Survey Paper:
Using Machine Learning Methods to do Fake News Automatic Detection

Jiahua You
Department of Computer Science and Engineering, UC Santa Cruz
jiyou@ucsc.edu

1 Research Overview and Summary

1. Introduction

Fake news has been in the public eye since 2016, they have been spread by prominent politicians, known media houses and through other sources such as social media and word of mouth. The impact has been felt by most, where the validity of news stories and claims have been challenged both politically and scientifically. The trustworthiness of news agencies have been heavily disputed, and the use of the "fake news" has turned into a shouting match about what points of view that are accepted by different people, and thus becoming filled with emotions instead of facts.

The amount of fake information is increasing and spreads to more and more topics, but overall, the more technical and complex a topic is, the harder it is to produce false claims and information for it. The fakes produced changes just as the normal news changes, and is often based on the same topics. For example, during the United States presidential election in 2016, massive amounts of political news was published and spread, and therefore the amount of politically loaded fakes were also in- creasing. Fake news is increasing, but the fight against it is also increasing, and the overall awareness about it and how to spot it as well. Tools are needed as they evolve, both to minimize, but also to combat it.

1.2 Fake News Characterization

1.2.1 Definition of Fake News

A narrow definition of fake news is news articles that are intentionally and verifiably false and could mis-lead readers. There are two key features of this definition: *authenticity* and *intent*. First, fake news includes false information that can be verified as such. Second, fake news is created with dishonest intention to mis-lead consumers.

1.2.2 Fake News on Traditional Media

Fake news itself is not a new problem. The media ecology of fake news has been changing over time from newsprint to radio/television and, recently, online news and social media. We denote "traditional fake news" as the fake news problem before social media had important effects on its production and dissemination. Next, we will describe several psychological and social science foundations that describe the impact of fake news at both the individual and social information ecosystem levels.

Psychological Foundations of Fake News

Humans are naturally not very good at differentiating between real and fake news. There are several psychological and cognitive theories that can explain this phenomenon and the influential

power of fake news. Traditional fake news mainly targets consumers by exploiting their individual vulnerabilities.

There are two major factors which make consumers naturally vulnerable to fake news:

(i) ***Naïve Realism***: consumers tend to believe that their perceptions of reality are the only accurate views, while others who disagree are regarded as uninformed, irrational, or biased;

(ii) ***Confirmation Bias***: consumers prefer to receive information that confirms their existing views.

Due to these cognitive biases inherent in human nature, fake news can often be perceived as real by consumers. Moreover, once the misperception is formed, it is very hard to correct it. Psychology studies shows that correction of false information (e.g., fake news) by the presentation of true, factual information is not only unhelpful

2 Theoretical Backgrounds and Methodology

1. Network Analysis

2.1.1 Network Properties

In this section, we outline the potential role of network properties for the study of fake news. First, users form groups with like-minded people, resulting in what are widely known as *echo chambers*. Second, *individual users* play different roles in the dissemination of fake news. Third, social media platforms allow users to personalize how information is presented to them, thus isolating users from information outside their personalized *filter bubbles*. Finally, highly active *malicious user accounts* become powerful sources and proliferators of fake news.

Echo Chambers

The process of seeking and consuming information on social media is becoming less mediated. Users on social media tend to follow like-minded people and thus receive news that promotes their preferred, existing narratives. This may increase social polarization, resulting in an *echo chamber* effect. The echo chamber effect facilitates the process by which people consume and believe fake news based on the following psychological factors:

(i) ***social credibility***, which means people are more likely to perceive a source as credible if others perceive it as such, especially when there is not enough information available to assess the truthfulness of that source; and

(ii) ***frequency heuristic***, which means that consumers may naturally favor information they hear frequently, even if it is fake news. In echo chambers, users share and consume the same information, which creates segmented and polarized communities.

Individual Users

During the fake news dissemination process, individual users play different roles. For example,

(i) ***persuaders*** spread fake news with supporting opinions to persuade and influence others to believe it;

(ii) ***gullible users*** are credulous and easily persuaded to believe fake news; and

(iii) ***clarifiers*** propose skeptical and opposing viewpoints to clarify fake news.

Social identity theory suggests that social acceptance and affirmation is essential to a person's identity and selfesteem, making persuaders likely to choose "socially safe" options when consuming and disseminating news information. They follow the norms established in the community even if the news being shared is fake news. The cascade of fake news is driven not only by influential persuaders but also by a critical mass of easily influenced individuals, i.e., gullible users. Gullibility is a different concept from trust.

In psychological theory, general trust is defined as the default expectations of other people's trustworthiness. High trusters are individuals who assume that people are trustworthy unless proven otherwise. Gullibility, on the other hand, is insensitivity to information revealing untrustworthiness. Reducing the diffusion of fake news to gullible users is critical to mitigating fake news. Clarifiers can spread opposing opinions against fake news and avoid one-sided viewpoints. Clarifiers can also spread true news which can:

(i) immunize users against changing beliefs before they are affected by fake news;

(ii) further propagate and spread true news to other users.

Filter Bubbles

A filter bubble is an intellectual isolation that occurs when social media websites use algorithms to personalize the information a user would want to see. The algorithms make assumptions about user preferences based on the user's historical data, such as former click behavior, browsing history, search history, and location. Given these assumptions, the website is more likely to present information that will support the user's past online activities. A filter bubble can reduce connections with contradicting viewpoints, causing the user to become intellectually isolated. A filter bubble will amplify the individual psychological challenges to dispelling fake news. These challenges include:

(i) ***Naïve Realism***: consumers tend to believe that their perceptions of reality are the only accurate views, while others who disagree are regarded as uninformed, irrational, or biased;

(ii) ***Confirmation Bias***: consumers prefer to receive information that confirms their existing views.

Malicious Accounts

Social media users can be malicious, and some malicious users may not even be real humans. Malicious accounts that can amplify the spread of fake news include social bots, trolls, and cyborg users. Social bots are social media accounts that are controlled by a computer algorithm. The algorithm automatically produces content and interacts with humans (or other bot users) on social media. Social bots can be malicious entities designed specifically for manipulating and spreading fake news on social media. Trolls are real human users who aim to disrupt online communities and provoke consumers to an emotional response. Trolls enable the easy dissemination of fake news among otherwise normal online communities. Finally, cyborg users can spread fake news in a way that blends automated activities with human input. Cyborg accounts are usually registered by a human as a disguise for automated programs that are set to perform activities on social media. The easy switch between humans and bots offer cyborg users unique opportunities to spread fake news.

