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 analyses 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, and show that we can achieve accuracies of up to 76%. 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 misinformation online increases, particularly in media outlets such as social media feeds, news blogs, and online newspapers. A recent report by the Jumpshot Tech Blog showed that Facebook referrals accounted for 50% of the total traffic to fake news sites and 20% total traffic to reputable websites.[1] Since as many as 62% of U.S. adults consume news on social media (Jeffrey and Elisa, 2016), being able to identify fake content in online sources is a pressing need.

Until now, computational approaches for fake news detection have relied on satirical news sources such as “The Onion” (Rubin et al., 2016), viral news tracking websites such as BuzzFeed (Potthast et al., 2017) and fact-checking websites such as ”politiFact” (Wang, 2017) and “Snopes” (Popat et al., 2016). 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). Moreover, fact-checking websites are usually constrained to a particular domain of interest, such as politics, and require human expertise to verify the news claims making it difficult to obtain datasets that provide some degree of generalization over other domains (Chen et al., 2015).

In this paper, we develop computational resources and models for the task of fake news detection. We introduce two novel datasets covering seven different domains. One of the datasets is collected by combining manual and crowdsourced annotation approaches, 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 news content, and we build fake news detectors relying on linguistic features that achieve accuracies of up to 76%. To place our results in perspective, we compare the performance of the developed classifiers with an empirical human baseline.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

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

Proceedings of the 27th International Conference on Computational Linguistics, pages 3391–3401
Santa Fe, New Mexico, USA, August 20-26, 2018.

2 Related Work

To date, there are two important lines of research into the automated classification of genuine and 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). Second, on an operational level, linguistic and fact-checking based approaches have been proposed to discriminate between real and fake news content (Conroy et al., 2015).

The linguistic approach attempts to identify text properties, such as writing style and content, that can help to discriminate real from fake news articles. The underlying assumption for this approach is that linguistic behaviors such as punctuation usage, word type choices, part-of-speech tags, and emotional valence of a text are rather involuntary and therefore outside of the author's control, thus revealing important insights into the nature of the text. The linguistic approach has yielded promising results in differentiating satire from real news (Rubin et al., 2016). Relying in 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), the authors explored the use of several linguistic features to discriminate between real and satirical news content. The best classification performances were achieved with feature sets representing absurdity, punctuation, and grammar.

On the other hand, fact-checking approaches rely on automated verification of propositions made in the news articles (e.g., "Barack Obama assumed office on a Tuesday") to assess the truthfulness of their claims (Conroy et al., 2015). Knowledge databases such as DBpedia [2] have been used to query the Web in a structured manner. The results of such queries can then be used to test whether different sources also contain information confirming the news claim (e.g., that Barack Obama assumed office on a Tuesday). Other works have used social network activity (e.g., tweets) on a specific news item to assess its credibility, for instance by identifying tweets voicing skepticism about the truthfulness of a claim made in a news article (Hannak et al., 2014; Jin et al., 2014). Although fact-checking approaches are becoming increasingly powerful, a major drawback is that they are built on the premise that the information can be verified using external sources, for instance FakeCheck.org and Snopes.com. However, this is not a straightforward task, as external sources might not be available, particularly for just-published news items. Therefore, the fact-checking approach is predominantly useful for the detection of deception in texts for which external, verifiable information is available.

Furthermore, also related to the current paper is work on the automatic identification of deceptive content, which has explored domains such as forums, consumer reviews websites, online advertising, online dating, and crowdfunding platforms (Warkentin et al., 2010; Ott et al., 2011a; Zhang and Guan, 2008; Toma and Hancock, 2010; Shafqat et al., 2016). While fake news detection is closely related to deception detection (i.e. determining whether or not someone is lying), there are important differences between the two tasks. First, fake news producers usually seek political or financial gain as well as self-promotion while deceivers have motivations that are more socially driven such as self protection, conflict or harm avoidance, impression management or identity concealment. Second, they differ significantly in their target and in the form they propagate: fake news items are usually disseminated at larger scale through the Internet and social media whereas deception is more specifically targeted at individuals. However, since both tasks deal with deceptive content, we hypothesize that there are linguistic aspects that might be shared between these tasks. Thus, we focus on the linguistic approach and build upon an emerging body of research on computer-automated verbal deception detection (Fitzpatrick et al., 2015).

