Fake News Early Detection: A Theory-driven Model

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

Massive dissemination of fake news and its potential to erode democracy has 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. Nevertheless, to detect fake news at an early stage, i.e., when it is published on a news outlet but not yet spread on social media, one cannot rely on news propagation information as it does not exist. Hence, there is a strong need to develop approaches that can detect fake news by focusing on news content. In this article, 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 the proposed method can outperform the state-of-the-art and enable fake news early detection 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, click-bait, feature engineering, interdisciplinary research

ACM Reference format:
Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2020. Fake News Early Detection: A Theory-driven Model. *Digit. Threat.: Res. Pract.* 1, 2, Article 12 (June 2020), 25 pages.
https://doi.org/10.1145/3377478

1 INTRODUCTION

Fake news is now viewed as one of the greatest threats to democracy and journalism [61]. The reach of fake news was best highlighted during the critical months of the 2016 U.S. presidential election campaign, where the top 20 frequently discussed fake election stories (see Figure 1 for an example) generated 8,711,000 shares, reactions, and comments on Facebook, which is larger than the total of 7,367,000 for the top 20 most-discussed election stories posted by 19 major news websites [52]. Our economies are not immune to the spread of fake

Authors' addresses: X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, EECS Department, Syracuse University, Syracuse, NY, 13244; emails: zhouxinyi@data.syr.edu, {atjain, vvphoha}@syr.edu, reza@data.syr.edu.

2576-5337/2020/06-ART12 $15.00
https://doi.org/10.1145/3377478

Fig. 1. Fake news.[1] (1) This fake news story originally published on Ending the Fed got ~754,000 engagements in the final three months of the 2016 U.S. presidential campaign, which is in the top-three-performing fake election news stories on Facebook [52]; (2) A fake news story with clickbait.

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 $130B in stock value [43].

Meanwhile, humans have been proven to be not proficient in 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 better than chance: typical accuracy rates are in the range of 55%–58%, with a mean accuracy of 54% over 1K participants in over 100 experiments [45]. Many expert-based (e.g., PolitiFact[2] and Snopes[3]) and crowd-sourced (e.g., Fiskkit[4] and TextThresher [62]) manual fact-checking websites, tools, and platforms thus have emerged to serve the public on this matter.[5] Nevertheless, manual fact-checking does not scale well with the volume of newly created information, especially on social media [60]. Hence, automatic fake news detection has been developed in recent years, where current methods can be generally grouped into *content-based* and *propagation-based* methods.

Content-based fake news detection aims to detect fake news by analyzing the content of news articles. Within a machine learning framework, researchers often detect fake news relying on either latent (via neural networks) [58, 64] or non-latent (usually hand-crafted) features [12, 40, 48, 53] of the content (see Section 2 for details). 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 explainable machine learning models for fake news detection [15]. For example, *Undeutsch hypothesis* [55] states that a fake statement differs in writing style and quality from a true one. Such theories, as will be summarized in Section 2.2, can refer to either *deception/disinformation* [24, 30, 55, 67], i.e., information that is intentionally and verifiably false, or *clickbaits* [28], the headlines whose main purpose is to attract the attention of readers and encourage them to click on a link to a particular webpage [65]. Compared to existing features, relying on such theories allows to introduce features that are explainable, 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 that includes fake news articles,

[1]Direct source: https://bit.ly/2uE5eaB.
[2]https://www.politifact.com/.
[3]https://www.snopes.com/.
[4]http://www.fiskkit.com/.
[5]Comparison among common fact-checking websites is provided in Reference [65] and a comprehensive list of fact-checking websites is available at https://reporterslab.org/fact-checking/.

fake statements, fake reviews, and so on. Hence, the characteristics attached to deception/disinformation might or might not be consistent with that of fake news, which motivates us to explore the relationships between fake news and types of deception. Meanwhile, clickbaits have been shown to be closely correlated to fake news [11]. 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 newsworthy [65]. Unfortunately, clickbaits help fake news attract more clicks (i.e., visibility) and further gain public trust, as indicated by the *attentional bias* [29], which states that the public trust to a certain news article will increase with more exposure, as facilitated by clickbaits. However, 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.

Unlike content-based fake news detection, 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 reasonable performance [7, 19, 23, 47, 51, 63, 66]. However, propagation-based methods face a major challenge when detecting fake news. Within a life cycle of any news article, there are three basic stages: being created, being published on news outlet(s), and being spread on social media (medium) [65]. Propagation-based models relying on social context information are difficult to be applied in predicting fake news before its third stage, i.e., before fake news has been propagated on social media. To detect fake news at an early stage, i.e., when it is published on a news outlet but has not yet spread on social media sites, to take early actions for fake news intervention (i.e., *fake new early detection*) motivates us to deeply mine news content. 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 that it is difficult to correct one's cognition after fake news has gained their trust (i.e., *Semmelweis reflex* [3], *confirmation bias* [34], and *anchoring bias* [54]).

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. Here, we aim to address these challenges by developing a theory-driven fake news detection model that concentrates on news content to be able to detect fake news before it has been propagated on social media. 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 study. Features are then utilized for fake news detection within a supervised machine learning framework. The specific contributions of this article are as follows:

(1) We propose a model that enables fake news early detection. By solely relying on news content, the model allows to conduct detection when fake news has been published on a news outlet while it has not been disseminated on social media. Experimental results on real-world datasets validate the model effectiveness when limited news content information is available.
(2) We conduct an interdisciplinary fake news study by broadly investigating social and psychological theories, which presents a systematic framework for fake news feature engineering. Within such framework, each news article is represented, respectively, at the lexicon-, syntax-, semantic-, and discourse-level within language. Compared to latent features, such features can enhance model interpretability, help fake news pattern discovery, and help the public better understand fake news.
(3) We explore the relationships among fake news, types of deception, and clickbaits. By empirically studying their characteristics in, e.g., content quality, sentiment, quantity, and readability, some news patterns unique to fake news or shared with deception or clickbaits are revealed.