2. Network Types

In this section, we introduce several network structures that are commonly used to detect and mitigate fake news. Then, following the three dimensions of the news dissemination ecosystem outlined above, we illustrate how *homogeneous* and *heterogeneous* networks can be built within a specific dimension and across dimensions.

Homogeneous Networks

(a) Friendship Network (b) Diffusion Network (c) Credibility Network

Node *u* indicate a user, and *s* represent a social media post.

Homogeneous networks have the same node and link types. As shown in Figure 2, we introduce three types of homogeneous networks: friendship networks, diffusion networks, and credibility networks. Each of these types is potentially useful in detecting and mitigating fake news.

Friendship Networks

Homophily theory [13] suggests that users tend to form relationships with like-minded friends, rather than with users who have opposing preferences and interests. Likewise, social influence theory [12] predicts that users are more likely to share similar latent interests towards news pieces. Thus, the friendship network provides the structure to understand the set of social

relationships among users. The friendship network is the basic route for news spreading and can reveal community information.

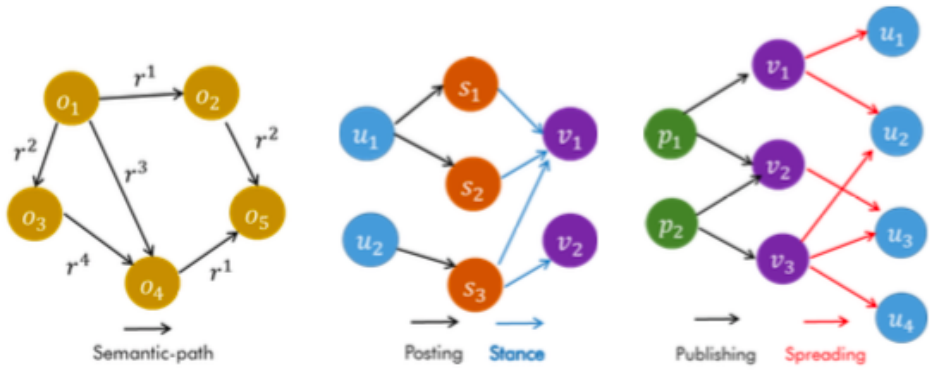
Diffusion Networks

The diffusion network is important for learning about representations of the structure and temporal patterns to help identify fake news. By discovering the sources of fake news and the spreading paths among the users, we can also better mitigate fake news problem.

Credibility Networks

Users express their viewpoints toward original news pieces through social media posts. In these posts, they can either express the same viewpoints (which mutually support each other), or conflicting viewpoints (which may reduce their credibility scores). By modeling these relationships, the credibility network can be used to evaluate the overall truthfulness of news by leveraging the credibility scores of each social media post relevant to the news.

Heterogeneous Networks



(a) Knowledge Network (b) Stance Network (c) Interaction Network

Node *o* indicate a knowledge entity, *v* represents a news item, *p* means a news publisher.

Heterogeneous networks have a different set of node and link types. The advantages of heterogeneous networks are the abilities to represent and encode information and relationships from different perspectives. During the news dissemination process, different types of entities are involved, including users, the social media posts, the actual news, etc. Figure shows the common types of heterogeneous networks for analyzing fake news:

- knowledge networks,
- stance networks,
- interaction networks.

Knowledge Networks

The knowledge network integrates linked open data, such as DBdata and Google Relation Extraction Corpus (GREC), as a heterogeneous network topology. Fact-checking using a

knowledge graph checks whether the claims in news content can be inferred from existing facts in the knowledge networks.

Stance Networks

Stances (or viewpoints) indicate the users' opinions towards the news, such as supporting, opposing, etc. Typically, fake news pieces will provoke tremendous controversial views among social media users, in which denying and questioning stances are found to play a crucial role in signaling claims as being fake.

Interaction Networks

The interaction networks can represent the correlations among different types of entities, such as publisher, news, and social media post, during the news dissemination process. The characteristics of publishers and users, and the publisher-news and news-users interactions have potential to differentiate fake news.

2.2 Fake News Detection

Fake news detection evaluates the truth value of a news piece, which can be formalized as a classification problem. The common procedure is feature extraction and model construction. In feature extraction, we capture the differentiable characteristics of news pieces to construct effective representations; Based on these representations, we can construct various models to learn and transform the features into a predicted label. To this end, we introduce how features and models can be extracted and constructed in different types of networks.

2.2.1 Interaction Network Embedding

Interaction networks describe the relationships among different entities such as publishers, news pieces, and users. Given the interaction networks the goal is to embed the different types of entities into the same latent space, by modeling the interactions among them. We can leverage the resultant feature representations of news to perform fake news detection.

News Embedding

We can use news content to find clues to differentiate fake news and true news. Using non-negative Matrix Factorization (NMF) we can attempt to project the document-word matrix to a joint latent semantic factor space with low dimensionality, such that the document-word relations are modeled as the inner product in the space.

User Embedding

On social media, people tend to form relationships with like-minded friends, rather than with users who have opposing preferences and interests. Thus, connected users are more likely to

share similar latent interests in news pieces. To obtain a standardized representation, we use nonnegative matrix factorization to learn the user's latent representations

User-News Embedding

The user-news interactions can be modeled by considering the relationships between user attributes and the level of veracity of news items. Intuitively, users with low credibility are more likely to spread fake news, while users with high credibility scores are less likely to spread fake news.

Publisher-News Embedding

The publisher-news interactions can be modeled by incorporating the characteristics of the publisher and news veracity values. Fake news is often written to convey opinions or claims that support the partisan bias of the news publisher.

2.2.2 Friendship Network Embedding

News temporal representations can capture the evolving patterns of news spreading sequences. However, we lose the direct dependencies of users, which plays an important role in fake news diffusion. The fact that users are likely to form echo chambers, strengthens our need to model user social representations and to explore its added value for a fake news study.

Credibility Network Propagation

The basic assumption is that the credibility of a given news event is highly related to the credibility of its relevant social media posts. To classify whether a news item is true of fake, we can collect all relevant social media posts. Then, we can evaluate the news veracity score by averaging the credibility scores of all the posts.

Network Initialization

Network initialization consists of two parts: node initialization and link initialization.

Network Optimization

Posts with supporting relations should have similar credibility values; posts with opposing relations should have opposing credibility values.

