Fake News Early Detection: A Theory-driven Model

XINYI ZHOU, ATISHAY JAIN, VIR V. PHOHA, and REZA ZAFARANI, Syracuse University, USA

The explosive growth of fake news and its erosion of democracy, justice, and public trust has significantly increased the demand for accurate fake news detection. Recent advancements in this area have proposed novel techniques that aim to detect fake news by exploring how it propagates on social networks. However, to achieve fake news early detection, one is only provided with limited to no information on news propagation; hence, motivating the need to develop approaches that can detect fake news by focusing mainly on news content. In this paper, a theory-driven model is proposed for fake news detection. The method investigates news content at various levels: lexicon-level, syntax-level, semantic-level and discourse-level. We represent news at each level, relying on well-established theories in social and forensic psychology. Fake news detection is then conducted within a supervised machine learning framework. As an interdisciplinary research, our work explores potential fake news patterns, enhances the interpretability in fake news feature engineering, and studies the relationships among fake news, deception/disinformation, and clickbaits. Experiments conducted on two real-world datasets indicate that the proposed method can outperform the state-of-the-art and enable fake news early detection, even when there is limited content information.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; • **Computing methodologies** → **Natural language processing**; **Machine learning**; • **Security and privacy** → *Social aspects of security and privacy*; • **Applied computing** → *Sociology*; *Computer forensics.*

Additional Key Words and Phrases: Fake news, fake news detection, news verification, disinformation, clickbait, feature engineering, interdisciplinary research

ACM Reference Format:
Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2019. Fake News Early Detection: A Theory-driven Model. 1, 1 (April 2019), 24 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Fake news is now viewed as one of the greatest threats to democracy and journalism [60]. The reach of fake news was best highlighted during the critical months of the 2016 U.S. presidential election campaign, where the top twenty frequently-discussed fake election stories, one of which has been illustrated in Figure 1, generated 8,711,000 shares, reactions, and comments on Facebook, ironically, larger than the total of 7,367,000 for the top twenty most-discussed election stories posted by 19 major news websites [47]. Our economies are not immune to the spread of fake news either, with fake news being connected to stock market fluctuations and massive trades. For example, fake news claiming that Barack Obama was injured in an explosion wiped out $130 billion in stock value [38].

Meanwhile, humans have been proven to be irrational and vulnerable differentiating between truth and falsehood when overloaded with deceptive information. Studies in social psychology and communications have demonstrated that human ability to detect deception is only slightly

Authors' address: Xinyi Zhou, zhouxinyi@data.syr.edu; Atishay Jain, atjain@syr.edu; Vir V. Phoha, vvphoha@syr.edu; Reza Zafarani, reza@data.syr.edu, EECS Department, Syracuse University, Syracuse, NY, 13244, USA.

XXXX-XXXX/2019/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

IT'S OVER: Hillary's ISIS Email Just Leaked & It's Worse Than Anyone Could Have Imagined...

POSTED BY FRIENDSOFSYRIA IN WAR CRIMES ≈ 351 COMMENTS

– Hillary Clinton, Friend of the Syria people? Like the USA is friends of the people of Iraq, Afghanistan, Pakistan, Libya, Somalia, Yemen...?

Today Wikileaks released what is, by far, the most devastating leak of the entire campaign. This makes Trump's dirty talk video looks like an episode of Barney and Friends.

Even though when Trump called Hillary the 'founder' of ISIS he was telling the truth and 100% accurate, the media has never stopped ripping him apart over it.

Today the media is forced to eat their hats because the newest batch of leaked emails show Hillary, in her own words, admitting to doing just that, funding and running ISIS.

John Podesta, Hillary's campaign chair, who was also a counselor to President Obama at the time, was the recipient of the 2014 email which was released today.

Assange promised his latest batch of leaks would lead to the indictment of Hillary, and it looks like he was not kidding. The email proves Hillary knew and was complicit in the funding and arming of ISIS by our 'allies' Saudi Arabia and Qatar!

Fig. 1. Fake News[5]. (1) This fake news story originally published on Ending the Fed has got ∼754,000 engagements in the final three months of the 2016 U.S. presidential campaign, which is the top-three-performing fake election news story on Facebook [47]; (2) It is a fake news story with clickbait.

better than chance: typical accuracy rates are in the range of 55%-58%, with a mean accuracy of 54% over 1,000 participants in over 100 experiments [41]. Many expert-based (e.g., PolitiFact[1] and Snope[2]) and crowd-sourced (e.g., Fiskkit[3] and TextThresher [56]) manual fact-checking websites, tools and platforms thus have emerged to serve the public on this matter[4]. Nevertheless, manual fact-checking does not scale well with the volume of newly created information, especially on social media [55]. Hence, automatic fake news detection has been developed in recent years, where current methods can be generally grouped into (I) *content-based* and (II) *propagation-based* methods.

I. Content-based Fake News Detection aims to detect fake news by analyzing the content of news articles, often formed by a title (headline) and body-text and sometimes accompanied by the author(s), image(s) and/or video(s). To detect fake news using content, researchers often rely on the *knowledge* (i.e., SPO (Subject, Predicate, Object) tuples) [11, 44], *style* [34] or *latent features* [52] of the content. When relying on the knowledge within a news article to detect whether or not it is fake, one can compare the knowledge extracted from it to that of stored in a *Knowledge Graph (KG)* as a source of ground truth. The construction of such a knowledge graph is still an open problem, particularly for fake news detection. First, based on Oxford Dictionaries, news is defined as "newly received or noteworthy information especially about recent events", which implies that such knowledge graphs should be time-sensitive [58]. Second, knowledge graphs are often far from complete, which requires developing approaches for *knowledge inference* [30]. Third, fake news is defined as "news that is intentionally and verifiably false" [45]; such knowledge-based approaches can help verify news authenticity however cannot verify the intentions being creating news articles [58]. When using the style of the news content to detect fake news, current techniques aim to capture some [non-latent] characteristics within news content, e.g., word-level statistics based on Term Frequency-Inverse Document Frequency (TF-IDF) [34], *n*-gram distribution [34]

[1] https://www.politifact.com/
[2] https://www.snopes.com/
[3] http://www.fiskkit.com/
[4] Comparison among common fact-checking websites is provided in [58] and a comprehensive list of fact-checking websites is available at https://reporterslab.org/fact-checking/.
[5] Direct source: https://bit.ly/2uE5eaB

and/or utilize Linguistic Inquiry and Word Count (LIWC) features [33]. Finally to detect fake news using the latent characteristics within news content, neural networks such as Convolutional Neural Network (CNN) [49, 52] have been developed to automatically select content features.

Nevertheless, in all such techniques, fundamental theories in social and forensic psychology have not played a significant role. Such theories can significantly improve fake news detection by highlighting some potential fake news patterns and facilitating interpretable machine learning models for fake news detection [28, 54]. For example, *Undeutsch hypothesis* [48] states that a fake statement differs in writing style and quality from a true one. Such theories, as will be discussed later, can refer to either *deception/disinformation* [22, 26, 48, 61], i.e., information that is intentionally and verifiably false, or *clickbaits* [24], the headlines whose main purpose is to attract the attention of readers and encourage them to click on a link to a particular webpage [58]. Compared to existing style features and latent features, relying on such theories allows on to introduce *theory-driven* features that are interpretable, can help the public well understand fake news, and help explore the relationships among fake news, deception/disinformation and clickbaits. Theoretically, deception/disinformation is a more general concept which includes fake news articles, fake statements, fake reviews, etc. Hence the characteristics attached to deception/disinformation might or might not be consistent with that of fake news, which motivates to explore the relationships between fake news and deceptions. Meanwhile, clickbaits have been shown to be closely correlated to fake news [5, 10]. The fake election news story in Figure 1 is an example of a fake news story with a clickbait. When fake news meets clickbaits, we observe news articles that can attract eyeballs but are rarely news worthy [32]. Unfortunately, clickbaits help fake news attract more clicks (i.e., visibility) and further gain public trust, as indicated by the *attentional bias* [25], which states that the public trust to a certain news article will increase with more exposure, as facilitated by clickbaits. On the other hand, while news articles with clickbaits are generally unreliable, not all such news articles are fake news, which motivates to explore the relationships between fake news and clickbait.

II. Propagation-based Fake News Detection aims to detect fake news by exploring how news propagates on social networks. Propagation-based methods have gained recent popularity where novel models have been proposed exhibiting good performance. For example, Jin et. al [20, 21] construct a *stance graph* based on user posts, and detects fake news by exploring stance correlations within a graph optimization framework. By exploring relationships among news articles, publishers, users (spreaders) and user posts, propagation-based methods often rely on matrix/tensor factorization [16, 46] and Recurrent Neural Networks (RNNs) [43, 57] to detect fake news. However, to detect fake news at an early stage (i.e., before it becomes wide-spread) in order to take early actions for fake news intervention (i.e., *fake new early detection*), one has to rely on news content and a limited amount of social context information, which can negatively impact the performance of propagation-based fake news detection models, in particular, those that are based on a deep learning framework. Such early detection is particularly crucial for fake news as more individuals become exposed to some fake news, the more likely they may trust it [4]. Meanwhile, it has been demonstrated theoretically [3] and empirically [40] that it is difficult to correct one's cognition after fake news has gained their trust.

