Natural Language Processing Research
Vol. **1(1-2)**, 2020, *pp.* 1–13
DOI: https://doi.org/10.2991/nlpr.d.200522.001; eISSN: 2666-0512
https://www.atlantis-press.com/journals/nlpr/

Review

Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective

Qi Su[1,*], Mingyu Wan[1,2], Xiaoqian Liu[1], Chu-Ren Huang[2]

[1] *Peking University, Beijing, China*
[2] *The Hong Kong Polytechnic University, Hong Kong, China*

ARTICLE INFO

Article History

Received 20 Oct 2019
Accepted 14 May 2020

Keywords

Misinformation detection
Information credibility
Feature representations
Modeling and predicting

ABSTRACT

The rise of misinformation online and offline reveals the erosion of long-standing institutional bulwarks against its propagation in the digitized era. Concerns over the problem are global and the impact is long-lasting. The past few decades have witnessed the critical role of misinformation detection in enhancing public trust and social stability. However, it remains a challenging problem for the Natural Language Processing community. This paper discusses the main issues of misinformation and its detection with a comprehensive review on representative works in terms of detection methods, feature representations, evaluation metrics and reference datasets. Advantages and disadvantages of the key techniques are also addressed with focuses on content-based analysis and predicative modeling. Alternative solutions to anti-misinformation imply a trend of hybrid multi-modal representation, multi-source data and multi-facet inference, e.g., leveraging the language complexity. In spite of decades' efforts, the dynamic and evolving nature of misrepresented information across different domains, languages, cultures and time spans determines the openness and uncertainty of this restless adventure in the future.

1. INTRODUCTION

Information is crucial for human's decision-making and has impact on life behaviors. Early information exchanges happened in interactive communications through daily conversations, or come from traditional media (e.g., books, newspapers, radio and television). Such information is more trustful as it is either self-vetted or controlled by authorities. Nowadays, people are exposed to massive information through a variety of sources (e.g., web pages, blogs, posts), especially with the popularity of Internet and social media platforms. The easiness of Internet access has caused the explosive growth of all sorts of misinformation, e.g., rumor, deception, hoaxes, fake news, spam opinion, which diffuses rapidly and uncontrollably in the human society. The erosion of misinformation to democracy, justice, and public trust becomes a global problem, which has gained an increasing number of research interests in its detection as well as the combat toward its propagation among a wide range of communities [1–8].

The inhibition of misinformation dissemination is never easy. One essential problem for anti-misinformation is the identification of information credibility. In the conventional journalism time, information credibility is established through esteemed publishers and the refereeing process. Nowadays, the credibility of information varies enormously, which can be true, false or has different degrees of reliability. Users of any kinds, prestigious or notorious, can freely post almost any information online and spread it without any cost. This should be alerted that rampant misinformation, especially maliciously fabricated news, poses great risks to the society, harms the public trust, misleads people's decision-making and may even lead to global tragedies. An example is the salt panic in China in 2011 which was caused by the rumor about salt radiation due to the Fukushima nuclear disaster. More examples include the recent consecutive cases of telecommunications scams, the medical scandal on the web, and the spread of false information about candidates of presidential election [9,10].

Various approaches, including human-crafted rules, traditional machine learning models, and neural networks, have been exploited to detect misinformation in an automatic way. With a broad definition, misinformation detection is the task of assessing the appropriateness (truthfulness/credibility/veracity/authenticity) of claims in a piece of information using a multidisciplinary approach. It could be investigated from various perspectives, such as Data Mining [11], Social Media Pragmatics [12,13], Linguistic Analysis [9], Psychological Experiments [14] and Natural Language Processing (NLP) [15]. Although efforts have been made for decades, misinformation detection remains a challenging problem due to limited datasets and lack of standard assessment of the diversified nature of misinformation. In this paper, we survey automated misinformation detection from the perspective of NLP with a comprehensive review of misinformation analysis, detection and inhibition by introducing the key methods, features, models and datasets for standard reference. It also elicits the challenges and opportunities with implications for future NLP research on this subject area.