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., "poli-

[2] http://wiki.dbpedia.org/about

Dataset	Class	Entries	Average Words/Sent	Words
FakeNewsAMT	Fake	240	132/5	31,990
	Legitimate	240	139/5	33,378
Celebrity	Fake	250	399/17	39,440
	Legitimate	250	700/33	70,975

Table 1: Class distribution and word statistics for fake news datasets

tiFact" or "Snopes"), which are typically focused on only one domain (generally politics). To address these shortcomings, we decided to construct two novel datasets containing fake news covering several news domains and specifically model the deceptive property of fake news. One dataset is collected via crowdsourcing covering six news domains (e.g., business, education). The second dataset is obtained directly from the web and covers celebrity news.

3.1 Crowdsourced Fake News 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.

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.	**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.

Table 2: 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 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."

Table 3: Sample legitimate and web fake news in the Celebrity domain

Crowdsourcing Fake News. To generate fake versions of the legitimate news items, we made use of crowdsourcing via Amazon Mechanical Turk (AMT). Despite having been successfully used in the past to collect deceptive 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), the use of

AMT on the news domain poses additional challenges. 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 so we could obtain news with homogeneous writing style. To simplify the fake news production task (and also address the latter challenge), we also opted for working with a shorter version of the original news article. Thus, we manually select a news excerpt – about two or three paragraphs – that summarizes the news article. This process resulted in 240 news excerpts derived from the legitimate news dataset collected earlier.

Next, we set up an AMT task that asked workers to generate a fake version of a given 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, to ensure crowdsourcing quality, we restricted participation to workers who had an AMT approval rate of at least 95% for previous tasks.

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. The corpus statistics, including class distribution and word/sentence statistics, are shown in Table 1. Table 2 shows an excerpt of a fake news article in our dataset, along with its legitimate version, in the technology domain.

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 that drive individuals to produce utterances that match the grammatical structure of sentences they have recently read (Ireland and Pennebaker, 2010).

Importantly, note that the AMT process of generating news mirrors the fake news production process quite well: similar to the AMT workers, the producers of fake news write for the purpose of generating quick money, and do not undergo the same professional writing training that the writers of legitimate news do.

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

3.2 Web Dataset Celebrity

For our second dataset, we sought to collect news from web sources 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, and politicians) and our sources include online magazines such as Entertainment Weekly, People Magazine, RadarOnline, among other tabloid and entertainment-oriented publications. The data was 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 also cross-referenced with information from other entertainment news sources on the web.

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 limiting 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.

Using this approach, we collected a total of 500 news articles, with an even distribution for fake and legitimate news. The corpus statistics, including class distribution and word/sentence statistics, are shown in Table 1. Table 3 shows a example excerpt of a celebrity fake/legitimate news 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 twelve 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; Ott et al., 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).[3]

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 from 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 –noun phrase– the parent node, and "commissions" the child node). CFG-based features have been previously shown to be useful for linguistic deception detection (Feng et al., 2012). Features in this set are also encoded as tf-idf values.

5 Automatic Fake News Detection

We conduct several experiments with different combinations of feature sets to explore their predictive separately and jointly. We use a linear SVM classifier and conduct our evaluations using five-fold cross-validation, with accuracy, precision, recall, and F-score as performance metrics. We use the machine learning algorithms implementation available in the caret (Kuhn et al., 2016) and e1071 packages (Meyer et al., 2015) with their default parameters.

Tables 4 and 5 show the results obtained for the different feature sets and the two datasets. Since our datasets contain an even distribution between fake and real news items, we use a random baseline of 50% as reference value. As seen in the tables most of the classifiers obtain performances well above the baseline, which indicates that the task of fake news detection can be effectively addressed using linguistic

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

features. For the FakeNewsAMT dataset, the best performing classifiers are the ones that rely on stylistic features (i.e., Punctuation and Readability), followed by the ones build using psycholinguistic features drawn from the LIWC lexicon. The classifiers build with the Celebrity dataset show the best performance when using the LIWC features, followed by the ngrams and syntactic features (CFG). Overall, our results suggest that fake news differ from real news mainly in aspects such as writing style (punctuation, readability, syntactic structure) and aspects related to writer's internal processes (LIWC features). Regarding the differences in performance between the two datasets, we believe that they can be attributed to the domain in which the news are generated. For instance, to spot fake news in more serious topics such as technology or education we might need to pay more attention to linguistic aspects of writing whereas to spot fake news in the celebrity domain we might need to focus on writing differences related to people's feelings and perceptions.