In summary, current development in fake news detection strongly motivates the need for techniques that deeply mine news content and rely less on how fake news propagates. Such techniques should investigate how social and forensic theories can help detect fake news for interpretablity reasons [59]. Here, we aim to address these challenges by developing a theory-driven fake news detection model that solely relies on news content. The model represents news articles by a set of manual features, which capture both content structure and style across language levels (i.e., lexicon-level, syntax-level, semantic-level and discourse-level) via conducting an interdisciplinary research. Features are then utilized for fake news detection within a supervised machine learning framework. The specific contributions of this paper are as follows:

(1) The proposed model enables fake news early detection. First, by solely relying on news content, the model allows to conduct detection before fake news has been disseminated on social media. Second, experimental results on real-world datasets indicate the model performs comparatively well among content-based models when limited news content information is available.
(2) The proposed model identifies fake news characteristics, which are inspired by well-established social and psychological theories, and captured respectively at the lexicon-, syntax-, semantic- and discourse-level within language. Compared to latent features, such theory-driven features can enhance model interpret-ability, help fake news pattern discovery, and help the public better understand fake news. Experimental results indicate that the proposed model can outperform baselines including which uses both news content and propagation information.
(3) Our work explores the relationships among fake news, deception/disinformation and clickbaits. By empirically studying their characteristics in, e.g., content quality, sentiment, quantity and readability, some fake news patterns unique or shared with deceptions or clickbait are revealed.

The rest of this paper is organized as follows. Literature review is presented in Section 2. The proposed model is specified in Section 3. In Section 4, we evaluate the performance of our model on two real-world datasets. Section 5 concludes the whole work.

2 RELATED WORK

Depending on whether the approaches detect fake news by exploring its content or by exploring how it propagates on social networks, current fake news detection studies can be generally grouped into content-based and propagation-based methods. We review recent advancements on both fronts.

2.1 Content-based Fake News Detection

In general, current content-based approaches detect fake news by representing news content in terms of features within a machine learning framework. Such representation of news content can be from the perspective of (I) knowledge or (II) style, or can be a (III) latent representation.

I. Knowledge is often defined as a set of SPO (Subject, Predicate, Object) tuples extracted from text. An example of such knowledge (i.e., SPO tuples) is (DonaldTrump, Profession, President) for the sentence “Donald Trump is the president of the U.S.” Knowledge-based fake news detection aims to directly evaluate news authenticity by comparing the knowledge extracted from to-be-verified news content with that within a Knowledge Graph (KG) such as Knowledge Vault [12]. KGs, often regarded as ground truth datasets, contain massive manually-processed relational knowledge from the open Web. However, one has to face various challenges within such a framework. Firstly, KGs are often far from *complete*, often demanding further postprocessing approaches for knowledge inference [30]. Second, news, as newly received or noteworthy information especially about recent events, demands knowledge to be *timely* within KGs. Third, knowledge-based approaches can only evaluate if the to-be-verified news article is false instead of being *intentionally* false, where the former refers to false news while the latter refers to fake news [58].

II. Style is a set of self-defined [non-latent] machine learning features that can represent fake news and differentiate it from the truth [58]. For example, such style features can be word-level statistics based on TF-IDF, n-grams and/or LIWC features [34, 36], rewrite-rule statistics based on TF-IDF [34], rhetorical relationships based on Rhetorical Structure Theory (RST) [35, 42], and content readability [34, 36].

III. Latent features represent news articles via automatically generated features often obtained by matrix/tensor factorization or deep learning techniques, e.g., Text-CNN [52]. Though these style and latent features can be comprehensive and perform well in detecting fake news, their selection or extraction is driven by experience or techniques that are often not supported by theories, which brings challenges to promoting the public's understanding of fake news and comprehending the generated features.

2.2 Propagation-based Fake News Detection

Propagation-based fake news detection further utilizes social context information to detect fake news, e.g., how fake news propagates on social networks, who spreads the fake news, and how spreaders connect with each other [29].

A direct way of presenting news propagation is using a *news cascade* [58] - a tree structure presenting post-repost relationships for each news article on social media, e.g., tweets and retweets on Twitter. Based on news cascades, Vosoughi et al. investigate the differential diffusion of true and fake news stories distributed on Twitter from 2006 to 2017, where the data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times [50]. The authors discover that fake news diffuses significantly farther, faster, more broadly, and can involve more individuals than the truth. They observe that these effects are more pronounced for fake political news than for fake news about terrorism, natural disasters, science, urban legends, or financial information. Wu et al. [53] extend news cascades by introducing user roles (i.e., opinion leaders or normal users), stance (i.e., approval or doubt) and sentiments expressed in user posts. By assuming that the overall structure of fake news cascades differs from true ones, the authors develop a random walk graph kernel to measure the similarity among news cascades and detect fake news based on such similarity. In a similar study, by selecting features from user profiles, tweets and news cascades, Castillo et al. [6] evaluate news credibility within a supervised machine learning framework.

In addition to news cascades, some self-defined graphs that can indirectly represent news propagation on social networks are also constructed for fake news detection. Jin et al. [20, 21] build a stance graph based on user posts, and detect fake news by mining the stance correlations within a graph optimization framework. By exploring relationships among news articles, publishers, users (spreaders) and user posts, PageRank-like algorithm [15], matrix and tensor factorization [16, 46], or Recurrent Neural Networks (RNN) [43, 57] have been developed for fake news detection.

While remarkable progress has been made, to detect fake news early, one cannot rely on social context information (e.g., propagation patterns) and in turn, propagation-based methods, as only limited or no social context information is available at the time of posting for fake news articles. Hence, to design a fake news early detection technique, we solely rely on mining news content.

3 METHODOLOGY

As suggested by *Undeutsch hypothesis* [48], fake news potentially differs in *writing style* from true news. Thus, we represent news content by capturing its writing style respectively at lexicon-level (Section 3.1), syntax-level (Section 3.2), semantic-level (Section 3.3) and discourse-level (Section 3.4). Such representation then can be utilized to predict fake news within a machine learning framework.

3.1 Lexicon-level

To capture news writing style at lexicon-level, we investigate the frequency of words being used in news content, where such frequency can be simply obtained by a *Bag-Of-Word (BOW)* model. However, BOW representation can only capture the absolute frequencies of terms within a news article rather than their *relative (standardized) frequencies* which have accounted for the impact of content length (i.e., the overall number of words within the news content); the latter is more

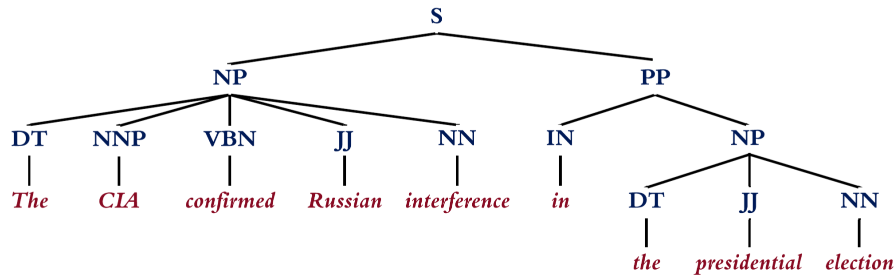


Fig. 2. PCFG Parsing Tree for the sentence "The CIA confirmed Russian interference in the presidential election" within a fake news article. The rewrite rules of this sentence should be the following: S → NP PP, NP → DT NNP VBN JJ NN, PP → IN NP, NP → DT JJ NN, DT → 'the', NNP → 'CIA', VBN → 'confirmed', JJ → 'Russian', NN → 'inference', IN → 'in', JJ → 'presidential' and NN → 'election'.

representative when extracting writing style features based on the words or topics that authors prefer to use or involve. Therefore, we use a standardized BOW model to represent the writing style of each news article at the lexicon-level.

3.2 Syntax-level

Syntax-level style features can be further grouped into shallow syntactic features and deep syntactic features [13], where the former investigates the frequency of *Part-Of-Speech (POS) tags* (e.g., nouns, verbs and determiners) and the latter investigates the frequency of productions (i.e., *rewrite rules*). The rewrite rules of a sentence within a news article can be obtained based on Probability Context Free Grammar (PCFG) parsing trees. An illustration is shown in Figure 2. Here, we also compute the frequencies of POS tags and rewrite rules of a news articles in a relative (standardized) way, which removes the impact of news content length (i.e., divides the overall number of POS tags or rewrite rules within the news content).

3.3 Semantic-level

Style features at semantic-level investigate some psycho-linguistic attributes, e.g., sentiments, expressed in news content. Such attributes defined and assessed in our work are basically inspired by fundamental theories initially developed in forensic- and social-psychology, where clickbait-related attributes target news headlines (Section 3.3.1) and deception/disinformation-related ones are mainly concerned with news body-text (Section 3.3.2). A detailed list of semantic-level features defined and selected in our study is provided in Appendix A.

3.3.1 ClickBait-related Attributes (CBAs). Clickbaits have been suggested to have a close relationship with fake news, where clickbaits help enhance *click-through rates* for fake news articles and in turn, further gain public trust [25]. We aim to extract a set of features that can well represent clickbaits to capture fake news headlines, which also provides an opportunity to empirically study the relationship between fake news and clickbaits. We evaluate news headlines from the following four perspectives.

