* and *Complete LIWC*), as well as the classifier relying on all the features (*All Features*). Table 5 shows the results obtained in cross-domain experiments between the FakeNewsAMT dataset and the Celebrity dataset. Perhaps not surprisingly, there is a significant loss in accuracy as compared to the within-domain results shown in Tables 3 and 4.

The metrics suggest that the generalization from the crowdsourced data to the celebrity news is biased towards the truth (i.e., the classifier almost exclusively predicted the 'true' class). Possible explanations for the drop in performance might be (1) that the linguistic properties of deception in one domain are structurally different from those of deception in a second domain, and (2) that the feature sets applied for the cross-domain evaluation, in particular the readability feature set (accuracy = 0.50), were not performing well in the respective domain in the first place. To test this idea, we also applied cross-domain evaluation where we trained the classifier of domain A on the with the best feature set of domain B and tested in on domain B (here: *Complete LIWC* and *Readability* for the Celebrity and FakeNewsAMT data, respectively). The readability feature set classifier of the Celebrity data yielded an accuracy of 0.61 on the FakeNewsAMT data (compared to the original 0.78), and, vice versa, the *Complete LIWC*

Training	Testing	Feature set	Acc.	$F1_{Legitimate}$	$F1_{Fake}$
Celebrity	FakeNewsAMT	Complete LIWC	0.60	0.62	0.57
		Readability	0.61	0.60	0.67
		All Features	0.56	0.63	0.47
FakeNewsAMT	Celebrity	Complete LIWC	0.61	0.62	0.57
		Readability	0.51	0.67	0.06
		All Features	0.51	0.67	0.08

Table 5: Cross-domain analysis for best performing feature sets

Domain	Readability			Complete LIWC			All features		
	Acc.	$F1_{Legitimate}$	$F1_{Fake}$	Acc.	$F1_{Legitimate}$	$F1_{Fake}$	Acc.	$F1_{Legitimate}$	$F1_{Fake}$
Technology	0.90	0.90	0.90	0.62	0.57	0.64	0.80	0.78	0.81
Education	0.84	0.86	0.81	0.68	0.66	0.69	0.84	0.84	0.83
Business	0.53	0.14	0.67	0.76	0.75	0.77	0.85	0.84	0.86
Sports	0.51	0.26	0.64	0.73	0.74	0.70	0.81	0.81	0.81
Politics	0.91	0.92	0.90	0.73	0.73	0.73	0.75	0.75	0.75
Entertainment	0.61	0.51	0.68	0.70	0.71	0.69	0.75	0.74	0.76

Table 6: Cross-domain classification accuracy for the complete LIWC and readability feature sets

classifier resulted in an accuracy of 0.61 (compared to 0.70). These findings indicate that different linguistic properties underlying different kinds of deception are more likely to explain cross-domain performance decreases than poorly performing feature sets.

We also assess the cross-domain classification performance for the six news domains in the FakeNewsAMT dataset. We do this by training on five of the six domains in the datset, and testing the remaining one. Table 6 shows the results obtained in these experiments. The politics, education, and technology domains appear to be rather robust against classifiers trained on other domains. The technology and politics domains, moreover, are classified both with a high accuracy of 0.91 with the *Readability* feature set, which may suggest that fake and legitimate news in each of these three domains might be structurally similar to the fake and legitimate content in the other five domains. By contrast, domains such as sports, business and entertainment are less generalizable and might therefore be more domain-dependent. Although further research is needed to consolidate these findings, a possible explanation could be the rather unique content and style of these domains

6 Human Performance

To identify a human baseline for the fake news detection task, we conducted a study to evaluate the human ability to spot fake news on the two developed datasets. We created an annotation interface that shows an annotator either a fake or a legitimate news article, and asks them to judge its credibility. We asked annotators to select a label of either "Fake" or "Legitimate" according to their own perceptions. We also asked them to indicate whether or not they have read or heard about the presented news in the past; overall, the annotators read less than 5% of the news before, which we considered to be a negligible fraction.

	Agreement	Kappa
FakeNewsAMT	70%	0.38
Celebrity	73%	0.45

Table 7: Agreement among two human annotators on the FakeNewsAMT and the Celebrity datasets.

	FakeNewsAMT	Celebrity
A1	0.71	0.80
A2	0.70	0.77
Sys	0.74	0.73

Table 8: Performance of two annotators (A1, A2) and the developed automatic system (Sys) on the fake news datasets

Two annotators labeled the news in each dataset. In both cases, the news articles were presented in a random order to avoid annotation bias. Annotators evaluated 480 and 200 news for the FakeNewsAMT and Celebrity datasets respectively. Annotators were not offered a monetary reward and we consider their judgments to be honest as they participated voluntarily in this experiment. Table 7 shows the observed agreement and Kappa statistics for each dataset. Resulting Kappa values show moderate agreement values with slightly lower Kappa for the FakeNewAMT dataset. The results suggest that humans are better at identifying fake news in the celebrity domain than fake news in other domains.

In addition, we evaluate the performance of the automatic fake news classifiers against the human capability to spot fake news. Thus, we compare the accuracy of our system to that of human annotators. Table 8 summarizes the accuracies obtained by the human annotators and our system on the two fake news datasets. Results confirm that humans are better at detecting fake content in the Celebrity domain. Notably, our system outperforms humans while detecting fake news in more serious and diverse news sources.