2.3 Fake News Mitigation

Fake news mitigation aims to reduce the negative effects brought by fake news. From a network analysis perspective, the goal is to minimize the scope of fake news spreading on social media. To achieve this, key spreaders of fake news need to be discovered such as provenances and persuaders. In addition, estimating the potential population affected by a fake news is useful for decision-makers to mitigate otherwise influential fake news. Moreover, choosing specific users to block the cascade of fake news, and even to start mitigation campaigns to immunize users are required to minimize the influence of fake news.

2.3.1 User Identification

Identifying key users on social media is important to mitigate the effect of fake news. For example, the provenances of fake news indicates the sources or originators. Provenances can help answer questions such as whether the piece of news has been modified during its propagation, and how an "owner" of the piece of information is connected to the transmission of the statement. In addition, it's necessary to identify influential persuaders to limit the spread scope of fake news by blocking the information flow from them to their followers on social media.

2.3.2 Network Size Estimation

The fake news diffusion process has different stages in terms of people's attention and reactions over time, resulting in a unique life cycle different from that of in-depth news. The impact of fake news on social media can be estimated as the number of users that are potentially affected by the news piece, an amount we want to assess and then minimize. We can adapt the network

scale-up based method for estimating the size of uncountable populations to estimate the size of the population affected by fake news on the provenance paths discussed previously.

2.3.3 Network Intervention

The goal of network intervention is to develop strategies to control the widespread dissemination of fake news before it goes viral. Network intervention mainly consists of two perspectives as follows:

Influence Minimization

Limiting the spread of fake news can be seen as analogous to inoculation in the face of an epidemic. Models of epidemics generally assume that a global parameter describes the probability that a user is infected by a neighbor. This assumption is violated in real-world situations of information exchange where users have varying degrees of willingness to accept information from their neighbors. Thus, the Independent Cascade Model (ICM) is proposed to alleviate this problem by assuming each edge has its specific activation probability. ICM is denoted as a sender-centric model.

Mitigation Campaign

Limiting the impact of fake news is not only to minimize the spread of fake news but also maximize the spread of true news. The campaign to mitigate fake news and to maximize true news forms during the information diffusion process. The network activities of fake news and real news can be represented as Multivariate Hawks Processes (MHP) with self and mutual excitations, where the control incentivizes more spontaneous mitigation

events. The influence of fake and real news is quantified using event exposure counts, represented by the number of times users are exposed to the news. The goal is to optimize the activity policy of a set of campaigner users to mitigate a fake news process stemming from another set of users. The whole idea is to optimize the performance of real news propagation (through the campaigner users) in diffusion network, ensuring that people who are exposed to fake news are also exposed to real news, so that they are less likely to be convinced by fake news.

2.4 Data Mining

This section will review fake news detection approaches from data mining perspective, including feature extraction and model construction.

2.4.1 News Content Features

Fake news detection on traditional news media mainly relies on news content, while in social media, extra social context auxiliary information can be used to as additional information to help detect fake news. The following is the detail of how to extract and represent useful features from *news content* and *social context*.

- **Source**: Author or publisher of the news article
- **Headline**: Short title text that aims to catch the attention of readers and describes the main topic if the article

- **Body Text**: Main text that elaborates the details of the news story; there is usually a major claim that is specifically highlighted and that shapes the angle of the publisher
- **Image / Video**: Part of the body content of a news article that provides visual cues to frame the story

Linguistic-based: It is reasonable to exploit linguistic features that capture the different writing styles and sensational headlines to detect fake news. Linguistic-based features are extracted from the text content in terms of document organizations from different levels, such as characters, words, sentences, and documents. In order to capture the different aspects of fake news and real news, existing work utilized both common linguistic features and domain-specific linguistic features.

    - **Lexical features**: character-level and word-level features, such as total words, characters per word, frequency of large words, and unique words;
    - **Syntactic features**: sentence-level features, such as frequency of function words and phrases, bag-of-words approaches, punctuation and parts-of-speech tagging.

Visual-based: Fake news exploits the individual vulnerabilities of people and thus often relies on sensational or even fake images to provoke anger or other emotional response of consumers. Visual-based features are extracted from visual elements (e.g. images and videos) to capture the different characteristics for fake news. Faking images were identified based on various user-level and tweet-level hand-crafted features using classification framework.

- **Visual features**: clarity score, coherence score, similarity distribution histogram, diversity score, and clustering score.
- **Statistical features**: count, image ratio, multi-image ratio, hot image ratio, long image ratio.

2.4.2 Social Context Features

User-based: Capturing users' profiles and characteristics by user-based features can provide useful information for fake news detection. User-based features represent the characteristics of those users who have interactions with the news on social media.

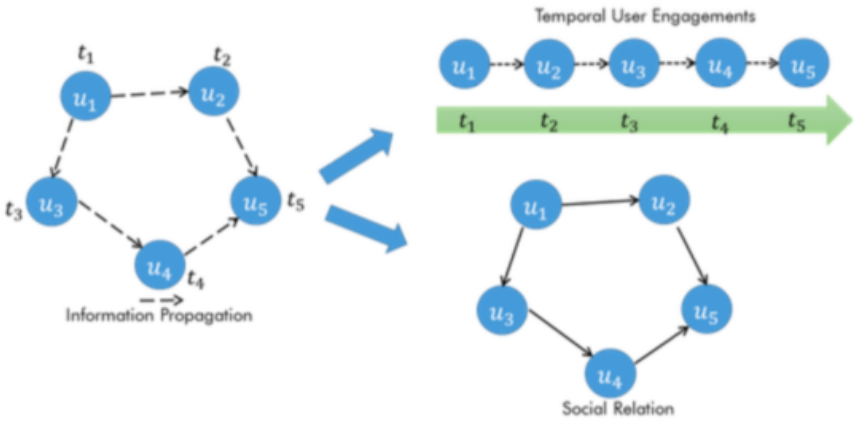
- **Individual level**: extracted to infer the credibility and reliability for each user using various aspects of user demographic. (registration age, number of followers, number of tweets the user has authored.)
- **Group level**: capture overall characteristics of groups of users related to the news.

Post-based: focus on identifying useful in- formation to infer the veracity of news from various aspects of relevant social media posts.

- **Post level**: generate feature values for each post.
- **Group level**: aim to aggregate the feature values for all relevant post for specific news articles by using "wisdom of crowds"
- **Temporal level**: consider the temporal variations of post level feature values.