A. General Clickbait Patterns. We have utilized two public dictionaries[6] that provide some common clickbait phrases and expressions such as "can change your life" and "will blow your mind" [14]. A general way of representing news headlines based on these dictionaries is to verify

[6]https://github.com/snipe/downworthy

if a news headline contains any of the common clickbait phrases and/or expressions listed, or how frequent such common clickbait phrases and/or expressions are in the news headline. Due to the length of news headlines, here the frequency of each clickbait phrase or expression is not considered in our feature set as it leads to many zeros in our feature matrix. Such dictionaries have been successfully applied in clickbait detection [8, 18, 37].

B. Readability. Psychological research has indicated that a clickbait attracts public eyeballs and encourages clicking behavior by creating an *information gap* between the knowledge within the news headline and individuals' existing knowledge [24]. Such information gap has to be produced on the basis that the readers have understood what the news headline expresses. Therefore, we investigate the readability of news headlines by employing several well-established metrics developed in education, e.g., Flesch Reading Ease Index (FREI), Flesch-Kincaid Grade Level (FKGL), Automated Readability Index (ARI), Gunning Fog Index (GFI), and Coleman-Liau Index (CLI). We also separately consider and include as features the parameters within these metrics, i.e., the number of characters, syllables, words, and long (complex) words.

C. Sensationalism. To produce an information gap [24], further attract public attention, and encourage users to click, expressions with exaggeration and sensationalism are common in clickbaits. As having been suggested in clickbait dictionaries [14], clickbait creators prefer to use "can change your life" which might actually "not change your life in any meaningful way"; or use "will blow your mind" to replace "might perhaps mildly entertain you for a moment", where the former rarely happens compared to the latter and thus produces the information gap. We evaluate the sensationalism degree of a news headline from the following aspects.

- *Sentiment.* Extreme sentiment expressed in a news headline is assumed to indicate a higher degree of sensationalism. Hence, we measure the frequencies of positive words and negative words within a news headline by using LIWC, as well as its sentiment polarity by computing the average sentiment scores[7] of the words it contains.
- *Punctuations*: Some punctuations can help to express sensationalism or extreme sentiments, e.g., ellipses ('...'), question ('?') and exclamation marks ('!'). Hence the frequencies of these three are also counted when representing news headlines.
- *Similarity.* Similarity between the headline of a news article and its body-text is assumed to be positively correlated to the degree of *relative* sensationalism expressed in the news headline [5]. Capturing such similarity requires firstly embedding the headline and body-text for each news article into the same space. To achieve this goal, we respectively utilize WORD2VEC [27] model at the word-level and train SENTENCE2VEC [2] model at the sentence-level, considering that one headline often refers to one sentence. For the headline or body-text containing more than one words or sentences, we compute the average of its word embeddings (i.e., vectors) or sentence embeddings. The similarity between a news headline and its body-text then can be computed based on various similarity measures, where we use cosine distance in experiments.

D. News-worthiness. While clickbaits can attract eyeballs they are rarely newsworthy with (I) low quality and (II) high informality [32]. We capture both characteristics in news:

- *I. Quality*: The title of high quality news articles is often a *summary* of the whole news event described in body-text [5]. To capture this property, one can assess the similarity between the headline of a news article and its body-text, which has been already captured when analyzing sensationalism. Secondly, such titles should be a *simplified* summary of the whole news event described in body-text, where meaningful words should occupy its main proportion [7]. From

[7] https://www.nltk.org/api/nltk.sentiment.html

this perspective, the frequencies of content words, function words, and stop words within each news headline are counted and included as features.

- *II. Informality*: LIWC [33] provides five dimensions to evaluate such informality of language: (1) *swear words* (e.g., 'damn'); (2) *netspeaks* (e.g., 'btw' and 'lol'); (3) *assents* (e.g., 'OK'); (4) *nonfluencies* (e.g., 'er', 'hm', and 'umm'); and (5) *fillers* (e.g., 'I mean' and 'you know'). Hence, we measure the informality for each news headline by investigating its word or phrase frequencies within every dimension and include them as features.

3.3.2 DisInformation-related Attributes (DIAs). Deception/disinformation is a more general concept compared to fake news, which additionally includes fake statements, fake reviews, and the like. [60]. Thus, we aim to extract a set of features inspired from patterns of deception/disinformation to represent news content, which also provides an opportunity to empirically study the relationships between fake news and deception/disinformation. Such patterns, often explained by fundamental theories initially developed in forensic psychology, are with respect to:

Quality: In addition to writing style, *Undeutsch hypothesis* [48] states that a fake statement also differs in quality from a true one. Here, we evaluate news quality from three perspectives:

- *Informality*: Basically, the quality of a news article should be negatively correlated to its informality. As having been specified, LIWC [33] provides five dimensions to evaluate the informality of language. Here, we investigate the word or phrase numbers (proportions) on each dimension within news content (as apposed to headline) and include them as features.
- *Diversity*: At a higher level, such quality can be assessed by investigating the writing and expression ability of news author(s), those of whom with a higher writing ability often possess a greater reserve of vocabularies. Thus, the number (proportion) of unique words, content words, nouns, verbs, adjectives and adverbs being used in news content are computed and included as features to evaluate the quality of news content.
- *Subjectivity*: When a news article becomes hyperpartisan and biased, its quality should also be considered to be lower compared with those that maintain objectivity [36]. Benefiting from the work done by Recasens et al. [39], which provides the corpus of biased lexicons, here we evaluate the subjectivity of news articles by counting their number (proportion) of biased words. On the other hand, factive verbs (e.g., 'observe') [17] and report verbs (e.g., 'announce') [39], as the opposite of biased ones, their numbers (proportions) are also included in our feature set, which are negatively correlated to content subjectivity.

Sentiment: Sentiment expressed within news content is suggested to be different within fake news and true news [61]. Here, we evaluate such sentiments for each news article by measuring the number (proportion) of positive words and negative words, as well as its sentiment polarity.

Quantity: *Information manipulation theory* [26] reveals that extreme information quantity (too much or too little) often exists in deception. We assess such quantity for each news article at character-level, word-level, sentence-level and paragraph-level, respectively, i.e., the overall number of characters, words, sentences and paragraphs; and the average number of characters per word, words per sentence, sentences per paragraph.

Specificity: Fictitious stories often lack the details of cognitive and perceptual processes, as indicated by *reality monitoring* [22] and *four-factor* [61] theories. Based on LIWC dictionary [33], for *cognitive processes*, we investigate the frequencies of terms related to (1) *insight* (e.g., 'think'), (2) *causation* (e.g., 'because'), (3) *discrepancy* (e.g., 'should'), (4) *tentative language* (e.g., 'perhaps'), (5) *certainty* (e.g., 'always') and (6) *differentiation* (e.g., 'but' and 'else'); for *perceptual processes*, we investigate the frequencies of terms referring to vision, hearing, and feeling.

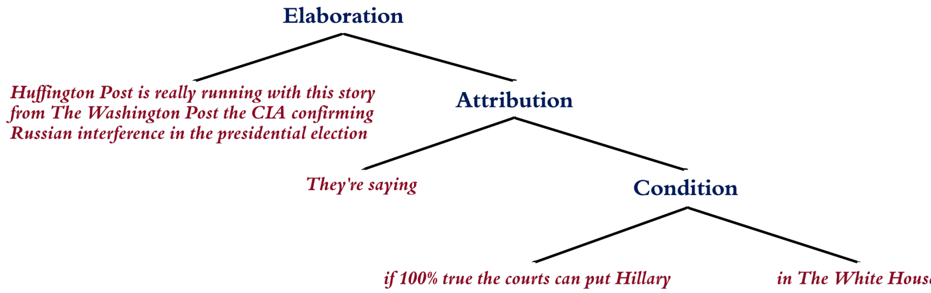


Fig. 3. Rhetorical Structure for the partial content "Huffington Post is really running with this story from The Washington Post about the CIA confirming Russian interference in the presidential election. They're saying if 100% true, the courts can PUT HILLARY IN THE WHITE HOUSE!" within a fake news article. Here, one elaboration, attribution and condition rhetorical relationships exist.

3.4 Discourse-level

Style features at discourse-level investigate the (relative/standardized) frequencies of rhetorical relationships among sentences within a news article. Such rhetorical relationships can be obtained through a RST parser[8] [19], where an illustration is provided in Figure 3.

We have detailed how each news article can be represented across language levels with theory-driven computational features. These features then can be utilized by a supervised learning framework, e.g., Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forests (RF) and XGBoost [9], for fake news prediction.

4 EXPERIMENTS

We conduct empirical studies to evaluate the proposed model, where experimental setup is detailed in Section 4.1, and the performance is presented and evaluated in Section 4.2.

4.1 Experimental Setup

Real-world datasets used in our experiments are specified in Section 4.1.1 followed by baselines that our model will be compared with in Section 4.1.2.