The rest of this article 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 article.

2 RELATED WORK

Our work is mainly related to the detection of fake news, with the investigation of its characteristics as well as that of types of deception/disinformation and clickbaits. Next, we review the development of fake news detection in Section 2.2.1 and summarize the characteristics of fake news, deception, and clickbait in Section 2.2.

2.1 Fake News Detection

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.

2.1.1 Content-based Fake News Detection. In general, current content-based approaches detect fake news by representing news content within different frameworks. 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 usually develops link prediction algorithms [12, 48] with the goal of directly evaluating news authenticity by comparing (inferring) the knowledge extracted from to-be-verified news content with that within a Knowledge Graph (KG) such as Knowledge Vault [14]. 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. First, KGs are often far from *complete*, often demanding further post-processing approaches for knowledge inference [33]. 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 [65].

II. Style is a set of self-defined [non-latent] machine learning features that can represent fake news and differentiate it from the truth [65]. For example, such style features can be word-level statistics based on TF-IDF, *n*-grams and/or LIWC features [6, 40, 41], and rewrite-rule statistics based on TF-IDF [40]. Though these style features can be comprehensive in detecting fake news, their selection or extraction is driven by experience that is rarely supported by fundamental theories across disciplines. An example of such a machine learning framework is the interesting study by Rubin and Lukoianova, which identifies fake news by combining rhetorical structures with vector space model [46].

III. Latent features represent news articles via automatically generated features often obtained by matrix/tensor factorization or deep learning techniques, e.g., Text-CNN [25, 58, 64]. Though these latent features can perform well in detecting fake news, they are often difficult to comprehend, which brings challenges to promote the public's understanding of fake news.

2.1.2 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 [32].

A direct way of presenting news propagation is using a *news cascade*—a tree structure presenting post-repost relationships for each news article on social media, e.g., tweets and retweets on Twitter [7, 65]. Based on news cascades, for example, Wu et al. [59] 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. Liu and Wu model

news cascades as multivariate time series. Based on that, fake news is detected by incorporating both Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) [27].

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. [23] 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 [18], matrix and tensor factorization [19, 51], or RNN [47, 63] have been developed for fake news detection.

While remarkable progress has been made, to detect fake news at an early stage, i.e., when it is published on a news outlet and before it has been spread on any social media, one cannot rely on social context information 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.

2.2 Fake News, Deception, and Clickbait Characteristics

We review the studies that reveal the characteristics of fake news, deception, and clickbait, respectively, in Section 2.2.1 to Section 2.2.3.

2.2.1 Fake News. Most current studies focus on investigating the patterns and characteristics in the propagation of fake news compared to that of the truth [38]. Vosoughi et al. investigate the differential diffusion of true and fake news stories distributed on Twitter from 2006 to 2017, where the data are ~126K stories tweeted by ~3M people more than 4.5M times [56]. 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. Recently, Zhou and Zafarani reveal that fake news spreaders often form a denser social network compared to true news spreaders [66].

2.2.2 Deception. Deception (disinformation) is the information that is intentionally false [65]. Deception has various forms, where fake (deceptive) statements and justifications are referred by most studies. Fundamental theories in psychology and social science have revealed some linguistic cues when a person lies compared to when he or she tells the truth. For example, *Undeutsch hypothesis* [55] states that a statement based on a factual experience differs in content style and quality from that of fantasy; *reality monitoring* [24] indicates that actual events are characterized by higher levels of sensory-perceptual information; *four-factor theory* [67] reveals that lies are expressed differently in terms of emotions and cognitive processes from truth; and *information manipulation theory* [30] validates that extreme information quantity often exists in deception.

2.2.3 Clickbait. Clickbait is the headlines whose main purpose is to attract the attention of readers and encourage them to click on a link to a particular Web page. Examples of clickbait are "33 Heartbreaking Photos Taken Just Before Death," "You Won't Believe Obama Did That No President Has Ever Done!" and "IT'S OVER: Hillary's ISIS Email Just Leaked & It's Worse Than Anyone Could Have Imagined…" (Figure 1). To achieve the purpose, clickbait creators make great efforts to produce an *information gap* [28] between the headlines and individuals' knowledge. Such information gaps produce the feeling of deprivation labeled curiosity, which motivates individuals to obtain the missing information to reduce such feeling.

3 METHODOLOGY

In this section, we detail the proposed method of predicting fake news. Before further elaboration, we formally define the target problem as below:

Problem Definition. Assume a to-be-verified news article can be represented as a feature vector $\mathbf{f} \in \mathbb{R}^n$, where each entry of $\mathbf{f}$ is a linguistic machine learning feature. The task to classify the news article based on its content