7 Further Insights

Our experiments suggest important differences in fake news content as compared to legitimate news content. Particularly, we observe that classifiers relying on the semantic information encoded in the LIWC lexicon show consistently good performance across domains. To gain further insights into the semantic classes that are associated with fake and legitimate content, we evaluate which classes show significant differences between the two groups of news. To compare both types of content, we subtract the average percentage of words in each LIWC category in the fake news from its corresponding values in the legitimate news set. Therefore, a positive result indicates an association between a LIWC class and legitimate content, and a negative result indicates an association between a LIWC class and fake content. Results for the FakeNewsAMT and Celebrity datasets are shown in Figures 3 and 4 respectively. All the differences shown in the graphs are statically significant (one tailed t-test, $p < 0.5$).

Figure 3 indicates that the language used to report legitimate content in the FakeNewsAMT dataset, often includes words associated with cognitive processes such as insight and differentiation. In addition, legitimate content includes more function words such as he, she, and negations, and expresses relativity. On the other hand, language used when reporting fake content uses more social and positive words, expresses more certainty and focuses on present and future actions. Moreover, the authors of fake news use more adverbs, verbs, and punctuation characters than the authors of legitimate news. Likewise, results in Figure 4 show noticeable differences among legitimate and fake content on the celebrity domain. Specifically, fake content in tabloid and entertainment magazines seem to use more perceptual words, e.g., hear, see, feeling, and positive emotions categories. In addition, fake content in this domain has a predominant use of the "I" pronoun and prepositions. In contrast, legitimate content uses words that indicate cognitive processes such as insight, cause, discrepancy, and tentative language.

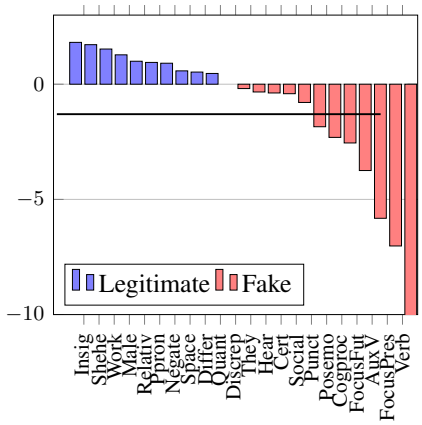

Figure 3: Language differences in fake and legitimate content in the FakeNewsAMT dataset

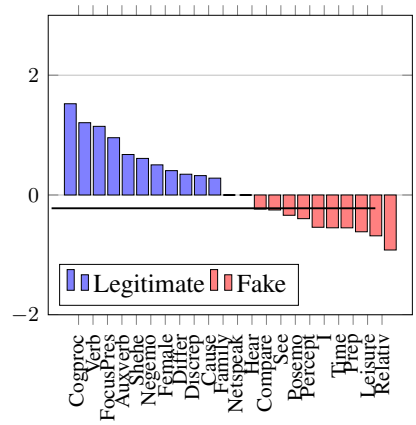

Figure 4: Language differences in fake and legitimate content in the Celebrity dataset

8 Conclusions

In this paper, we addressed the task of automatic identification of fake news. We introduced two new fake news datasets, one obtained through crowdsourcing and covering six news domains, and another one obtained from the web covering celebrities. We developed classification models that rely on a combination of lexical, syntactic, and semantic information, as well features representing text readability properties. Our best performing models achieved accuracies that are comparable to human ability to spot fake content.

References

Gary D Bond and Adrienne Y Lee. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology* 19(3):313–329.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology* 99(3):549.

Gottfried Jeffrey and Shearer Elisa. 2016. News use across social media platforms 2016. In *Pew Research Center Reports*. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '03, pages 423–430. https://doi.org/10.3115/1075096.1075150.

Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. 2016. *caret: Classification and Regression Training*. R package version 6.0-70. https://CRAN.R-project.org/package=caret.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2015. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. https://CRAN.R-project.org/package=e1071.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '05, pages 531–538. https://doi.org/10.3115/1220575.1220642.

M. Newman, J. Pennebaker, D. Berry, and J. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29.

Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *HLT-NAACL*. pages 497–501.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey Hancock. 2011a. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 309–319.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011b. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 309–319. http://www.aclweb.org/anthology/P11-1032.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1120–1125. http://aclweb.org/anthology/D15-1133.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *CoRR* abs/1702.05638. http://arxiv.org/abs/1702.05638.

T. Qin, J. K. Burgoon, J. P. Blair, and J. F. Nunamaker. 2005. Modality effects in deception detection and applications in automatic deception-detection. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology* 52(1):1–4. https://doi.org/10.1002/pra2.2015.145052010083.

Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of NAACL-HLT*. pages 7–17.

Wafa Shafqat, Seunghun Lee, Sehrish Malik, and Hyun-chul Kim. 2016. The language of deceivers: Linguistic features of crowdfunding scams. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 99–100.

C. Toma and J. Hancock. 2010. Reading between the lines: linguistic cues to deception in online datin In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, New York, NY, USA, CSCW '10, pages 5–8. https://doi.org/10.1145/1718918.1718921.

Byron C Wallace. 2015. Computational irony: A survey and new perspectives. *Artificial Intelligence Review* 43(4):467–483.

D. Warkentin, M. Woodworth, J. Hancock, and N. Cormier. 2010. Warrants and deception in computer mediated communication. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, pages 9–12.

Linfeng Zhang and Yong Guan. 2008. Detecting click fraud in pay-per-click streams of online advertising networks. In *Distributed Computing Systems, 2008. ICDCS'08. The 28th International Conference on*. IEEE, pages 77–84.