Network-based: extracted via constructing specific networks among the users who published related social media posts. Different types of networks can be constructed.

- **Stance network**: can be built with nodes indicating all the tweets relevant to the news and the edge indicating the weights of similarity of stance.
- **Co-occurrence network**: based on user engagements by counting whether those users write post relevant to the same news articles.
- **Friendship network**: indicating the structure of users who post related tweets.
- **Diffusion network**: tracks the trajectory of the spread of news.



A diffusion network consists of temporal user engagements and a friendship network.

2.4.3 News Content Models

Knowledge-based: aim to use external sources to fact-check proposed claims in news content. The goal of fact-checking is to assign a truth value to a claim in a particular context.

- **Expert-oriented**: fact-checking heavily relies on human domain experts to investigate relevant data and documents to construct the verdicts of claim veracity
- **Crowdsourcing-oriented**: fact-checking exploits the "wisdom of crowd" to enable normal people to annotate news content; these annotations are then aggregated to produce an overall assessment of the news veracity.
- **Computational-oriented**: To identify check-worthy claims, factual claims in news content are extracted that convey key statements and viewpoints, facilitating the subsequent fact-checking process
-

Style-based: try to detect fake news by capturing the manipulators in the writing style of news content.

- **Deception-oriented**: stylometric methods capture the deceptive statements or claims from news content
    - **Deep Syntax**, which sentence can be transformed into rules that describe the syntax structure. Based on using probabilistic context free grammars

(FCPG), different rules can be developed for deception detection, such as unlexicalized/lexicalized production rules and grandparent rules.

- **Rhetorical structure,** capture the difference between deceptive and truthful sentence. Deep network models, such as Convolutional Neural Network (CNN), have also been applied to classify fake news veracity.

- **Objectivity-oriented**: capture style signals that can indicate a decreased objectivity of news con- tent and thus the potential to mislead consumers,
  - **Hyperpartisan styles**, represent extreme behavior in favor of a particular political party, which often correlates with a strong motivation to create fake news. Linguistic- based features can be applied to detect hyperpartisan articles.
  - **Yellow-journalism**, represents those articles that do not contain well researched news, but instead rely on eye-catching headlines (i.e., clickbait) with a propensity for exaggeration, sensationalization, scare-mongering, etc.

2.4.4 Social Context Models

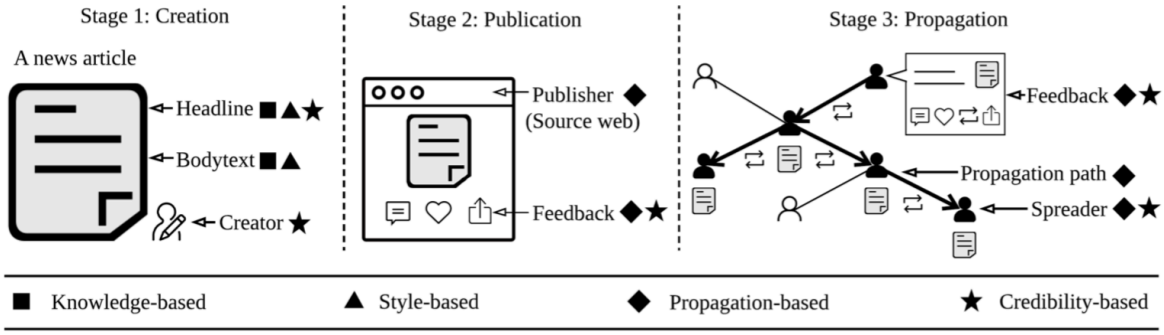
Stance-based: Stance detection is the task of automatically determining from a post whether the user is in favor of, neutral toward, or against some target entity, event, or idea. Stance-based approaches utilize users' view- points from relevant post contents to infer the veracity of original news articles.

- **Explicit stance**, are direct expression of emotion or opinion, such as the "thumbs up" and "thumbs down" reaction expressed in Facebook.
- **Implicit stance**, can automatically extracted from social media posts.

Propagation-based: Propagation-based approaches for fake news detection reason about the interrelations of relevant social media posts to predict news credibility. The basic assumption is that the credibility of a news event is highly related to the credibility of relevant social media posts.

- **Homogeneous credibility networks**, consist of a single type of entities, such as post or event.

- **Heterogeneous credibility networks**, involve different types of entities, such as post, sub-events, and events.


Stage 1: Creation
A news article
Headline ■▲★
Bodytext ■▲
Creator ★
Stage 2: Publication
Publisher ◆
(Source web)
Feedback ◆★
Stage 3: Propagation
Feedback ◆★
Propagation path ◆
Spreader ◆★
■ Knowledge-based
▲ Style-based
◆ Propagation-based
★ Credibility-based

2.5 Stance detection task

Identifying public misinformation is a complicated and challenging task. An important part of checking the veracity of a specific claim is to evaluate the stance different news sources take towards the assertion. Automatic stance evaluation, i.e. stance detection, would arguably facilitate the process of fact checking. In this section, we present our stance detection system which claimed third place in Stage 1 of the Fake News Challenge. Despite our straightforward approach, our system performs at a competitive level with the complex ensembles of the top two winning teams. We therefore propose our system as the 'simple but tough-to-beat baseline' for the Fake News Challenge stance detection task.

Automating stance evaluation has been suggested as a valuable first step towards assisting human fact checkers to detect inaccurate claims. The Fake News Challenge initiative thus recently organised the first stage of a competition (FNC-1) to foster the development of systems for automatically evaluating what a news source is saying about a particular issue.