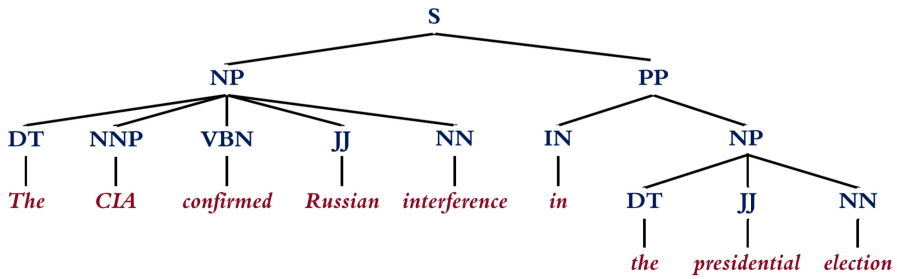


Fig. 2. PCFG parsing tree for the sentence "The CIA confirmed Russian interference in the presidential election" within a fake news article. The rewrite rules of this sentence should be the following: S → NP PP, NP → DT NNP VBN JJ NN, PP → IN NP, NP → DT JJ NN, DT → "the," NNP → "CIA," VBN → "confirmed,"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

representation is to identify a function $\mathcal{A}$, such that $\mathcal{A} : \mathbf{f} \xrightarrow{TD} \hat{y}$, where $\hat{y} \in \{0, 1\}$ is the predicted news label; 1 indicates that the news article is predicted as fake news, and 0 indicates it is true news. $TD = \{(\mathbf{f}^{(k)}, y^{(k)}) : \mathbf{f}^{(k)} \in \mathbb{R}^n, y^{(k)} \in \{0, 1\}, k \in \mathbb{N}_+\}$ is the training dataset. The training dataset helps estimate the parameters within $\mathcal{A}$ and consists of a set of news articles represented by the same set of features ($\mathbf{f}^{(k)}$) with known news labels ($y^{(k)}$).

Within the aforementioned traditional supervised learning framework, an explainable and well-performed method of predicting fake news relies on (1) the way that a news article is represented ($\mathbf{f}$), and (2) the classifier used to predict fake news ($\mathcal{A}$). Next, we will specify each in Section 3.1 and Section 3.2, respectively.

3.1 News Representation

As suggested by *Undeutsch hypothesis* [55], 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 3.1.1), syntax-level (Section 3.1.2), semantic-level (Section 3.1.3), and discourse-level (Section 3.1.4).

3.1.1 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-Words (BOW)* model. However, BOW representation can only capture the absolute frequencies of terms within a news article rather than their *relative (standardized) frequencies* that have accounted for the impact of content length (i.e., the overall number of words within the news content); the latter is more representative when extracting writing style features based on the words or topics that authors prefer to use or involve. Therefore, we first use a standardized BOW model to represent the writing style of each news article at the lexicon-level. Mathematically, assume a corpus contains p news articles $M = \{m_1, m_2, \ldots, m_p\}$ with a total of q words $W = \{w_1, w_2, \ldots, w_q\}$. x_j^i denotes the number of w_j appearing in m_i. Then the standardized frequency of w_j for news m_i is $x_j^i / \sum_{j=1}^{q} x_j^i$.

3.1.2 Syntax-level. Syntax-level style features can be further grouped into shallow syntactic features and deep syntactic features [16], 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 news articles in a relative (standardized) way, which removes the impact of news content length (i.e., instead of denoting the ith word, w_i defined in the last section here indicates the ith POS tag or rewrite rule).

3.1.3 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 well-established fundamental theories initially developed in forensic and social psychology. As specified in

Section 2.2, most these theories are not accurately developed fake news, but for deception/disiformation or clickbait that includes or closely relates to fake news. We detail our feature engineering at semantic level by separating news content as headlines and body-text, where clickbait-related attributes target news headlines and disinformation-related ones are mainly concerned with news body-text. A detailed list of semantic-level features defined and selected in our study is provided in Appendix A.

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 [29]. Hence, we aim to extract a set of features that can well represent clickbaits to capture fake news headlines. 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" [17]. A general way of representing news headlines based on these dictionaries is to verify 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 [9, 21, 42].

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 [28]. Such an 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 [28], 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 [17], 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 the news headline sentiment polarity by computing the average sentiment scores of the words it contains.
- *Punctuation*: Some punctuation can help express sensationalism or extreme sentiments, e.g., quotes (" . . . "), question marks ("?"), 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 [13]. Capturing such similarity requires first embedding the headline and body-text for each news article into the same space. To achieve this goal, we, respectively, utilize WORD2VEC [31] 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 word or sentence, we compute the average of its

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

word embedding (i.e., vectors) or sentence embedding. 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. To the best of our knowledge, similarity between the headlines and their body-text is first captured in such a way.

D. Newsworthiness. While click-baits can attract eyeballs, they are rarely newsworthy with (I) low quality and (II) high informality [65]. 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 the body-text [13]. 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. Second, such titles should be a *simplified* summary of the whole news event described in body-text, where meaningful words should occupy its main proportion [8]. From 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 [39] provides five dimensions to evaluate such informality of language: (1) *swear words* (e.g., "damn"); (2) *netspeak* (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.

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 [5]. Here, we aim to extract a set of features inspired by disinformation-related theories such as *Undeutsch hypothesis* [55], *reality monitoring* [24], *four-factor theory* [67], and *information manipulation theory* [30] to represent news content. Such features are with respect to:

A. Quality. In addition to writing style, *Undeutsch hypothesis* [55] 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 [39] 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 opposed to headline) and include them as features.
- *Diversity*: Such quality can possibly be assessed by investigating the number (proportion) of unique (non-repeated) words, content words, nouns, verbs, adjectives, and adverbs being used in news content. We compute and include them as features as well.
- *Subjectivity*: When a news article becomes hyperpartisan and biased, its quality should also be considered to be lower compared with those that maintain objectivity [41]. Benefiting from the work done by Recasens et al. [44], which provides the corpus of biased lexicons, here, we evaluate the subjectivity of news articles by counting their number (proportion) of biased words. However, factive verbs (e.g., "observe") [20] and report verbs (e.g., "announce") [44], as the opposite of biased ones, their numbers (proportions) are also included in our feature set, which are negatively correlated to content subjectivity.