Finally, our results show that when using all the features on the two datasets we achieve the best accuracies, with 0.74 and 0.76 respectively. These results suggest that an integrated use of linguistic, syntactic and semantic features is useful to discriminate between real and fake news content.

Features (# features)	Acc.	$\text{F1}_{Legit.}$	F1_{Fake}
Punctuation (12)	0.71	0.69	0.72
LIWC-Summ (7)	0.61	0.58	0.64
LIWC-LingProc. (21)	0.67	0.66	0.66
LIWC-PsyProc. (40)	0.56	0.56	0.55
LIWC (80)	0.70	0.70	0.70
Readability (26)	0.78	0.77	0.79
Ngrams (634)	0.62	0.62	0.62
CFG (1377)	0.65	0.64	0.65
All Features (2140)	0.74	0.74	0.74

Table 4: Classification results for the FakeNewsAMT dataset collected via crowdsourcing.

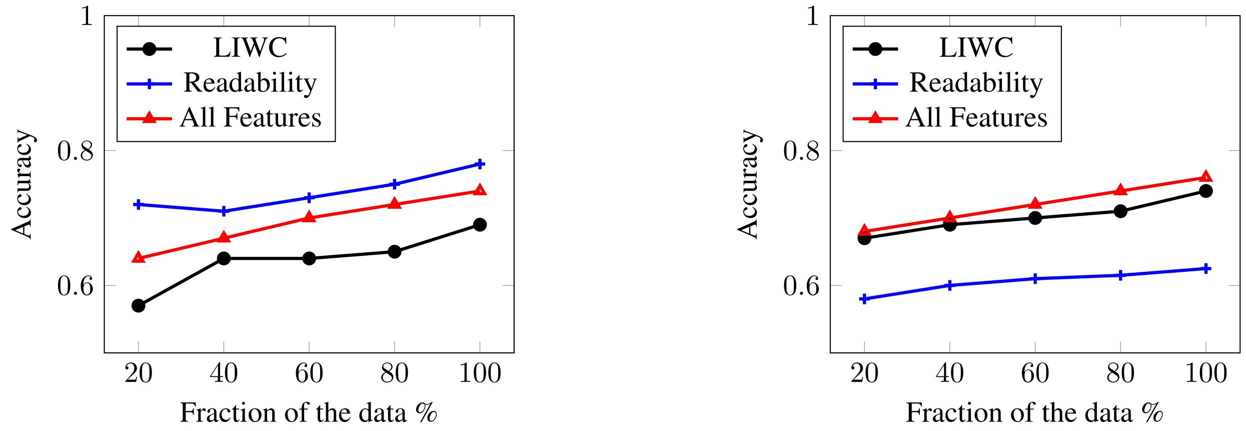
Features (# features)	Acc.	$\text{F1}_{Legit.}$	F1_{Fake}
Punctuation (12)	0.69	0.69	0.69
LIWC-Summ. (7)	0.67	0.66	0.69
LIWC-LingProc (21)	0.72	0.72	0.71
LIWC-PsyProc (40)	0.67	0.68	0.66
LIWC (80)	0.74	0.74	0.74
Readability (28)	0.62	0.61	0.63
Ngrams (1317)	0.71	0.72	0.71
CFG (2599)	0.72	0.72	0.72
All Features (4048)	0.76	0.77	0.76

Table 5: Classification results for the Celebrity news dataset.

Learning Curves. Next, seeking to investigate whether larger amounts of training data can improve the identification of fake content, we analyzed the learning trend of our best classifiers. Thus, we plot the learning curves of the LIWC features, readability features, and the combination of all features sets using incremental amounts of data as shown in Figures 1a and 1b. Overall, the learning trend in both datasets shows steady improvement, thus suggesting that larger quantities of training data could improve the classification performance.

Cross-domain Analyses. We also explore the applicability of our methods across domains, using the two best feature sets identified during our previous experiments, (*Readability* and *LIWC*), as well as the classifier relying on all the features (*All Features*).