4.1.1 Datasets. Our experiments are conducted on two well-established public benchmark datasets of fake news detection[9] [46]. News articles in these datasets are collected from PolitiFact and BuzzFeed, respectively. Ground truth labels (*fake* or *true*) of news articles in both datasets are provided by fact-checking experts, which guarantees the quality of news labels (*fake* or *true*). In addition to news content and labels, both datasets also provide massive information on social network of users involved in spreading true/fake news on Twitter which contains (1) users and their following/follower relationships (*user-user relationships*) and (2) how the news has been propagated (tweeted/re-tweeted) by Twitter users, i.e., *news-user relationships*. Such information is valuable for our comparative studies. Statistics of two datasets are provided in Table 1. Note that the original datasets are balanced with 50% true news and 50% fake news. As few reference studies have provided the actual ratio between true news and fake news, we design an experiment in Section 4.2.5 to evaluate our work within unbalanced datasets by controlling this ratio.

[8]https://github.com/jiyfeng/DPLP
[9]https://www.dropbox.com/s/gho59cezl43sov8/FakeNewsNet-master.zip?dl=0

Table 1. Data Statistics

Data	PolitiFact	BuzzFeed
# Users	23,865	15,257
# News–Users	32,791	22,779
# Users–Users	574,744	634,750
# News Stories	240	180
# True News	120	90
# Fake News	120	90

4.1.2 Baselines. We compare the performance of the proposed method with several state-of-the-art fake news detection methods on the same datasets. These methods detect fake news by (1) analyzing news content (i.e., content-based fake news detection) [34], or (2) exploring news dissemination on social networks (i.e., propagation-based fake news detection) [6], or (3) utilizing both information within news content and news propagation information [46].

I. Pérez-Rosas et al. [34] propose a comprehensive linguistic model for fake news detection, involving the following features: (i) n-grams (i.e., uni-grams and bi-grams) and (ii) CFGs based on TF-IDF encoding; (iii) word and phrase proportions referring to all categories provided by LIWC; and (iv) readability. Features are computed and used to predict fake news within a supervised machine learning framework.

II. Castillo et al. [6] design features to exploit information from user profiles, tweets and propagation trees to evaluate news credibility within a supervised learning framework. Specifically, these features are based on (i) quantity, sentiment, hash-tag and URL information from user tweets, (ii) user profiles such as registration age, (iii) news topics through mining tweets of users, and (iv) propagation trees (e.g., the number of propagation trees for each news topic).

III. Shu et al. [46] detect fake news by exploring and embedding the relationships among news articles, publishers and spreaders on social media. Specifically, such embedding involves (i) news content by using non-negative matrix factorization, (ii) users on social media, (iii) news-user relationships (i.e., user engagements in spreading news articles), and (iv) news-publisher relationships (i.e., publisher engagements in publishing news articles). Fake news detection is then conducted within a semi-supervised machine learning framework.

Additionally, fake news detection based on latent representation of news articles is also investigated in comparative studies. Compared to style features, such latent ones are less explainable but have been empirically shown to be remarkably useful [31, 51]. Here we consider as baselines supervised classifiers with the input of **(IV) word2vec** [27] and **(V) Doc2Vec** [23] embedding of news articles.

4.2 Performance Evaluation

In our experiments, several supervised classifiers have been used, among which SVM (with linear kernel), Random Forest (RF) and XGBoost[10] [9] perform best compared to the others (e.g., LR, Logistic Regression and NB, Naïve Bayes) within both our model and baselines. The performance of experiments are provided in terms of accuracy, precision, recall and F_1 scores based on 5-fold cross-validation. In this section, we first present and evaluate the general performance of the proposed model by comparing it with baselines in Section 4.2.1. As news content is represented at

[10] https://github.com/dmlc/XGBoost

Table 2. General Performance of Fake News Detection Models[10]. Among the baselines, (1) the propagation-based model ([6]) can perform relatively well compared to content-based ones ([23, 27, 34]); and (2) the hybrid model ([46]) can outperform both types of techniques. Compared to the baselines, (3) our model [slightly] outperforms the hybrid model and can outperform the others in predicting fake news.

Method	PolitiFact				BuzzFeed			
	Acc.	**Pre.**	**Rec.**	**F_1**	**Acc.**	**Pre.**	**Rec.**	**F_1**
Perez-Rosas et al. [34]	.811	.808	.814	.811	.755	.745	.769	.757
n-grams+TF-IDF	.755	.756	.754	.755	.721	.711	.735	.723
CFG+TF-IDF	.749	.753	.743	.748	.735	.738	.732	.735
LIWC	.645	.649	.645	.647	.655	.655	.663	.659
Readability	.605	.609	.601	.605	.643	.651	.635	.643
word2vec [27]	.688	.671	.663	.667	.703	.714	.722	.718
Doc2Vec [23]	.698	.684	.712	.698	.615	.610	.620	.615
Castillo et al. [6]	.794	.764	.889	.822	.789	.815	.774	.794
Shu et al. [46]	.878	.867	.893	.880	.864	.849	.893	.870
Our Model	**.892**	**.877**	**.908**	**.892**	**.879**	**.857**	**.902**	**.879**

the lexicon, syntax, semantic and discourse levels, we evaluate the performance of the model within and across different levels in Section 4.2.2. The detailed analysis at the semantic-level follows, which provides opportunities to investigate the potential and understandable patterns of fake news, as well as its relationships with deception/disinformation (Section 4.2.3) and clickbaits (Section 4.2.4). Next, we assess the impact of news distribution on the proposed model in Section 4.2.5. Finally, we investigate the performance of the proposed method for fake news early detection in Section 4.2.6.

4.2.1 General Performance in Predicting Fake News. Here, we provide the general performance of the proposed model in predicting fake and compare it with baselines. Results are presented in Table 2, which indicate that among baselines, (1) the propagation-based fake news detection model ([6]) can perform comparatively well compared to content-based ones ([23, 27, 34]); and (2) the hybrid model ([46]) can outperform fake news detection models that use either news content or propagation information. Compared to the baselines, (3) our model [slightly] outperforms the hybrid model in predicting fake news, while not relying on propagation information. For fairness of comparison, we report the best performance of the methods that rely on supervised classifiers by using SVM, RF, XGBoost, LR and NB.

4.2.2 Fake News Analysis Across Language Levels. As being specified in Section 3, features representing news content are extracted at lexicon-level, syntax-level, semantic-level and discourse-level. We first evaluate the performance of such features within or across language levels in predicting fake news in (**E1**), followed by feature importance analysis at each level in (**E2**).

E1: *Feature Performance Across Language Levels.* Table 3 presents the performance of features within each level and across levels for fake news detection. Results indicate that within single level, (1) features at lexicon-level (BOWs) and deep syntax-level (CFGs) outperform the others, which can achieve above 80% accuracy rate and F_1 score, where (2) the performance of features at semantic-level (DIAs and CBAs) and shallow syntax-level (POS tags) follows with an accuracy and F_1 score above 70% while below 80%. However, (3) fake news prediction by using the standardized frequencies of rhetorical relationships (discourse-level) do not perform well within the framework. It should be noted that the number of features based on BOWs and CFGs is in the order of a thousand,

[10]For each dataset, the maximum value is underlined, that in each column is bold, and that in each row is colored in gray.

Table 3. Feature Performance across Language Levels[10]. Lexicon-level and deep syntax-level features outperform the others, where the performance of semantic-level and shallow syntax-level ones follows. When combining features (exclude RRs) across levels, it enhances the performance compared to when separately using them in predicting fake news.

	Language Level	Feature Group	PolitiFact				BuzzFeed			
			XGBoost		RF		XGBoost		RF	
			Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Within Levels	**Lexicon**	BOW	.856	.858	.837	.836	.823	.823	.815	.815
	Shallow Syntax	POS	.755	.755	.776	.776	.745	.745	.732	.732
	Deep Syntax	CFG	.877	.877	.836	.836	.778	.778	.845	.845
	Semantic	DIA+CBA	.745	.748	737	.737	.722	.750	.789	.789
	Discourse	RR	.621	.621	.633	.633	.658	.658	.665	.665
Across Two Levels	**Lexicon+Syntax**	BOW+POS+CFG	.858	.860	.822	.822	.845	.845	**.871**	**.871**
	Lexicon+Semantic	BOW+DIA+CBA	.847	.820	.839	.839	.844	.847	.844	.844
	Lexicon+Discourse	BOW+RR	.877	.877	.880	.880	.872	.873	.841	.841
	Syntax+Semantic	POS+CFG+DIA+CBA	.879	.880	.827	.827	.817	.823	.844	.844
	Syntax+Discourse	POS+CFG+RR	.858	.858	.813	.813	.817	.823	.844	.844
	Semantic+Discourse	DIA+CBA+RR	.855	.857	.864	.864	.844	.841	.847	.847
Across Three Levels	**All-Lexicon**	All-BOW	.870	.870	.871	.871	.851	.844	.856	.856
	All-Syntax	All-POS-CFG	.834	.834	.822	.822	.844	.844	.822	.822
	All-Semantic	All-DIA-CBA	.868	.868	.852	.852	.848	.847	.866	.866
	All-Discourse	All-RR	**.892**	**.892**	**.887**	**.887**	**.879**	**.879**	.868	.868
		Overall	.865	.865	.845	.845	.855	.856	.854	.854

much more than the others that are within the order of a hundred; and (4) when combining features (exclude RRs) across levels, it enhances the performance compared to when separately using features within each level in predicting fake news. Such performance can achieve ~88% to ~89% accuracy and F_1 score. In addition, it can be observed from Table 2 and Table 3 that though the assessment of semantic-level features (DIAs and CBAs) that we defined and selected based on psychological theories rely on LIWC, their performance in predicting fake news is better than directly utilizing all word and phrase categories provided by LIWC without supportive theories.