B. Sentiment. Sentiment expressed within news content is suggested to be different within fake news and true news [67]. 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.

C. Quantity. Information manipulation theory [30] reveals that extreme information quantity (too much or too little) often exists in deceptive content. 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.

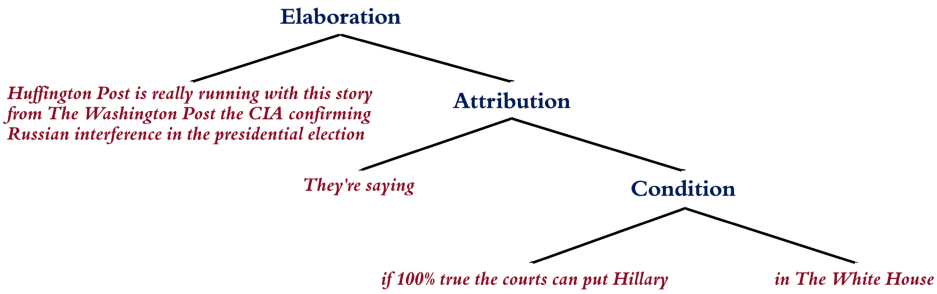


Fig. 3. Rhetorical structure for the partial content "Huffington Post is really running with this story from The Washington Post about the CIA confirming Russian interference in the presidential election. They're saying if 100% true, the courts can PUT HILLARY IN THE WHITE HOUSE!" within a fake news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

D. Specificity. Fictitious stories often differ in cognitive and perceptual processes, as indicated by *reality monitoring* [24] and *four-factor theory* [67]. Based on LIWC dictionary [39], 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.

3.1.4 Discourse-level. We first extract style features at discourse-level by investigating the standardized frequencies of rhetorical relationships among phrases or sentences within a news article. Instead of denoting the ith word, w_i defined in Section 3.1.1 here indicates the ith rhetorical relationship. Such relationships can be obtained through an RST parser[7] [22], where an example is given in Figure 3. Specifically, for a given piece of content, its rhetorical relationships among phrases or sentence can form a tree structure, where each leaf node is a phrase or sentence while non-leaf node is the corresponding rhetorical relationship between two phrases or sentences.

3.2 News Classification

We have detailed how each news article can be represented across language levels with computational features inspired by fundamental theories. These features then can be utilized by supervised classifiers that are widely accepted and well-established, e.g., Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forests (RF), and XGBoost [10], for fake news prediction. As classifiers perform best for machine learning settings they were initially designed for (i.e., *no free lunch theorem*), it is illogical to determine algorithms that perform best for fake news detection [65]. Following the common machine learning setting, we experiment with each of the aforementioned classifiers based on our features and experimental settings; results will be comprehensively presented by multiple classifiers performing the best [66].

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 to in Section 4.1.2.

[7] https://github.com/jiyfeng/DPLP.

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

4.1.1 Datasets. Our experiments are conducted on two well-established public benchmark datasets of fake news detection[8] [49–51]. 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 containing (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 on the 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.

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) [40], (2) exploring news dissemination on social networks (i.e., propagation-based fake news detection) [7], or (3) utilizing both information within news content and news propagation information [51].

I. Pérez-Rosas et al. [40] 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. [7] 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. [51] 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 other studies. Compared to style features, such latent features are less explainable but have been empirically shown to

[8] https://github.com/KaiDMML/FakeNewsNet/tree/old-version.

Table 2. General Performance of Fake News Detection Models[9]

Method	PolitiFact				BuzzFeed			
	Acc.	Pre.	Rec.	F_1	Acc.	Pre.	Rec.	F_1
Perez-Rosas et al. [40]	.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 [31]	.688	.671	.663	.667	.703	.714	.722	.718
DOC2VEC [26]	.698	.684	.712	.698	.615	.610	.620	.615
Castillo et al. [7]	.794	.764	.889	.822	.789	.815	.774	.794
Shu et al. [51]	.878	.867	.893	.880	.864	.849	.893	.870
Our Model	**.892**	**.877**	**.908**	**.892**	**.879**	**.857**	**.902**	**.879**

Among the baselines, (1) the propagation-based model [7] can perform relatively well compared to content-based ones [26, 31, 40]; and (2) the hybrid model [51] 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.

be remarkably useful [36, 57]. Here, we consider as baselines classifiers that use as features **(IV) WORD2VEC** [31] and **(V) DOC2VEC** [26] embeddings of news articles.

4.2 Performance Evaluation

In our experiments, each dataset is randomly divided into training and testing datasets with the ratio 0.8 : 0.2. Several supervised classifiers have been used with five-fold cross-validation, among which SVM (with linear kernel), Random Forest (RF), and XGBoost[10] [10] perform best compared to the others (e.g., LR, Logistic Regression; and NB, Naïve Bayes) within both our model and baselines. The performance of the experiments is provided in terms of accuracy, precision, recall, and F_1 scores. 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 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 [7] can perform comparatively well compared to content-based ones [26, 31, 40]; and (2) the hybrid model [51] 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: WORD2VEC features [31] and features in the work of Perez-Rosas et al. [40] perform best with linear SVM; while features based on DOC2VEC [26], in the work of Castillo et al. [7], and in our work perform best by using XGBoost.

[9]For each dataset, the maximum value is underlined, that in each column is bold, and that in each row is colored in gray.
[10]https://github.com/dmlc/XGBoost.

Table 3. Feature Performance across Language Levels[9]

	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

Lexicon-level and deep syntax-level features outperform the others, where the performance of semantic-level and shallow syntax-level ones follow. When combining features (excluding RRs) across levels, it enhances the performance compared to when separately using them in predicting fake news.