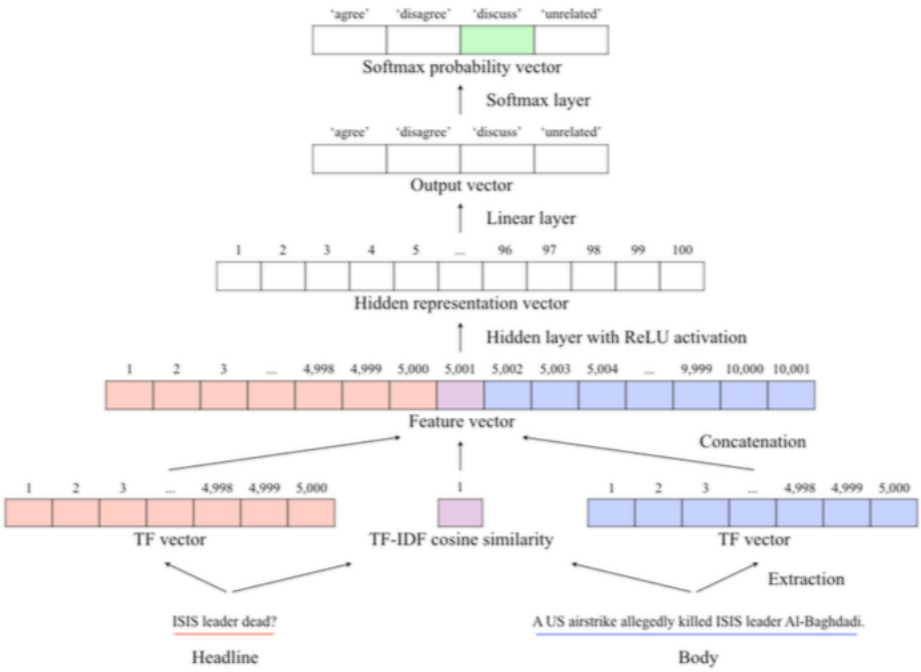
More specifically, FNC-1 involved developing a system that, given a news article headline and a news article body, estimates the stance of the body towards the headline. The stance label to be assigned could be one of the set: 'agree', 'disagree', 'discuss', or 'unrelated'.

Architecture Description

The single, end-to-end stance detection system consists of lexical and similarity features passed through a multi-layer perceptron (MLP) with one hidden layer. We use two simple bag-of-words (BOW) representations for the text inputs: term frequency (TF) and term frequency-inverse document frequency (TF-IDF).

Different vocabularies are used for calculating the TF and TF-IDF vectors. For the TF vectors, we extract a vocabulary of the 5,000 most frequent words in the training set and exclude stop words (the *scikit-learn* stop words for the English language with negation terms removed). For the TF-IDF vectors, a vocabulary of the 5,000 most frequent words is defined on both the training and test sets and the same set of stop words is excluded.

The TF vectors and the TF-IDF cosine similarity are concatenated in a feature vector of total size 10,001 and fed into the classifier.



Classifier

The classifier is a MLP [5] with one hidden layer of 100 units and a softmax on the output of the final linear layer. We use the rectified linear unit (ReLU) activation function [8] as non-linearity for the hidden layer. The system predicts with the highest scoring label (‘agree’, ‘disagree’, ‘discuss’, or ‘unrelated’). The classifier as described is fully implemented in TensorFlow [1].

2.6 Linguistic Analysis

Computational linguistics can aide in the process of identifying fake news in an automated manner well above the chance level. The proposed linguistics-driven approach suggests that to differentiate between fake and genuine content it is worthwhile to look at the lexical, syntactic and semantic level of a news item in question.

Linguistic Features

N-grams

We extract unigrams and bigrams derived from the bag of words representation of each news article. To account for occasional differences in content length, these features are encoded as term frequency-inverse document frequency (TF-IDF) values.

Punctuation

Previous research suggest that use of punctuation might be useful to differentiate deceptive from truthful texts. We construct a punctuation feature set consisting of twelve types of punctuation derived from the Linguistic Inquiry and Word Count software (LIWC, Version 1.3.1 2015)

(Pennebaker et al., 2015). This includes punctuation characters such as periods, commas, dashes, question marks and exclamation marks.

Psycholinguistic Features

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

Readability

We also extract features that indicate text understandability. These include content features such as the number of characters, complex words, long words, number of syllables, word types, and

number of paragraphs, among others content features. We also calculate several readability metrics, including the Flesch-Kincaid, Flesch Reading Ease, Gunning Fog, and the Automatic Readability Index (ARI).

Syntax

Finally, we extract a set of features derived from production rules based on context free grammars (CFG) trees using the Stanford Parser (Klein and Manning, 2003). The CFG derived features consist of all the lexicalized production rules (rules including child nodes) combined with their parent and grandparent node, e.g., *NNˆNP→commission (in this example NN –a noun– is the grandparent node, NP –noun phrase– the parent node, and "commissions" the child node). CFG-based features have been previously shown to be useful for linguistic deception detection (Feng et al., 2012). Features in this set are also encoded as tf-idf values.

Support Vector Machine

A support vector machine (SVM) is a classifier that works by separating a hyperplane (n-dimensional space) containing input. It is based on statistical learning theory. Given labeled training data, the algorithm outputs an optimal hyperplane which classifies new examples. The optimal hyperplane is calculated by finding the divider that minimizes the noise sensitivity and maximizes the generalization and margin of the model. A unique feature of the SVM is that the

hyperplane approach is based solely on the data points, and these points are called the support vectors. One of the major drawbacks with SVM is that it can only work with labeled data, and thus only work in a supervised training fashion.

SVM is not bound to linear separation, as they are able to transform input data into a high dimensional feature space, whereas a separating hyperplane can be found to work as an optimal classifier. One of the strengths of SVMs are that they can be used for very high dimensional problems, as long as their features can be mapped linearly in the feature space. The non-linear use of SVMs utilizes something called the *kernel trick*. The kernel trick works by replacing parts of the original algorithms with a kernel function instead of a dot function. Kernel methods can work in high-dimensional spaces because they compute the inner products between the data in the space instead of using the coordinates of the data. It is also worth mentioning that higher dimensional feature spaces increase the generalization error, but given enough samples, it still performs well.

2.7 Machine Learning

For machine learning (ML) training, there are three main approaches.

1. ***Supervised learning*** is learning where the learning data have both inputs and outputs so that we always know the correct answer to the input, and can train and adapt the ML algorithm to get the same output as the correct one.
2. ***Unsupervised learning***, where some of the data only contain input. Because of this, the system makes some assumptions and thus unsupervised classifies the input without the known correct answer.
3. ***Reinforcement learning*** is learning where there is no direct access to the correct output, but the quality of the output can be measured following input. Reinforcement learning uses rewards to quantify the output and over time the model is changed based on how the total reward changes.

This kind of heuristic approach where the model changes over time is similar to how ANNs work, and there is overlap between the methods, and also where they are applied.