E2: *Feature Importance Analysis.* RF (mean decrease impurity) is used to determine the importance of features, among which the top discriminating lexicons, POS tags, rewrite rules and RRs are provided in Table 4. It can be seen that (1) discriminating lexicons differ from one dataset to the other; (2) compared to the other POS tags, the standardized frequencies of POS (possessive ending), VBN (verb in a form of past participle) and JJ (adjective) are more powerful in differentiating fake news from true news in two datasets; (3) unsurprisingly, discriminating rewrite rules are often formed based on discriminating lexicons and POS tags, e.g., JJ → 'presidential' and ADVP (adverb phrase) → RB (adverb) NP (noun phrase); (4) compared to the other RRs, nucleus that contains basic information about parts of text and same_unit that indicates the relation between discontinuous clauses play a comparatively significant role in predicting fake news. It should be noted that though these features can capture news content style and perform well, they are not as easy to be understood as semantic-level features. Consider that, detailed analysis for DIAs (Section 4.2.3) and CBAs (Section 4.2.4) is conducted next.

4.2.3 Deceptions and Fake News. As discussed in Section 3.3.2, well-established forensic psychology theories on identifying deception/disinformation have inspired us to represent news content through measuring its [psycho-linguistic] attributes, e.g., sentiment. Such potential clues provided by these theories help reveal fake news patterns that are easy to understand. Opportunities are also provided

Table 4. Important Lexicon-level, Syntax-level and Discourse-level Features for Fake News Detection.

(a) Lexicons

Rank	PolitiFact	BuzzFeed
1	'nominee'	'said'
2	'continued'	'authors'
3	'story'	'university'
4	'authors'	'monday'
5	'hillary'	'one'
6	'presidential'	'trump'
7	'highlight'	'york'
8	'debate'	'daily'
9	'cnn'	'read'
10	'republican'	'donald'

(c) Rewrite Rules

Rank	PotiliFact	BuzzFeed
1	NN → 'story'	VBD → 'said'
2	NP → NP NN	ADVP → RB NP
3	VBD → 'said'	RB → 'hillary'
4	ROOT → S	NN → 'university'
5	POS → ''s'	NNP → 'monday'
6	NN → 'republican'	VP → VBD NP NP
7	NN → 'york'	NP → NNP
8	NN → 'nominee'	VP → VB NP ADVP
9	JJ → 'hillary'	S → ADVP VP
10	JJ → 'presidential'	NP → JJ

(b) POS Tags

Rank	PolitiFact	BuzzFeed
1	POS	NN
2	JJ	VBN
3	VBN	POS
4	IN	JJ
5	VBD	RB

(d) RRs

Rank	PolitiFact	BuzzFeed
1	nucleus	attribution
2	attribution	nucleus
3	textualorganization	satellite
4	elaboration	span
5	same_unit	same_unit

to compare deceptions/disinformation and fake news; theoretically, deception/disinformation is a more general concept compared to fake news, which additionally includes fake statements, fake reviews, and the like. In this section, we first evaluate the performance of these disinformation-related attributes (i.e., DIAs) in predicting fake news in (**E1**). Then in (**E2**), important features and attributes are identified, followed by a detailed feature analysis to reveal the potential patterns of fake news and compare them with that of deception (**E3**).

E1: *Performance of Disinformation-related Attributes in Predicting Fake News.* Table 5 presents the performance of disinformation-related attributes in predicting fake news. Results indicate that identifying fake news articles respectively based on their content quality, sentiment, quantity, and specificity performs similarly, with 60% (50%) to 70% (60%) accuracy and F_1 score using PolitiFact (BuzzFeed) data. Combining all attributes to detect fake news performs better than separately using each type of attribute, which can achieve 70% (60%) to 80% (70%) accuracy and F_1 score using PolitiFact (BuzzFeed) data.

E2: *Importance Analysis for Disinformation-related Features and Attributes.* RF (mean decrease impurity) is used to determine the importance of features, among which the top ten discriminating features are presented in Table 6. Results indicate that, in general, (1) content quality (i.e., informality, subjectivity and diversity), sentiments expressed, quantity and specificity (i.e., cognitive and perceptual process) all play a role in differentiating fake news articles from the true ones. Specifically, in both datasets, (2) fake news differs more significantly in diversity and quantity from the truth compared to the other attributes, where (3) cognitive process involved in news content and content subjectivity follow. (4) Content informality and sentiments play a comparatively weak role in predicting fake news compared to the others.

Table 5. Performance of Disinformation-related Attributes in Predicting Fake News[10]. Individual attributes perform similarly while combining all attributes perform better in predicting fake news.

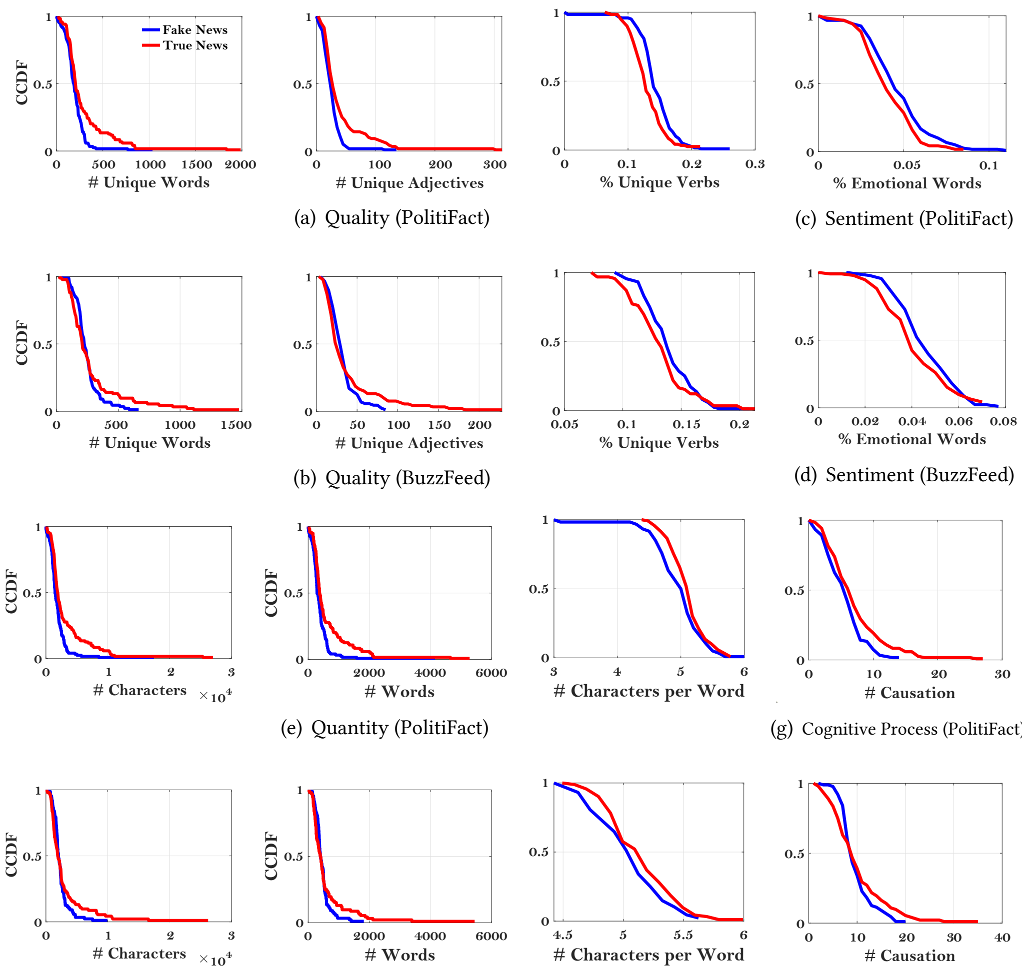
Disinformation-related Attribute(s)	**PolitiFact**				**BuzzFeed**			
	XGBoost		**RF**		**XGBoost**		**RF**	
	Acc.	**F_1**	**Acc.**	**F_1**	**Acc.**	**F_1**	**Acc.**	**F_1**
Quality	.667	.652	.645	.645	.556	.500	.512	.512
– Informality	.688	.727	.604	.604	.555	.513	.508	.508
– Subjectivity	.688	.706	.654	.654	.611	.588	.533	.530
– Diversity	.583	.600	.620	.620	.639	.552	.544	.544
Sentiment	.625	.591	.583	.583	.556	.579	.515	.525
Quantity	.583	.524	.638	.638	.528	.514	.584	.586
Specificity	.625	.609	.558	.558	.583	.571	.611	.611
– Cognitive Process	.604	.612	.565	.565	.556	.579	.531	.531
– Perceptual Process	.563	.571	.612	.612	.556	.600	.571	.571
Overall	**.729**	**.735**	**<u>.755</u>**	**<u>.755</u>**	**<u>.667</u>**	**<u>.647</u>**	**.625**	**.625**

Table 6. Important Disinformation-related Features and Attributes for Fake News Detection. In both datasets, content diversity and quantity are most significant in differentiating fake news from the truth; cognitive process involved and content subjectivity are second; content informality and sentiments expressed are third.