4.2.2 Fake News Analysis across Language Levels. As 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 that is between 70% to 80%. However, (3) fake news prediction using the standardized frequencies of rhetorical relationships (discourse-level) does 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, much more than others that are within the order of a hundred; and (4) when combining features (excluding RRs) across levels, it enhances the performance compared to when separately using features within each level in predicting fake news. Such performance can achieve an accuracy value and F_1 score around ~88%. 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

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"

(b) POS Tags

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

(c) Rewrite Rules

Rank	PolitiFact	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

(d) RRs

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

standardized frequencies of POS (possessive ending), VBN (verb in a form of past participle), and JJ (adjective) can better differentiate 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 understand as semantic-level features. Considering that, detailed analyses for DIAs (Section 4.2.3) and CBAs (Section 4.2.4) are conducted next.

4.2.3 Types of Deception and Fake News. As discussed in Section 3.1.3, well-established forensic psychology theories on identifying deception/disinformation have inspired us to represent news content by 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 to compare types of deception/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% to 70% accuracy and F_1 score using PolitiFact data, and 50% to 60% accuracy and F_1 score using BuzzFeed data. Combining all attributes to detect fake news performs better than separately using each type of attribute, which

Table 5. Performance of Disinformation-related Attributes in Predicting Fake News[9]

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**	**.755**	**.755**	**.667**	**.647**	**.625**	**.625**

Individual attributes perform similarly, while combining all attributes performs better in predicting fake news.

Table 6. Important Disinformation-related Features and Attributes for Fake News Detection

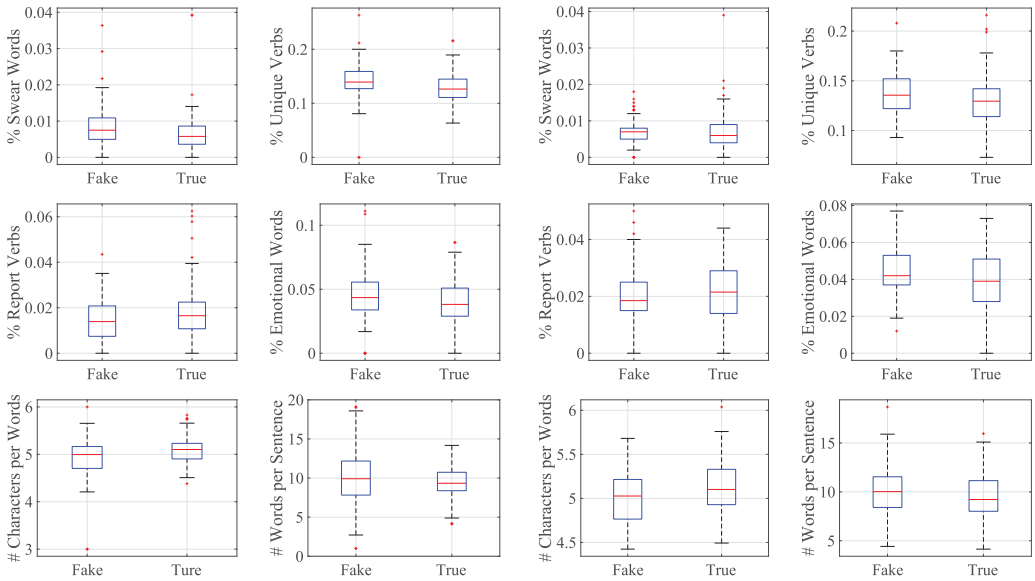
Rank	PolitiFact		BuzzFeed	
	Feature	Attribute	Feature	Attribute
1	# Characters per Word	Quantity	# Overall Informal Words	Informality
2	# Sentences per Paragraph	Quantity	% Unique Words	Diversity
3	% Positive Words	Sentiment	% Unique Nouns	Diversity
4	% Unique Words	Diversity	% Unique Content Words	Diversity
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

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.

can achieve 70% to 80% accuracy and F_1 score on PolitiFact data, and 60% to 70% accuracy and F_1 score on 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 10 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. Finally, (4) content informality and sentiments play a comparatively weak role in predicting fake news compared to the others.

E3: *Potential Patterns of Fake News Content.* We analyze each feature in DIA group, among which those that exhibit a consistent pattern in both datasets are presented in Figure 4. Specifically, we have the following observations:



(a) PolitiFact (b) BuzzFeed

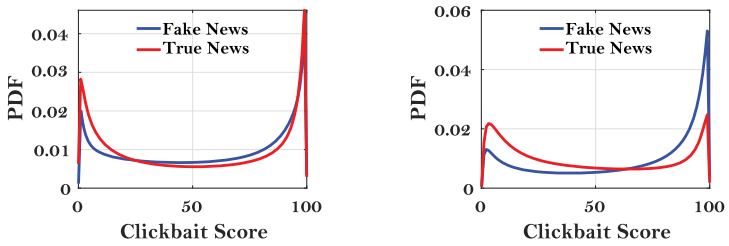
Fig. 4. Potential patterns of fake news. In both datasets, fake news shares higher (i) informality (% swear words), (ii) diversity (% unique verbs), and (iii) subjectivity (% report verbs), and is (iv) more emotional (% emotional words) with (v) longer sentences (# words per sentence) and (vi) shorter words (# characters per words) compared to true news.

- Similar to deception, fake news differs in content quality and sentiments expressed from the truth [55, 67]. Compared to true news, fake news often carries less report verbs and a greater proportion of unique verbs, swear words, and emotional (positive+negative) words.
- Compared to true news articles, fake news articles are characterized by shorter words and longer sentences.
- It is known that deception often does not involve cognitive and perceptual processes [24, 67]. However, the frequencies of lexicons related to cognitive and perceptual processes can hardly discriminate between fake and true news stories based on our datasets.