Table 6 shows the results obtained in cross-domain experiments, where we train our models using the



(a) FakeNewsAMT dataset

(b) Celebrity dataset

Figure 1: Learning curves using incremental fraction of the data and three feature sets

Training	Testing	Feature set	Accuracy
FakeNewsAMT	Celebrity	LIWC	0.48
		Readability	0.52
		All Features	0.50
Celebrity	FakeNewsAMT	Complete LIWC	0.60
		Readability	0.65
		All Features	0.64

Table 6: Cross-domain analysis for best performing feature sets.

FakeNewsAMT dataset and test on the Celebrity dataset. Perhaps not surprisingly, there is a significant loss in accuracy as compared to the within-domain results shown in Tables 4 and 5.

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 = 0.62), 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 classifiers using the celebrity domain (Celebrity) and tested in the other domain (FakeNewsAMT).

This time, the readability feature set classifier of the Celebrity data yielded an accuracy of 0.65 on the FakeNewsAMT data (compared to the original 0.78) and similarly, the *LIWC* classifier resulted in an accuracy of 0.60 (compared to 0.70). Likewise, the performance using all features dropped from 0.74 and 0.76 to 0.50 and 0.64 for the FakeNewsAMT and celebrity datasets, respectively. Overall, these findings hint at the important role of domain in the fake news detection.

As an additional experiment, we 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 dataset, and testing on the remaining one. Table 7 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 both classified with a high accuracy of 0.90 and 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.

Test Domain	Readability	LIWC	All features
Technology	0.90	0.62	0.80
Education	0.84	0.68	0.84
Business	0.53	0.76	0.85
Sports	0.51	0.73	0.81
Politics	0.91	0.73	0.75
Entertainment	0.61	0.70	0.75

Table 7: Cross-domain classification accuracy for the complete LIWC and readability feature sets. Training data consists of all but the test domains in the FakeNewsAMT dataset.

	Agreement	Kappa
FakeNewsAMT	70%	0.38
Celebrity	73%	0.45

Table 8: Agreement among two human annotators on the FakeNewsAMT and the Celebrity datasets.

6 Human Performance

Fake news detection is a challenging task for humans as readers frequently find themselves sharing fake news content or being lured by clickbait headlines. Seeking 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 "Fake" or "Legitimate" according to their own perceptions upon reading the news item. We also asked them to indicate whether or not they have read or heard about the presented news item in the past; overall, the annotators read less than 5% of the news before, which we considered a negligible fraction.

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 8 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.

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 9 summarizes the accuracies obtained by the human annotators and our system on the two fake news datasets. The findings indicate that humans are better at detecting fake content in the Celebrity domain than in the other fake news domain. Notably, our system outperforms humans while detecting fake news in more serious and diverse news sources.