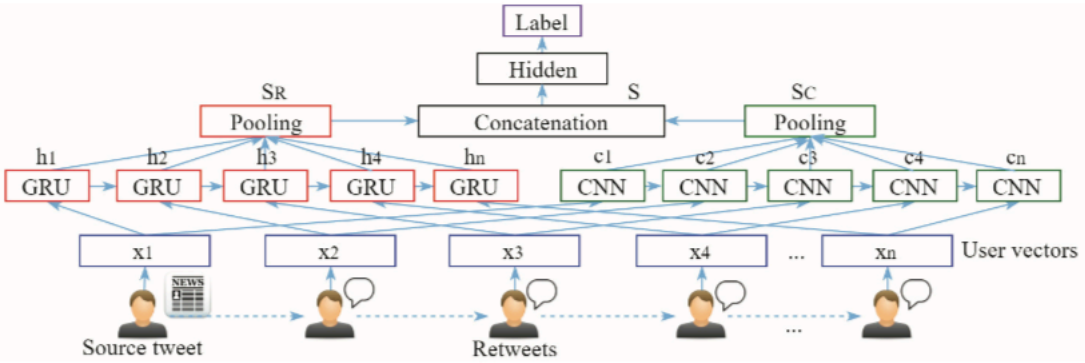
Propagation Path Classification with RNN and CNN

The proposed fake news detection model consists of four major components, i.e.,

- propagation path construction and transformation,
- RNN-based propagation path representation,

- CNN-based propagation path representation,
- Propagation path classification,

which are integrated together to detect fake news at the early stage of its propagation.



The architecture if the proposed fake news detection model

Propagation Path Construction and Transformation

Given a news story propagating on social media, we first construct its propagation path by first identifying the users who engaged in propagating the news. Then, its propagation path denoted as a *variable-length multivariate time series* is constructed by extracting user characteristic from relevant user profiles. And Then transform it into *fixed-length multivariate sequence*.

RNN-Based Propagation Path Representation

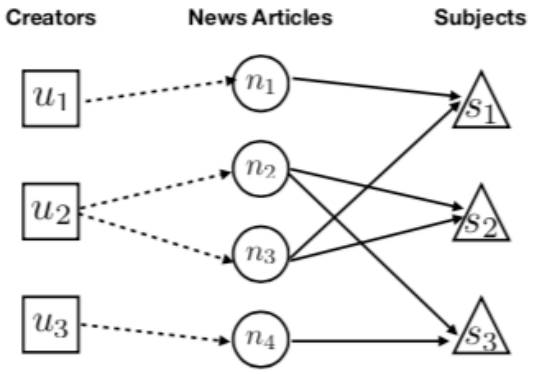
We utilize a variant of RNN called *Gated Recurrent Unit (GRU)* to learn a vector representation for each transformed propagation path.

CNN-Based Propagation Path Representation

We also use convolutional networks (CNN) to learn another vector representation.

2.8 Deep Diffusive Network Model

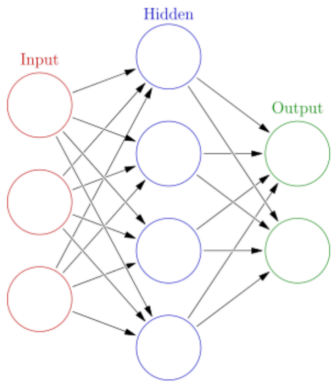
The credibility of news articles are highly correlated with their subjects and creators. The relationships among news articles, creators and subjects are illustrated with an example in Figure below,



For each creator, they can write multiple news articles, and each news article has only one creator. Each news article can belong to multiple subjects, and each subject can also have multiple news articles taking it as their main topics. To model the correlation among news articles, creators and subjects, we will introduce the deep diffusive network model as follow.

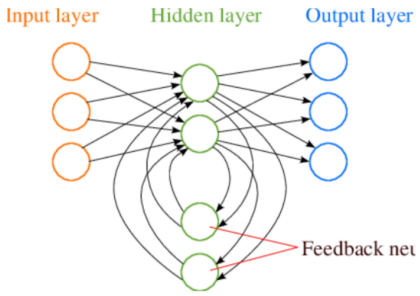
2.9 Artificial Neural Networks

Neural networks is in short a programming paradigm where computers are enabled to learn from observational data, and thereby increase in efficiency and accuracy over time.



An Artificial neural network (ANN) is a system which imitates how the biological nervous systems process information, such as the brain. The brain itself is a series of interconnected neurons, which each works on solving the same problem, and through the learning potential is able to work out a better solution given more input data and time. ANNs are mostly configured against a single application, where it can specialize its learning potential within a single topic and the more specialized it becomes, the better it becomes to find data that does not fit the model it creates, and outliers and other abnormalities are then more easily detected. This can, for instance, be used in tax analysis, where systems can process the normal tax forms, but the moment something is not within the thresholds it can notify a human to have a look at it.

The main difference between neural networks and conventional computing is that ANNs do not follow a set path of instructions to find a solution, but instead organically finds a solution and therefore can be unpredictable if not given the correct training and input data.



Neurons can be much more complicated than the ones stated above. They can have weighted inputs, where certain inputs take precedence over other and will fire if the total input if over a threshold. The networks come in many different forms. Feed-forward networks, always work in one direction, from input to output, and are mostly used in pattern recognition. Feedback, networks can have signals travel back and forward in the network, and contain loops. Because of this, the state of the entire network is always changing, and will only give an output when the system is in a stable state. More complex ANNs use neurons that are called perceptron, which are neurons with weighted inputs with some additional, fixed, pre-processing.

4 Related Works

This definition is widely adopted in recent studies.

- The fake news spreading plague: Was it preventable
- Automatic deception detection: Methods for finding fake news.
- A stylometric inquiry into hyper partisan and fake news.
- Fake news: A legal perspective.

Broader definitions of fake news focus on the either authenticity and intent of the news content. Some paper regard satire news as fake news since the content are false even though satire is often entertainment-oriented and reveals its own deceptiveness to the consumers.

- Fake news or truth? Using satirical cues to detect potentially misleading news.
- When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism.
- News verification by exploiting conflicting social viewpoints in microblogs.
- The impact of real news about fake news: Intertextual process and political satire.

Other literature directly treats deceptive news as fake news

- Deception detection for news: three types of fakes.

5 Open Issues and Future Works

In this section, we present some open issues in fake news detection and future research directions. Fake news detection on social media is a newly emerging research area, so we aim to point out promising research directions form a Data Mining perspectives, Detection and Mitigation, Linguistic Analysis and Machine Learning, and Network Model, etc. And the figure below better exemplifies the research direction in the future.

5.1 Data-oriented

Data-oriented fake news research is focusing on different kinds of data characteristic, such as: Data set, temporal, psychological.