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 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 pre-trained 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 clickbait scores, true news articles often occupy a greater proportion compared to fake

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



(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 high clickbait scores, fake news articles often occupy a greater proportion compared to true news articles.

Table 7. Performance of Clickbait-related Attributes in Predicting Fake News[9]

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
Newsworthiness	**.729**	**.711**	**.683**	**.683**	**.686**	**.686**	.694	.667
Overall	.604	.612	.652	.652	.638	.628	**.705**	**.705**

Based on the experimental setup, newsworthiness of headlines outperforms the other attributes in predicting fake news.

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

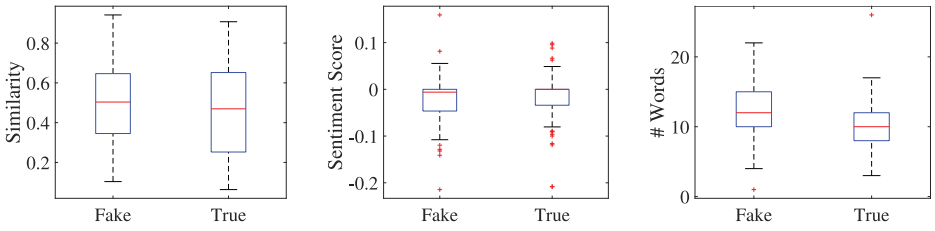
Rank	PolitiFact		BuzzFeed	
	Feature	Attribute	Feature	Attribute
1	Similarity (WORD2VEC)	S/N	Similarity (WORD2VEC)	S/N
2	Similarity (SENTENCE2VEC)	S/N	# Characters	R
3	% Netspeak	N	# Words	R
4	Sentiment Polarity	S	# Syllables	R
5	Coleman-Liau Index	R	Gunning-Fog Index	R

R: Readability; S: Sensationalism; N: Newsworthiness.

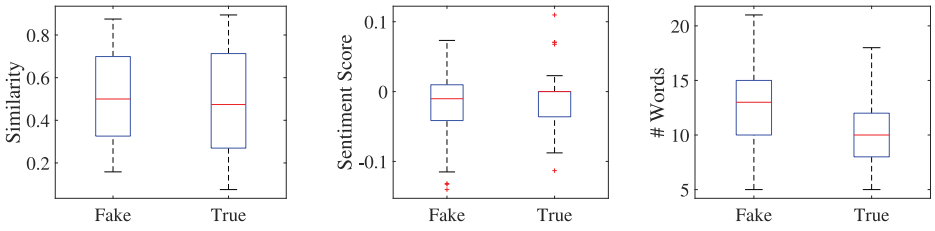
ones; while among news headlines with relatively high clickbait scores, a greater proportion often refers to fake news articles compared to true news articles.

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 newsworthiness, 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. Results indicate that (1) headline readability, sensationalism, and newsworthiness 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 newsworthiness of headlines rank relatively higher compared to that assessing headline readability and sensationalism.



(a) PolitiFact



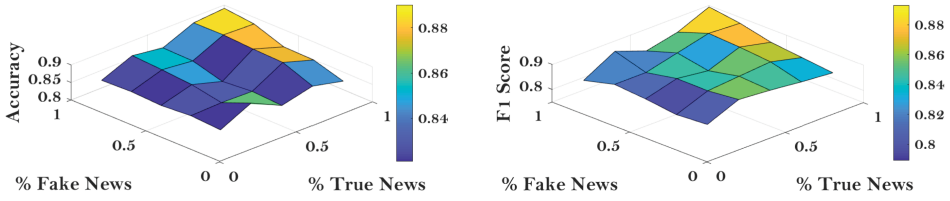
(b) BuzzFeed

Fig. 6. Potential patterns of fake news headlines. In both datasets, fake news headlines are generally (i) less similar to their body-text and contain (ii) more words when compared to true news. In addition, (iii) fake news headlines are slightly inclined to be negative with a broader scope of sentiment scores; while true news headlines are comparatively neutral with a more narrow scope of sentiment scores.

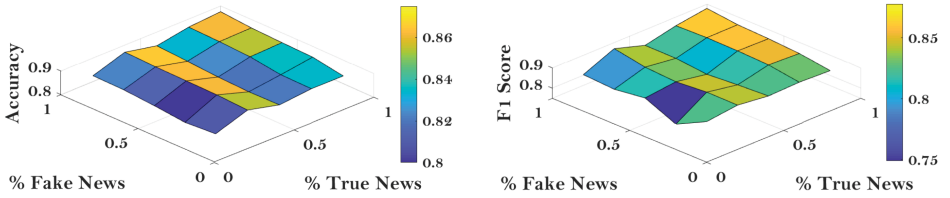
E4: *Potential Patterns of Fake News Headlines.* Using the boxplot of clickbait features within fake and true news, we examine whether fake news headlines share some potential patterns with clickbaits. Results are provided in Figure 6. Specifically,

- Figures in the left column present the box-plot of the cosine 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 newsworthiness 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 [13].
- Figures in the middle column present the box-plot of the average sentiment score of words within a news headline. Both figures reveal that, in general, fake news headlines are slightly inclined to be negative with a broader scope of sentiment scores; while true news headlines are comparatively neutral with a more narrow scope of sentiment scores. In other words, fake news headlines are more likely to be negative, or to be sensational with an extreme emotion, which matches with the characteristic of clickbaits [8].
- Figures in the right column present the box-plot 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.