Rank	**PolitiFact**		**BuzzFeed**	
	Feature	**Attribute**	**Feature**	**Attribute**
1	# Characters per Word	Quantity	# Overall Informal Words	Informality
2	# Sentences per Paragraph	Quantity	% Unique Words	Diversity
3	% Positive Words	Sentiment	% Unique Nouns	Diversity
4	% Unique Words	Diversity	% Unique Content Words	Diversity
5	% Causation	Cognitive Process	# Report Verbs	Subjectivity
6	# Words per Sentence	Quantity	% Insight	Cognitive Process
7	% Report Verbs	Subjectivity	% Netspeak	Informality
8	% Unique Verbs	Diversity	# Sentences	Quantity
9	# Sentences	Quantity	% Unique Verbs	Diversity
10	% Certainty Words	Cognitive Process	% Unique Adverbs	Diversity

E3: *Potential Patterns of Fake News.* Based on Complementary Cumulative Distribution Function (CCDF) [50], we analyze each feature to identify common patterns of fake news across both datasets. Results are illustrated in Figure 4. Note that each feature variable presented in Figure 4 meets the following requirements: in both datasets, (i) its distribution in fake news is different from that in true news; such differences (ii) can reveal various characteristics of fake news, e.g., fake news often has less unique words compared to true news across two datasets, and (iii) is significant with p-value less than 0.1 in two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test. Specifically, we have the following observations:

- Similar to deception, fake news differs in content quality and sentiments expressed from the truth [48, 61]. Compared to true news, fake news often carries (i) less unique words, adjectives and adverbs (the CCDFs of the number of unique adverbs are not presented in Figure 4 due to space limitation); while (ii) a greater proportion of unique verbs; and (iii) a greater proportion of emotional (positive+negative) words (see Figure 4(c) and 4(d)).

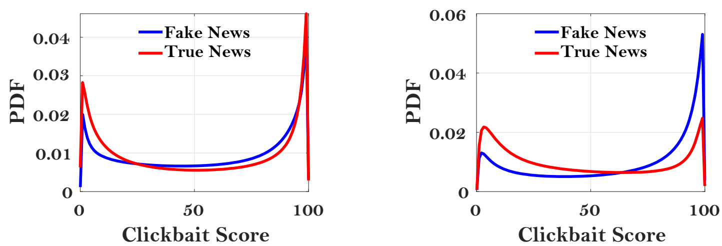


(a) Quality (PolitiFact)
(b) Quality (BuzzFeed)
(c) Sentiment (PolitiFact)
(d) Sentiment (BuzzFeed)
(e) Quantity (PolitiFact)
(f) Quantity (BuzzFeed)
(g) Cognitive Process (PolitiFact)
(h) Cognitive Process (BuzzFeed)

Fig. 4. Potential Patterns of Fake News. Compared to true news, fake news exhibits the following characteristics in both datasets: (a-b) lower diversity of words and adjectives while higher verb diversity; (c-d) a greater proportion of emotional words; (e-f) lower quantities of characters, words, and characters per word; and (g-h) a lower level of cognitive information.

- Compared to true news articles, fake news articles are characterized by (i) shorter words; and (ii) lower quantities of characters and words (see Figure 4(e) and 4(f)).
- It is known that deception often does not involve cognitive and perceptual processes [22, 61]. Consistent with this discovery, in general, lexicons related to cognitive processes, e.g., causation words (see Figure 4(g) and 4(h)), appear less frequently in fake news articles compared to the true ones. The frequencies of lexicons related to perceptual processes, however, can hardly discriminate between fake and true news stories.

4.2.4 Clickbaits and Fake News. We also explore the relationship between clickbaits and fake news by conducting four experiments: (**E1**) analyzes clickbait distribution within fake and true news



(a) PolitiFact (b) BuzzFeed

Fig. 5. Clickbait Distribution within Fake and True News Articles. Clickbaits are more common in fake news articles compared to true news articles: among news headlines with relatively low (high) clickbait scores, e.g., ≤ 50 (≥ 50), true (fake) news articles often occupy a greater proportion compared to fake (true) news articles.

Table 7. Performance of Clickbait-related Attributes in Predicting Fake News[10]. Based on the experimental setup, news-worthiness of headlines outperforms the other attributes in predicting fake news.

Clickbait-related Attributes	PolitiFact				BuzzFeed			
	XGBoost		RF		XGBoost		RF	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Readability	.708	.682	.636	.636	.529	.529	.528	.514
Sensationalism	.563	.571	.653	.653	.581	.581	.694	.645
News-worthiness	**.729**	**.711**	**.683**	**.683**	**.686**	**.686**	.694	.667
Overall	.604	.612	.652	.652	.638	.628	**.705**	**.705**

articles; (**E2**) evaluates the performance of clickbait-related attributes in predicting fake news, among which important features and attributes are identified in (**E3**); and (**E4**) examines if clickbait and fake news share some potential patterns.

E1: *Clickbait Distribution within Fake and True News Articles.* As few datasets, including PolitiFact and BuzzFeed, provide both news labels (*fake* or *true*) and news headline labels (*clickbait* or *regular headline*), we use a pretrained deep net, particularly, a Convolutional Neural Network (CNN) model[11] [1] to obtain the clickbait scores ($\in [0, 100]$) of news headlines, where 0 indicates not-clickbait (i.e., a regular headline) and 100 indicates clickbait. The model can achieve ~93.8% accuracy [1]. Using clickbait scores we obtain the clickbait distribution (i.e., Probabilistic Density Function, PDF) respectively within fake and true news articles, which is depicted in Figure 5. We observe that clickbaits have a closer relationship with fake news compared to true news: among news headlines with relatively low (high) clickbait scores, e.g., ≤ 50 (≥ 50), true (fake) news articles often occupy a greater proportion compared to fake (true) news articles.

E2: *Performance of Clickbait-related Attributes in Predicting Fake News.* Table 7 presents the performance of clickbait-related attributes in predicting fake news. Results indicate that identifying fake news articles based on their headline news-worthiness, whose accuracy and F_1 score are around 70%, performs better than based on either headline readability or sensationalism.

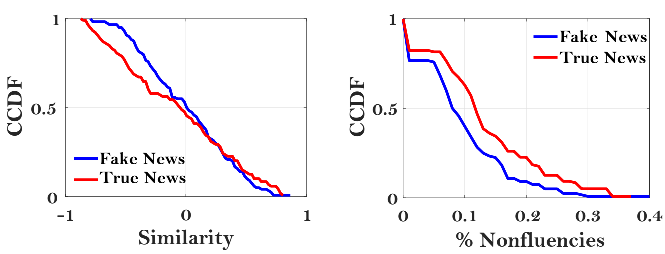
E3: *Importance Analysis for Clickbait-related Features and Attributes.* Random forest is used to identify most important features, among which the top five features are presented in Table 8.

[11] https://github.com/saurabhmathur96/clickbait-detector

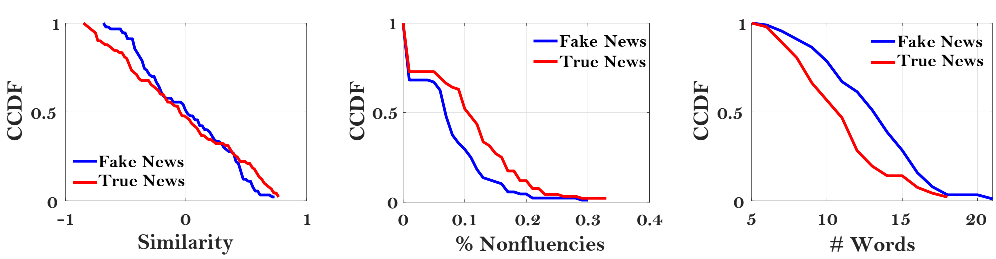
Table 8. Important Clickbait-related Features and Attributes for Fake News Detection.

Rank	**PolitiFact**		**BuzzFeed**	
	Feature	**Attribute**	**Feature**	**Attribute**
1	Similarity (WORD2VEC)	S/N	Similarity (WORD2VEC)	S/N
2	Similarity (SENTENCE2VEC)	S/N	# Characters	R
3	% Netspeak	N	# Words	R
4	Sentiment Polarity	S	# Syllables	R
5	Coleman-Liau Index	R	Gunning-Fog Index	R

R: Readability; S: Sensationalism; N: News-worthiness



(a) PolitiFact

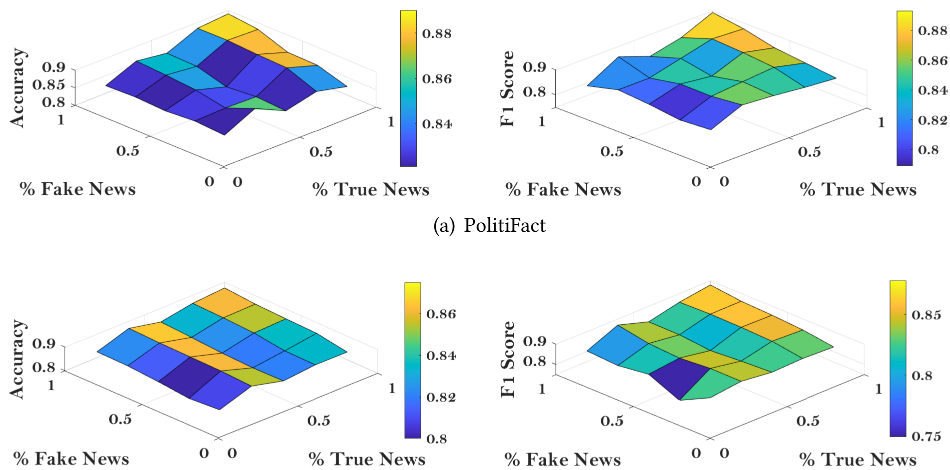


(b) BuzzFeed

Fig. 6. Potential Patterns of Fake News Headlines. **(Left)**: Similarity between Headline of a News Article and its Body-text. Figures indicate that, in general, fake news headlines are less similar to their body-text when compared to true news. **(Middle)**: Informality of News Headline. Figures indicate that, in general, nonfluencies occupy a less proportion in fake news headlines compared to true news haeadlines. **(Right)**: The Number of Words within News Headlines. Figures indicate that fake news headlines generally contain more words compared to true news headlines.