- **Dataset perspective**, a promising direction is to create a comprehensive and large-scale fake news benchmark dataset, which can be used by researchers to facilitate further research in this area.
- **Temporal perspective**, fake news dissemination on social media demonstrates unique temporal patterns different from true news. Along this line, one interesting problem is to perform *early fake news detection*, which aims to give early alerts of fake news during the dissemination process. For example, this approach could look at only social media posts within some time delay of the original post as sources for news verification.
- **Psychological perspective**, different aspects of fake news have been qualitatively explored in the social psychology literature. It's worth to explore how to use data mining methods to validate and capture psychology intentions. For example, the echo chamber effect plays an important role for fake news spreading in social media. Then how to capture echo chamber effects and how to utilize the pattern for fake news detection in social media could be an interesting investigation. Moreover, *intention detection* from news data is promising but limited as most existing fake news research focus on detecting the authenticity but ignore the intent aspect of fake news. Intention detection is very challenging as the intention is often explicitly unavailable.

5.2 Feature-oriented

Feature-oriented fake news research aims to determine effective features for detecting fake news from multiple data sources. There are two major data resource: *news content* and *social context*.

- **News content perspective**: linguistic-based and visual-based techniques to extract features from text information.
    - ***Linguistic-based***: this feature has been widely study from NLP tasks, such as text classification, clustering, and specific application such as author identification, deception detection.

      But, the underlying characteristic of fake news have not been fully understood. Moreover, embedding techniques, such as word embedding and deep neural networks, are attracting much attention for textual feature extraction, and has the potential to learn better representation.

    - ***Visual-based***: this feature extracted from images are also shown to be important indicator for fake news. However, very important research has been done to exploit effective visual features, including traditional local and global features, and newly emerging deep network-based features.

      Recently, it has been shown that advanced tools can manipulated video footage of public figures, synthesize high quality videos. Thus, it becomes much more

challenging and important to differentiate real and fake visual content, and more advanced visual-based features are needed for this research.

- **Social context perspective**: user-based, post-based, and network-based features.
    - ***User-based***: mainly focus on general user profiles, rather than differentiating account types separately and extracting user-specific features.
    - ***Post-based***: convolutional neural network (CNN) can better capture people's opinions and reactions toward fake news. Image in social media post can also be utilized to better understand users' sentiments toward news events.
    - ***Network-based features***: extracted to represent how different types of network are constructed.

This is important to extent this preliminary work to explore to future research:

- How other networks can be constructed in terms of different aspects of relationships among relevant users and post.
- Network embedding is also considered as other advance method of network representations.

5.3 Model-oriented

Model-oriented fake news research opens the door to building more effective and practical models for fake news detection. Most previously mentioned approaches focus on:

- extracting various features and incorporating the features into supervised classification models below, and selecting the classifier that performs the best.
    - naïve Bayes
    - decision tree
    - logistic regression
    - k nearest neighbor (KNN)
    - support vector machines (SVM)

More research can be done to build more complex and effective models and to better utilize extracted features, such as

- **Aggregation methods**, combine different feature representations into a weighted form and optimize the feature weights.

- **Probabilistic methods**, predict a probabilistic distribution of class labels (i.e., fake news versus true news) by assuming a generative model that pulls from the same distribution as the original feature space.

(since fake news may commonly mix true statements with false claims, it may make more sense to predict the likelihood of fake news instead of producing a binary value)

- **Ensemble methods**, build a conjunction of several weak classifiers to learn a strong classifier that is more successful than any individual classifier alone; ensembles have been widely applied to various applications in the *machine learning* literature. It may be beneficial to build ensemble models as news content and social context features each have supplementary information that has the potential to boost fake news detection performance.

(Since, one of the major challenges for fake news detection is the fact that each feature, such as source credibility, news content style, or social response, has some limitations to directly predict fake news on its own)

- **Projection methods**, refer to approaches that lean projection functions to map between original feature spaces (e.g., news content features and social context features) and the latent feature spaces that may be more useful for classification.

(Since, fake news content or social context information may be noisy in the raw feature space)

5.4 Application-oriented

Application-oriented fake news research encompass research that goes into other areas beyond fake news detection. There are two major directions along these lines: *fake news diffusion* and *fake news intervention*.

- **Fake news diffusion**, characterizes the diffusion paths and patterns of fake news on social media sites. Some early re- search has shown that true information and misinformation follow different patterns when propagating in online social networks

  Similarly, the diffusion of fake news in social media demonstrates its own characteristics that need further investigation, such as *social dimensions*, *life cycle*, *spreader identification*, etc.

  - ***Social dimensions***, refer to the heterogeneity and weak dependency of social connections within different social communities. Users' perceptions of fake news pieces are highly affected by their like-minded friends in social media (i.e., echo chambers), while the degree differs along different social dimensions. Thus, it is worth exploring why and how different social dimensions play a role in spreading fake news in terms of different topics, such as political, education, sports, etc.

- ***Life cycle***, The fake news diffusion process also has different stages in terms of people's attentions and reactions as time goes by, resulting in a unique life cycle. Research has shown that breaking news and in-depth news demonstrate different life cycles in social media. Studying the life cycle of fake news will provide deeper understanding of how particular stories "go viral" from normal public discourse. Tracking the life cycle of fake news on social media requires recording essential trajectories of fake news diffusion in general, as well as further investigations of the process for specific fake news pieces, such as graph-based models and evolution-based models.

- ***Spreader***, identifying key spreaders of fake news is crucial to mitigate the diffusion scope in social media. Note that key spreaders can be categorized in two ways, i.e., *stance* and *authenticity*.

    - Along the *stance* dimensions, spreaders can either be
        - (i) *clarifiers*, who propose skeptical and opposing viewpoints towards fake news and try to clarify them;
        - or (ii) *persuaders*, who spread fake news with supporting opinions to persuade others to believe it. In this sense, it is important to explore how to detect clarifiers and persuaders and better use them to control the dissemination of fake news.

        - From an *authenticity* perspective, spreaders could be either human, bot, or cyborg. Social bots have been used to intentionally spread fake news in social media, which motivates further research to better characterize and detect malicious accounts designed for propaganda.

- **Fake news intervention**, which aims to reduce the effects of fake news by *proactive* intervention methods that minimize the spread scope or reactive intervention methods after fake news goes viral.
    - (i) remove malicious accounts that spread fake news or fake news itself to isolate it from future consumers;
    - (ii) *immunize* users with true news to change the belief of users that may already have been affected by fake news.