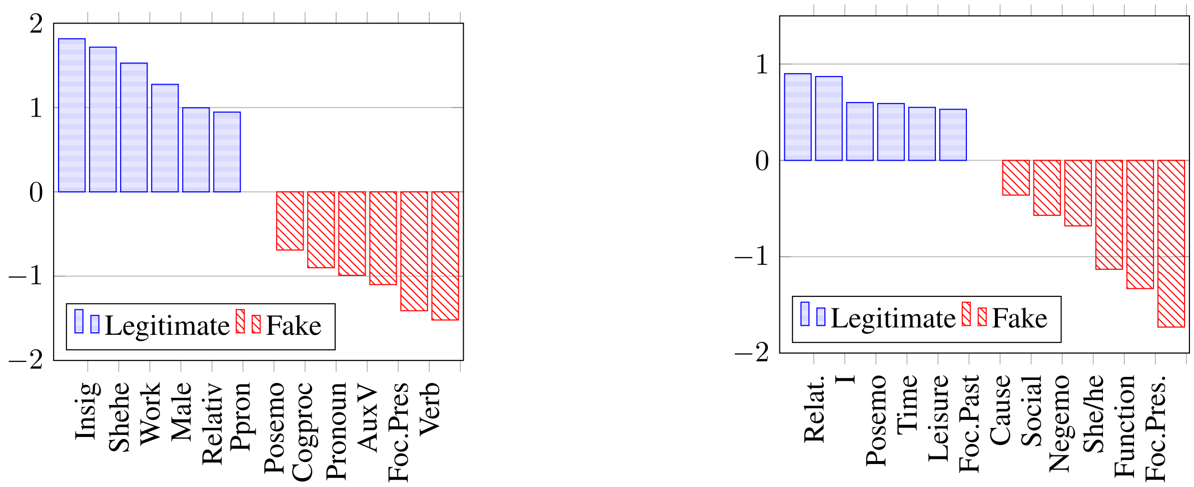
7 Further Insights

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 2a and 2b respectively. All the differences shown in the graphs are statically significant (one-tailed t-test, $p < 0.05$).

Figure 2a indicates that the language used to report legitimate content in the FakeNewsAMT dataset

	FakeNewsAMT	Celebrity
A1	0.71	0.80
A2	0.70	0.77
Sys	0.74	0.76

Table 9: Performance of two annotators (A1, A2) and the developed automatic system (Sys) on the fake news datasets



(a) FakeNewsAMT

(b) Celebrity

Figure 2: Language differences in each dataset using the LIWC word categories

often includes words associated with cognitive processes such as insight and differentiation. In addition, legitimate content includes more function words (e.g., pronouns such as he, she), negations, and expressions of 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, the results in Figure 2b show noticeable differences among legitimate and fake content on the celebrity domain. Specifically, legitimate news in tabloid and entertainment magazines seem to use more first person pronouns, talk about time (Relativity,Time, FocusPast), and use positive emotion words (posemo), which interestingly were also found as markers of truth-tellers in previous work on deception detection (Pérez-Rosas and Mihalcea, 2014). On the other hand, fake content in this domain has a predominant use of second person pronouns (he, she), negative emotion words (negemo) and focus on the present (Foc.Pres).

8 Conclusions

With an increasing focus of academic researchers and practitioners alike on the detection of online misinformation, the current investigation allows for two key conclusions.

First, computational linguistics can aide in the process of identifying fake news in an automated manner well above the chance level. The proposed linguistics-driven approach suggests that to differentiate between fake and genuine content it is worthwhile to look at the lexical, syntactic and semantic level of a news item in question. The developed system's performance is comparable to that of humans in this task, with an accuracy up to 76%. Nevertheless, while linguistics features seem promising, we argue that future efforts on misinformation detection should not be limited to these and should also include *meta* features (e.g., number of links to and from an article, comments on the article), features from different modalities (e.g., the visual makeup of a website using computer vision approaches), and embrace the increasing potential of computational approaches to fact verification (Thorne et al., 2018). Thus, future work might want to explore how hybrid decision models consisting of both fact verification and data-driven machine learning judgments can be integrated.

Second, we showed that it is possible to build resources for the fake news detection task by combining manual and crowsourced annotation approaches. Our paper presented the development of two datasets using these strategies and showed that they exhibit linguistic properties related to deceptive content. Furthermore, different from other available fake news datasets, our dataset consists of actual news excerpts, instead of short statements containing fake news information.

Finally, with the current investigation and dataset, we encourage the research community and practitioners to take on the challenge of tackling misinformation. The datasets introduced in this paper are publicly available at `http://lit.eecs.umich.edu/downloads.html`.

Acknowledgments

This material is based in part upon work supported by the Michigan Institute for Data Science and by the National Science Foundation (grant #1344257). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science or the National Science Foundation.

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.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. News in an online world: The need for an ”automatic crap detector”. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics.

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari. 2015. Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies*, 8(3):1–119.

Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, Ann Arbor, MI, June.

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*.

Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 230–239. IEEE.

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

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.

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.

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*, HLT '05, pages 531–538, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, HLT '11, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, pages 309–319, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

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. 2014. Cross-cultural deception detection. In *ACL (2)*, pages 440–445.

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*, pages 1120–1125, Lisbon, Portugal, September. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2173–2178, New York, NY, USA. ACM.

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

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.

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*, pages 99–100. International World Wide Web Conferences Steering Committee.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355.*

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

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

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 (Volume 2: Short Papers)*, pages 422–426. Association for Computational Linguistics.

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*, pages 9–12. ACM.

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*, pages 77–84. IEEE.