Results indicate that (1) headline readability, sensationalism and news-worthiness all play a role in differentiating fake news articles from the true ones; and (2) consistent with their performance in predicting fake news, features measuring news-worthiness of headlines rank relatively higher compared to that assessing headline readability and sensationalism.

E4: *Potential Patterns of Fake News Headlines.* Using the CCDFs of clickbait features within fake and true news, we examine whether fake news headlines share some potential patterns with clickbaits. Results are provided in Figure 6, where the values of each feature variable obtained from true news



(a) PolitiFact

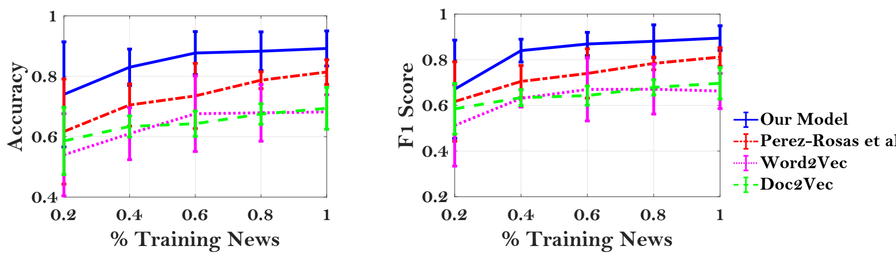
(b) BuzzFeed

Fig. 7. Performance Sensitivity to News Distribution (% Fake News vs. % True News)

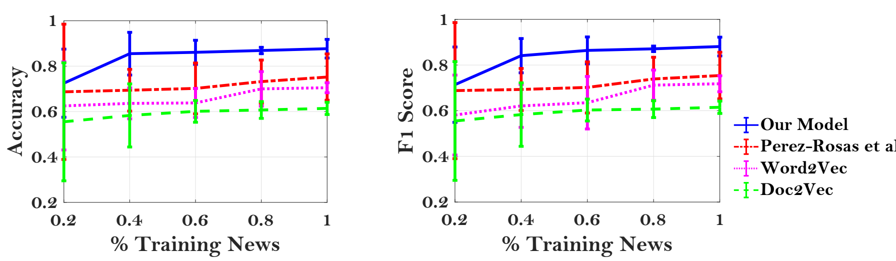
and fake news are not drawn from the same underlying continuous population with p-value less than 0.1 in two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test. Specifically,

- Figures on the left column present the CCDF of the similarity between news headlines and their corresponding body-text, which is computed using the SENTENCE2VEC model [2]. Such similarity is assumed to be positively correlated to the sensationalism and negatively correlated to the news-worthiness of news headlines. Both figures reveal that, in general, fake news headlines are less similar to their body-text compared to true news headlines, which matches with the characteristic of clickbaits [5].
- Figures on the middle column present the CCDF of the proportion of nonfluencies (e.g., 'hm') for fake and true news headlines, which is one of the features measuring the informality (as well as news-worthiness) of news headlines. Unexpectedly, we observe from both figures that nonfluencies (as well as netspeak) often occupy a smaller proportion in fake news headlines compared to true news headlines, which is inconsistent with the characteristic of clickbaits [37].
- Figures on the right column present the CCDF of the number of words within news headlines, as one of the parameters of readability criteria and features representing news readability. Though it cannot directly measure the readability of news headlines, we find that fake news headlines often contain more words (as well as syllables and characters) compared to true news. An interesting phenomenon that can be observed from Figure 4 and Figure 6 is that compared to true news, fake news is characterized by a longer headline yet a shorter body-text.

4.2.5 Impact of News Distribution on Fake News Detection. We assess the sensitivity of our model to the news distribution, i.e., the proportion of true vs. fake news stories within the population, which are initially equal in both PolitiFact and BuzzFeed datasets. Specifically, we randomly select a proportion ($\in (0, 1]$) of fake news stories and a proportion of true news stories in each dataset. The corresponding accuracy and F_1 scores by using XGBoost are presented in Figure 7. Results on both datasets indicate that the performance of the proposed model fluctuates between $\sim$0.75 and $\sim$0.9. However, in most cases, the model is resilient to such perturbations and the accuracy and F_1 scores are between $\sim$0.8 and $\sim$0.88.



(a) PolitiFact



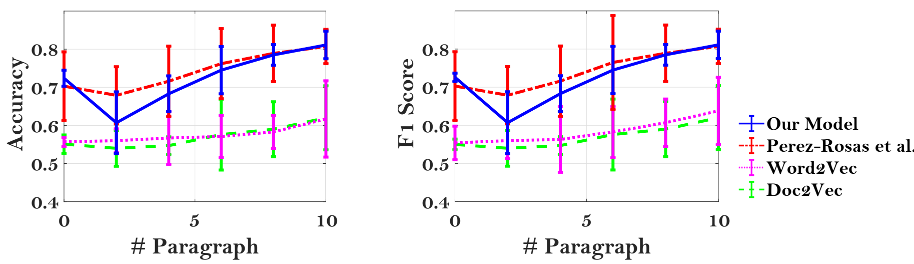
(b) BuzzFeed

Fig. 8. Impact of the Number of Training News Articles in Predicting Fake News.

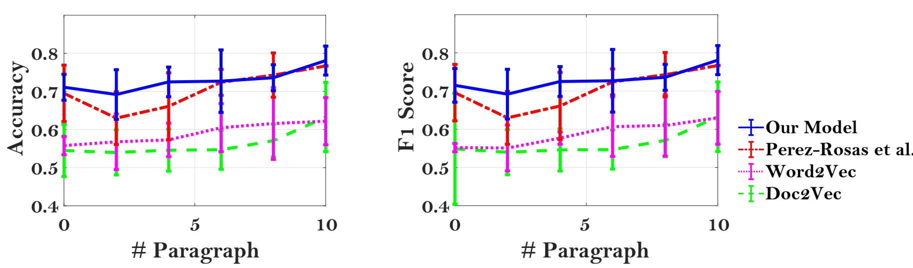
4.2.6 Fake News Early Detection. Compared to propagation-based models, content-based fake news detection models can detect fake news before it has been disseminated on social media. Among content-based fake news detection models, their early detection ability also depends on how much prior knowledge they require to accurately detect fake news [52, 58]. Here, we measure the number of such prior knowledge from two perspectives: (**E1**) the number of news articles available for learning and training a classifier, and (**E2**) the content for each news article available for training and predicting fake news.

E1: *Model Performance with Limited Number of Training News Articles.* In this experiment, we randomly select a proportion ($\in (0, 1]$) of news articles from each of the PolitiFact and BuzzFeed datasets. Performance of several content-based models in predicting fake news is then evaluated based on the selected subset of news articles, which has been presented in Figure 8. It can be observed from Figure 8 that with the change of the number of available training news articles, the proposed model performs best in most cases. Note that, compared to random sampling, sampling based on the time that news articles were published is a more proper strategy when evaluating the early detection ability of models; however, such temporal information has not been fully provided in the datasets.

E2: *Model Performance with Limited News Content Information.* In this experiment, we assess the performance of our fake news model when partial news content information is available. Specifically, such partial news content information ranges from the headline of the news article to the headline with n ($n = 1, 2, \cdots$) randomly selected paragraph(s) from the article. Results are presented in Figure 9, which indicate that (1) compared to the linguistic model proposed by Perez-Rosas et al. [34], our model generally has a comparable performance while can always outperform it when



(a) PolitiFact



(b) BuzzFeed

Fig. 9. Impact of the Available Information within News Content in Predicting Fake News.

only news headline information is available (i.e., # paragraphs is 0); and (2) our model can always perform better than the models based on the latent representation of news content [23, 27].