There is recent research that attempts to use content-based immunization and network-based immunization methods in misinformation intervention. One approach uses a multivariate Hawkes process to model both true news and fake news and mitigate the spreading of fake news in real-time.

- The aforementioned spreader detection techniques can also be applied to target certain users (e.g., persuaders) in social media to stop spreading fake news
- other users (e.g. clarifiers) to maximize the spread of corresponding true news.

6 Conclusion

The goal of this survey has been to comprehensively and extensively review, summarize, compare and evaluate the current research on fake news, which includes

1. The qualitative and quantitative analysis of fake news, as well as detection and intervention strategies for fake news from four perspectives: the false knowledge fake news communicates, its writing style, its propagation patterns, and its credibility;
2. Main fake news characteristics (authenticity, intention, and being news) that allow distinguishing it from other related concepts (e.g., misinformation, disinformation, or rumors);
3. Various news-related (e.g., headline, body-text, creator, and publisher) and social-related (e.g., comments, propagation paths and spreaders) information that can exploited to study fake news across its lifespan (being created, published, or propagated);
4. Feature-based and relation-based techniques for studying fake news; and
5. Available resources, e.g., fundamental theories, websites, tools, and platforms, to support fake news studies.

A summary and comparison of various perspectives to study fake news is provided in Table. The open issues and challenges are also presented in this survey with potential research tasks that can facilitate further development in fake news research.

	Knowledge-based	Style-based	Propagation-based	Credibility-based
Potential Research Task(s)	Fake news analysis and detection	Fake news analysis and detection	Fake news analysis, detection, and intervention	Fake news analysis, detection, and intervention
Fake News Stage(s) Studied	Creation, publication and propagation	Creation, publication and propagation	Propagation	Creation, publication and propagation
Information Utilized	News-related	News-related	Primarily social-related	News-related and social-related
Objective(s)	News Authenticity Evaluation	News Intention Evaluation	News Authenticity and Intention Evaluation	News Authenticity and Intention Evaluation
Techniques	Relation-based	Feature-based	Primarily Relation-based	Relation-based and Feature-based
Resources	Knowledge graphs, e.g., Knowledge Vault	Theories, e.g., reality monitoring; however, not many theories focus on fake news	Theories in Table 2	Theories in Table 2.
Related Topic(s)	Fact-checking	Deception analysis and detection	Epidemic modeling, rumor analysis and detection.	Clickbait analysis and detection, (review and Web) spam detection.

7 Reference

[] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[] Fake News Detection on Social Media: A Data Mining Perspective

[] Study Fake News via Network Analysis: Detection and Mitigation

[] A simple but tough-to-beat baseline for the Fake News Challenge stance detection task

[] Automated Fake News Detection Using Linguistic Analysis and Machine Learning

[] Automatic Detection of Fake News

[] Early Detection of Fake News on Social Media Through Propagation Path Classification with

[] Recurrent and Convolutional Networks

[] Automatic Online Fake News Detection Combining Content and Social Signals

[] Fake News Detection with Deep Diffusive Network Model

[] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data. AcM, 1247–1250.

[] Gary D Bond, Rebecka D Holman, Jamie-Ann L Eggert, Lassiter F Speller, Olivia N Garcia, Sasha C Mejia, Kohlby W Mcinnes, Eleny C Ceniceros, and Rebecca Rustige. 2017.

[] ‘Lyin’Ted’,‘Crooked Hillary’, and ‘Deceptive Donald’: Language of Lies in the 2016 US Presidential Debates. Applied Cognitive Psychology 31, 6 (2017), 668–677.

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

[] 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.

[] Chloé Braud and Anders Søgaard. 2017. Is writing style predictive of scientific fraud? arXiv preprint arXiv:1707.04095 (2017).

[] Michael T Braun and Lyn M Van Swol. 2016. Justifications offered, questions asked, and linguistic patterns in deceptive and truthful monetary interactions.

[] Group decision and negotiation 25, 3 (2016), 641–661.

[] Cody Buntain and Jennifer Golbeck. 2017. Automatically Identifying Fake News in Popular Twitter Threads. In Smart Cloud (SmartCloud), 2017 IEEE

International Conference on. IEEE, 208–215.

[] Chiyu Cai, Linjing Li, and Daniel Zeng. 2017. Detecting Social Bots by Jointly Modeling Deep Behavior and Content Information. In Proceedings of the

[] 2017 ACM on Conference on Information and Knowledge Management. ACM, 1995–1998.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending

[] language learning.. In AAAI, Vol. 5. Atlanta, 3.

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.

Tong Chen, Lin Wu, Xue Li, Jun Zhang, Hongzhi Yin, and Yang Wang. 2017. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. arXiv preprint arXiv:1704.05973 (2017).

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In CSCW: proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work, Vol. 2017. NIH Public Access, 1217.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014).

Peter Christen. 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 151–159.

Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE

Transactions on Dependable and Secure Computing 9, 6 (2012), 811–824.

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.

Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. Commun. ACM 54, 10 (2011), 66–71.

Aron Culotta and Andrew McCallum. 2005. Joint deduplication of multiple record types in relational data. In Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 257–258.

Douglas C Derrick, Thomas O Meservy, Jeffrey L Jenkins, Judee K Burgoon, and Jay F Nunamaker Jr. 2013. Detecting deceptive chat-based communication using typing behavior and message cues. ACM Transactions on Management Information Systems (TMIS) 4, 2 (2013), 9.

Morton Deutsch and Harold B Gerard. 1955. A study of normative and informational social influences upon individual judgment. The journal of abnormal and social psychology 51, 3 (1955), 629.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In International workshop on multiple classifier systems. Springer, 1–15.

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.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based
trust: Estimating the trustworthiness of web sources. Proceedings of the VLDB Endowment 8, 9 (2015), 938–949.

Zhicheng Dou, Ruihua Song, Xiaojie Yuan, and JiRong Wen. 2008. Are click-through data adequate for learning web search rankings?. In Proceedings of
the 17th ACM conference on Information and knowledge management. ACM, 73–82.

Norman R Draper and Harry Smith. 2014. Applied regression analysis. Vol. 326. John Wiley & Sons.

Nan Du, Yingyu Liang, Maria Balcan, and Le Song. 2014. Influence function learning in information diffusion networks. In International Conference on
Machine Learning. 2016–2024.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can Rumour Stance Alone Predict Veracity?. In Proceedings of the 27th
International Conference on Computational Linguistics. 3360–3370.