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

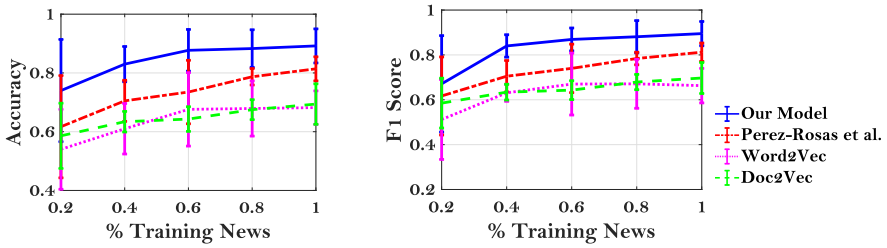


(a) PolitiFact

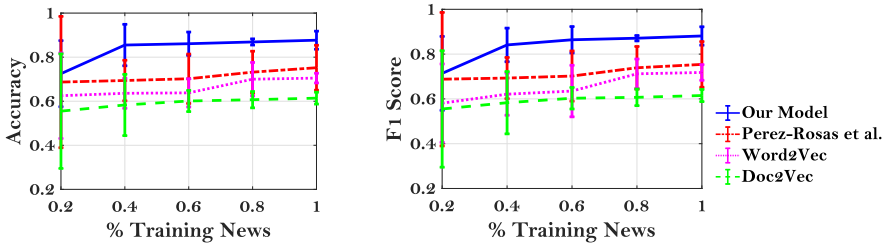


(b) BuzzFeed

Fig. 7. Performance sensitivity to news distribution (% fake news vs. % true news).



(a) PolitiFact

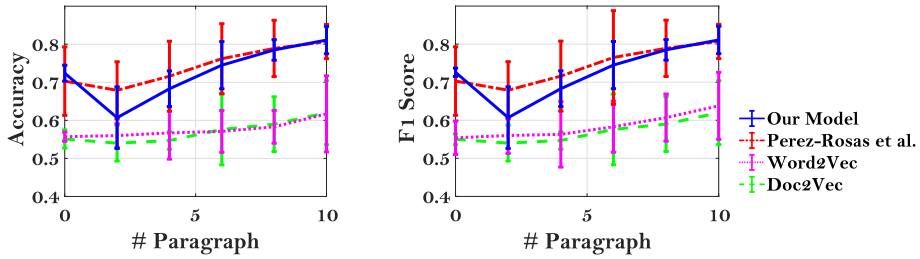


(b) BuzzFeed

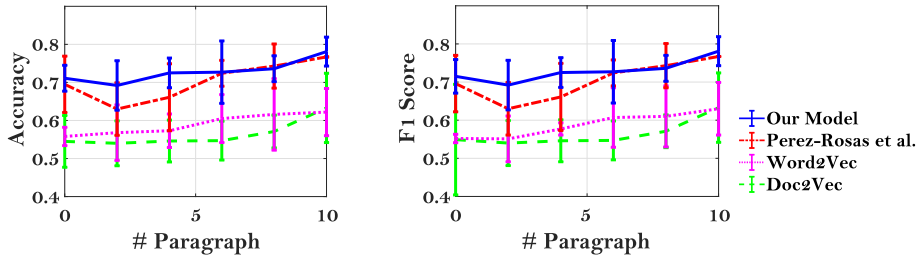
Fig. 8. Impact of the number of training news articles in predicting fake news.

fluctuates between ~0.75 and ~0.9. However, in most cases, the model is resilient to such perturbations and the accuracy and F_1 scores are between ~0.8 and ~0.88.

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 [58, 65]. Here, we measure the amount of such prior knowledge from two perspectives:



(a) PolitiFact



(b) BuzzFeed

Fig. 9. Impact of the available information within news content in predicting fake news.

(**E1**) the number of news articles available for learning and training a classifier, and (**E2**) the content quantity 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 ($n = 1, 2, \ldots$) 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. [40], our model generally has a comparable performance while can always outperform it when 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 [26, 31].

5 CONCLUSION

In this article, an interdisciplinary study is conducted for explainable fake news early detection. To predict fake news before it starts to propagate on social media, our work 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 style, quality, and sentiment from the truth, while it carries similar levels of cognitive and perceptual information compared to the truth. (5) Similar to clickbaits, fake news headlines present higher sensationalism and lower newsworthiness while their readability characteristics are complex and difficult to be directly concluded. In addition, fake news (6) is often matched with shorter words and longer sentences. With three stages of being created, being published on news outlet(s), and being propagated on any social media (medium), based on the proposed method, fake news proliferation can be mitigated before social media users have touched with it. Meanwhile, we should emphasize that a news article is likely, based on our observations, to be fake when it matches all (instead of any) potential patterns in its content. Note that our results do not indicate any news article sharing such characteristics is absolutely fake. To systematically reveal further patterns in fake news content compared to true news content, one has to involve (1) more fundamental theories and (2) empirical analyses on larger real-world datasets (see Reference [56] for an illustrated analysis for fake news propagation, and see Reference [35] for a recently released large-scale dataset). Datasets consisting of the ground truth of, e.g., both fake news and clickbaits, are invaluable to understand the relationships among different types of unreliable information; however, such datasets are so far rarely available. Furthermore, it should be pointed out that effective utilization of rhetorical relationships and utilizing news images [37] in an explainable way for fake news detection are still open issues. All aforementioned limitations will be part of our future work.