5 CONCLUSION

In this paper, a theory-driven model is proposed for fake news early detection. To predict fake news before it has been propagated on social media, the proposed model comprehensively studies and represents news content at four language levels: lexicon-level, syntax-level, semantic-level, and discourse-level. Such representation is inspired by well-established theories in social and forensic psychology. Experimental results based on real-world datasets indicate that the performance (i.e., accuracy and F_1 score) of the proposed model can (1) generally achieve ~88%, outperforming all baselines which include content-based, propagation-based and hybrid (content+propagation) fake news detection models; and (2) maintain ~80% and ~88% when data size and news distribution (% fake news vs. % true news) vary. Among content-based models, we observe that (3) the proposed model performs comparatively well in predicting fake news with limited prior knowledge. We also observe that (4) similar to deception, fake news differs in content quality and sentiment from the truth, carries poorer cognitive information while carries similar levels of perceptual information compared to the truth. (5) Similar to clickbaits, fake news headlines present higher sensationalism while their readability and news-worthiness characteristics are complex and difficult to be directly concluded. In addition, fake news (6) is often matched with shorter words and (7) often contains more characters and words in headlines while less in body-text. It should be pointed out that (1) effective utilization of rhetorical relationships and (2) utilizing news images in an interpretable way for fake news detection are still open issues, which will be part of our future work.

A SEMANTIC-LEVEL FEATURES

Table 9 provides a detailed list of semantic-level features involved in our study.

Table 9. Semantic-level Features

		Attribute	Feature(s)	Tool & Ref.
Disinformation-related Attributes (DIAs) (72)	Quality (30)	Informality (12)	#/% Swear Words	LIWC
			#/% Netspeak	
			#/% Assent	
			#/% Nonfluencies	
			#/% Fillers	
			Overall #/% Informal Words	
		Diversity (12)	#/% Unique Words	Self-implemented
			#/% Unique Content Words	LIWC
			#/% Unique Nouns	NLTK POS Tagger
			#/% Unique Verbs	
			#/% Unique Adjectives	
			#/% Unique Adverbs	
		Subjectivity (6)	#/% Biased Lexicons	[39]
			#/% Report Verbs	
			#/% Factive Verbs	[17]
	Sentiment (13)		#/% Positive Words	LIWC
			#/% Negative Words	
			#/% Anxiety Words	
			#/% Anger Words	
			#/% Sadness Words	
			Overall #/% Emotional Words	
			Avg. Sentiment Score of Words	NLTK.Sentiment Package
	Quantity (7)		# Characters	Self-implemented
			# Words	Self-implemented
			# Sentences	Self-implemented
			# Paragraphs	Self-implemented
			Avg. # Characters Per Word	Self-implemented
			Avg. # Words Per Sentence	Self-implemented
			Avg. # Sentences Per Paragraph	Self-implemented
	Specificity (22)	Cognitive Process (14)	#/% Insight	LIWC
			#/% Causation	
			#/% Discrepancy	
			#/% Tentative	
			#/% Certainty	
			#/% Differentiation	
			Overall #/% Cognitive Processes	
		Perceptual Process (8)	#/% See	
			#/% Hear	
			#/% Feel	
			Overall #/% Perceptual Processes	
	General Clickbait Patterns (3)		# Common Clickbait Phrases	[14]
			# Common Clickbait Expressions	
			Overall # Common Clickbait Patterns	
			Flesch Reading Ease Index (FREI)	Self-implemented
			Flesch-Kioncaid Grade Level (FKGL)	Self-implemented

Clickbait-related Attributes (CBAs) (44)		Readability (10)	Automated Readability Index (ARI)	Self-implemented
			Gunning Fox Index (GFI)	Self-implemented
			Coleman-Liau Index (CLI)	Self-implemented
			# Words	Self-implemented
			# Syllables	Self-implemented
			# Polysyllables	Self-implemented
			# Characters	Self-implemented
			# Long Words	Self-implemented
	Sensationalism (13)	Sentiments (7)	#/% Positive Words	LIWC
			#/% Negative Words	
			Overall #/% Emotional Words	
			Avg. Sentiment Score of Words	NLTK.Sentiment Package
		Punctuations (4)	# '!'	Self-implemented
			# '?'	Self-implemented
			# '...'	Self-implemented
			Overall # '!' '?' '...'	Self-implemented
		Similarity between Headline & Bodytext (2)	Word2Vec + Cosine Distance	[27]
			Sentence2Vec + Cosine Distance	[2]
	News-worthiness (20)	Quality (8)	Word2Vec + Cosine Distance	[27]
			Sentence2Vec + Cosine Distance	[2]
			#/% Content Words	LIWC
			#/% Function Words	
			#/% Stop Words	Self-implemented
		Informality (12)	#/% Swear Words	LIWC
			#/% Netspeak	
			#/% Assent	
			#/% Nonfluencies	
			#/% Fillers	
			Overall #/% Informal Words	

REFERENCES

[1] Amol Agrawal. 2016. Clickbait detection using deep learning. In *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*. IEEE, 268–272.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. (2016).

[3] Péter Bálint and Géza Bálint. 2009. The Semmelweis-reflex. *Orvosi hetilap* 150, 30 (2009), 1430.

[4] Lawrence E Boehm. 1994. The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin* 20, 3 (1994), 285–293.

[5] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. 84–89.

[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

[7] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 9–16.

[8] Abhijnan Chakraborty, Rajdeep Sarkar, Ayushi Mrigen, and Niloy Ganguly. 2017. Tabloids in the Era of Social Media?: Understanding the Production and Consumption of Clickbaits in Twitter. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 30.

[9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[10] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.
[11] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.
[12] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 601–610.
[13] 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*. Association for Computational Linguistics, 171–175.
[14] Alison Gianotto. 2014. Downworthy: A browser plugin to turn hyperbolic viral headlines into what they really mean. *downworthy. snipe. net* (2014).
[15] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 153–164.
[16] Shashank Gupta, Raghuveer Thirukovalluru, Manjira Sinha, and Sandya Mannarswamy. 2018. CIMTDetect: A Community Infused Matrix-Tensor Coupled Factorization Based Method for Fake News Detection. *arXiv preprint arXiv:1809.05252* (2018).
[17] J Hooper. 1975. On Assertive Predicates in Syntax and Semantics, Vol. 4. *New York* (1975).
[18] Kokil Jaidka, Tanya Goyal, and Niyati Chhaya. 2018. Predicting email and article clickthroughs with domain-adaptive language models. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 177–184.
[19] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 13–24.
[20] 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*. IEEE, 230–239.
[21] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs.. In *AAAI*. 2972–2978.
[22] Marcia K Johnson and Carol L Raye. 1981. Reality monitoring. *Psychological review* 88, 1 (1981), 67.
[23] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.
[24] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.
[25] Colin MacLeod, Andrew Mathews, and Philip Tata. 1986. Attentional bias in emotional disorders. *Journal of abnormal psychology* 95, 1 (1986), 15.
[26] Steven A McCornack, Kelly Morrison, Jihyun Esther Paik, Amy M Wisner, and Xun Zhu. 2014. Information manipulation theory 2: a propositional theory of deceptive discourse production. *Journal of Language and Social Psychology* 33, 4 (2014), 348–377.
[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[28] Sina Mohseni, Eric Ragan, and Xia Hu. 2019. Open Issues in Combating Fake News: Interpretability as an Opportunity. *arXiv preprint arXiv:1904.03016* (2019).
[29] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. *arXiv preprint arXiv:1902.06673* (2019).
[30] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2016), 11–33.
[31] Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A Survey on Natural Language Processing for Fake News Detection. *arXiv preprint arXiv:1811.00770* (2018).
[32] Supavich Fone Pengnate. 2016. Measuring emotional arousal in clickbait: eye-tracking approach. (2016).
[33] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
[34] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. *arXiv preprint arXiv:1708.07104* (2017).
[35] D Pisarevskaya. 2015. Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language. In *Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search (AINL-ISMW) FRUCT Conference, Saint-Petersburg, Russia*.
[36] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A Stylometric Inquiry into Hyperpartisan and Fake News. *arXiv preprint arXiv:1702.05638* (2017).

[37] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European Conference on Information Retrieval*. Springer, 810–817.
[38] K Rapoza. 2017. Can 'fake news' impact the stock market?
[39] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1650–1659.
[40] Arne Roets et al. 2017. 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* 65 (2017), 107–110.
[41] Victoria L Rubin. 2010. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the Association for Information Science and Technology* 47, 1 (2010), 1–10.
[42] Victoria L Rubin and Tatiana Lukoianova. 2015. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology* 66, 5 (2015), 905–917.
[43] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.
[44] Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems* 104 (2016), 123–133.
[45] 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, 1 (2017), 22–36.
[46] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 312–320.
[47] Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News* 16 (2016).
[48] Udo Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der psychologie* 11 (1967), 26–181.
[49] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 647–653.
[50] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[51] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).
[52] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.
[53] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 651–662.
[54] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *Companion of The Web Conference*. 155–158.
[55] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social media mining: an introduction*. Cambridge University Press.
[56] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 603–612.
[57] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. 2018. Fake News Detection with Deep Diffusive Network Model. *arXiv preprint arXiv:1805.08751* (2018).
[58] Xinyi Zhou and Reza Zafarani. 2018. Fake News: A Survey of Research, Detection Methods, and Opportunities. *arXiv preprint arXiv:1812.00315* (2018).
[59] Xinyi Zhou and Reza Zafarani. 2019. Fake News Detection: An Interdisciplinary Research. In *Companion of The Web Conference*. https://doi.org/10.1145/3308560.3316476
[60] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake News: Fundamental Theories, Detection Strategies and Challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 836–837.
[61] Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. 1981. Verbal and Nonverbal Communication of Deception1. In *Advances in experimental social psychology*. Vol. 14. Elsevier, 1–59.