APPENDIX

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	[44]
			#/% Report Verbs	
			#/% Factive Verbs	[20]
	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	

(Continued)

Table 9. Continued

	Attribute		Feature(s)	Tool & Ref.
Clickbait-related Attributes (CBAs) (44)		General Clickbait Patterns (3)	# Common Clickbait Phrases	[17]
			# Common Clickbait Expressions	
			Overall # Common Clickbait Patterns	
		Readability (10)	Flesch Reading Ease Index (FREI)	Self-implemented
			Flesch-Kioncaid Grade Level (FKGL)	Self-implemented
			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	Word2Vec + Cosine Distance	[31]
			Sentence2Vec + Cosine Distance	[2]
	Newsworthiness (20)	Headline & Bodytext (2) Quality (8)	Word2Vec + Cosine Distance	[31]
			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 *Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT'16)*. IEEE, 268–272.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *The International Conference on Learning Representations (ICLR'17)*.

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

[4] Lawrence E. Boehm. 1994. The validity effect: A search for mediating variables. *Person. Soc. Psychol. Bull.* 20, 3 (1994), 285–293.

[5] Finn Brunton. 2013. *Spam: A Shadow History of the Internet*. The Mit Press.

[6] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Proceedings of the World Wide Web Conference*. ACM, 975–980.

[7] 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.

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

[9] 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. *Proc. ACM on Hum.-comput. Interact.* 1, CSCW (2017), 30.

[10] 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.

[11] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the ACM Workshop on Multimodal Deception Detection*. ACM, 15–19.

[12] 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.

[13] Manqing Dong, Lina Yao, Xianzhi Wang, Boualem Benatallah, and Chaoran Huang. 2019. Similarity-aware deep attentive model for clickbait detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 56–69.

[14] 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.

[15] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

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

[17] Alison Gianotto. 2014. Downworthy: A browser plugin to turn hyperbolic viral headlines into what they really mean. downworthy.snipe.net/. (2014).

[18] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on Twitter. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 153–164.

[19] 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).

[20] Joan B. Hooper. 1974. On assertive predicates. In *Syntax and Semantics*, Vol. 4. Indiana University Linguistics Club.

[21] 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.

[22] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 13–24.

[23] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'16)*. 2972–2978.

[24] Marcia K. Johnson and Carol L. Raye. 1981. Reality monitoring.*Psychol. Rev.* 88, 1 (1981), 67.

[25] Junaed Younus Khan, Md Khondaker, Tawkat Islam, Anindya Iqbal, and Sadia Afroz. 2019. A benchmark study on machine learning methods for fake news detection. *Arxiv Preprint Arxiv:1905.04749* (2019).

[26] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*. 1188–1196.

[27] Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[28] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychol. Bull.* 116, 1 (1994), 75.

[29] Colin MacLeod, Andrew Mathews, and Philip Tata. 1986. Attentional bias in emotional disorders.*J. Abnorm. Psychol.* 95, 1 (1986), 15.

[30] 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. *J. Lang. Soc. Psychol.* 33, 4 (2014), 348–377.

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Arxiv Preprint Arxiv:1301.3781* (2013).
[32] 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).
[33] 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.
[34] Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 2 (1998), 175.
[35] Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 630–638.
[36] 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).
[37] Shivam B. Parikh and Pradeep K. Atrey. 2018. Media-rich fake news detection: A survey. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'18)*. IEEE, 436–441.
[38] Shivam B. Parikh, Vikram Patil, Ravi Makawana, and Pradeep K. Atrey. 2019. Towards impact scoring of fake news. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'19)*. IEEE, 529–533.
[39] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC'15*. Technical Report. The University of Texas at Austin.
[40] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *Arxiv Preprint Arxiv:1708.07104* (2017).
[41] 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).
[42] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *Proceedings of the European Conference on Information Retrieval*. Springer, 810–817.
[43] Kenneth Rapoza. 2017. Can "fake news" impact the stock market? Retrieved from www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/ (9. 7. 2018).
[44] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1650–1659.
[45] Victoria L. Rubin. 2010. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proc. Assoc. Inf. Sci. Technol.* 47, 1 (2010), 1–10.
[46] Victoria L. Rubin and Tatiana Lukoianova. 2015. Truth and deception at the rhetorical structure level. *J. Assoc. Inf. Sci. Technol.* 66, 5 (2015), 905–917.
[47] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news detection. In *Proceedings of the ACM Conference on Information and Knowledge Management*. ACM, 797–806.
[48] Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowl-based Syst.* 104 (2016), 123–133.
[49] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable fake news detection. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press.
[50] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A data repository with news content, social context, and dynamic information for studying fake news on social media. *Arxiv Preprint Arxiv:1809.01286* (2018).
[51] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. ACM, 312–320.
[52] Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News* 16 (2016).
[53] Niraj Sitaula, Chilukuri K. Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2019. Credibility-based fake news detection. *Arxiv Preprint Arxiv:1911.00643* (2019).
[54] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.
[55] Udo Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen. *Handb. Psychol.* 11 (1967), 26–181.
[56] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[57] William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Arxiv Preprint Arxiv:1705.00648* (2017).
[58] 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.
[59] Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on Sina Eeibo by propagation structures. In *Proceedings of the IEEE 31st International Conference on Data Engineering (ICDE'15)*. IEEE, 651–662.
[60] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.

[61] Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. 2019. Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 3207–3208.

[62] 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 *Proceedings of the Web Conference*. International World Wide Web Conferences Steering Committee, 603–612.

[63] 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).

[64] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer.

[65] Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *Arxiv Preprint Arxiv:1812.00315* (2018).

[66] Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *SIGKDD Explor.* 21, 2 (2019), 48–60.

[67] Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In *Proceedings of the Advances in Experimental Social Psychology*. Vol. 14. Elsevier, 1–59.

Received April 2019; revised November 2019; accepted December 2019