* *Corresponding author. Email: sukia@pku.edu.cn*

1.1. Elaboration on Similar Concepts

Misinformation refers to misrepresented information in a macro aspect, including a series of fabricated, misleading, false, fake, deceptive or distorted information. It is usually created by information creators with malicious intentions for achieving certain purposes. As such, the credibility of the information is usually undermined. Under the common umbrella of conveying misrepresented information, it closely relates to several similar concepts, such as **fake news**, **rumor**, **deception**, **hoaxes**, **spam opinion** etc. Despite being similar, there exists salient differences among them in terms of the degrees of wrongness, the contexts of usage and the functions of serving for different propagation purposes. Below will address the main concepts of the several varieties of misinformation.

Deception is generally defined as an intentionally misleading statement [16], as a means of conceptualizing deceptive communication both implicitly [1] and explicitly [17]. A deceptive behavior normally shows the following two characteristics: (a) the deceiver transmitting a false message (while hiding the true information) and (b) the act being intentional. Notably, unintentional behavior that leads to an untrue belief, such as honest mistakes, or misremembering, is not considered as deception [2].

Fake news is differentiated by the content that mimics news media in form but not in editorial processes [18]. This definition emphasizes two main characteristics of fake news: the false content of the news and the lack of editorial norms and processes for credibility control. Fake news may be misleading or even harmful, especially when they are disconnected from their original sources and contexts. Fake news detection has been investigated for decades and remains a popular issue in NLP, but there is still no congruent definition of "fake news," which is sometimes interchangeably used with phony press releases, hoaxes, rumor and opinion spam.

Opinion spam, also called review spam, are fabricated reviews that range from self-promotions to false announcements of the reviewed product, to deliberately mislead consumers to buy or avoid the product. According to Shu *et al.* [11], deceptive opinion spam has two distinct variations: *hyper spam*, where unwarranted positive reviews are given to products in order to unfairly promote them, and *defaming spam*, which gives unjustified negative reviews to competing products in order to damage their reputations.

A **rumor** is defined as a piece of circulating information whose veracity status is yet to be verified at the time of spreading. The function of a rumor is to make sense of an ambiguous situation, and the truthfulness value could be true, false or unverified [11]. Different from fake news, which usually refers to public news events that can be verified as true or false, rumors may include long-term rumors, such as conspiracy theories, as well as short-term emerging rumors.

Although misinformation is usually created and propagated intentionally, it is difficult to detect the intention of information creators due to the insufficient information of such metadata in most publicly available datasets. The research scope will be focused on the measurement and detection of the veracity/credibility of information with license to existing NLP technologies and released datasets which can help verify the information as true or false or partially true in a certain scale.

1.2. Related Tasks

In misinformation detection, there are some related tasks which can facilitate the identification process and help improve the performance to a certain degree. These tasks include stance detection, abstractive summarization, fact checking, rumor detection and sentiment analysis.

Stance detection is the task of assessing whether a document supports or opposes a specific claim. It aims to assess the consistency between a document and a claim rather than the veracity of information, which can help search evidence from a document [19] and extract credibility features for misinformation detection. Recent research has found that misinformation tends to provoke controversies compared to facts [20,21], thus there can be many obviously opposing responses to misinformation during its propagation [3,22]. Therefore, stance detection of responses can serve as a complementary credibility feature for misinformation detection. For instance, Wu *et al.* [23] design a sifted multi-task learning method to selectively capture valuable shared features between stance detection task and misinformation detection, by exploring a selected sharing layer relying on gate mechanism and attention mechanism, which achieves the state-of-the-art performance on two public datasets (RumourEval and PHEME).

Abstractive summarization is also a relevant task that can be useful for facilitating misinformation detection. Specifically, the summarization model can be applied to identify the central claims of the input texts and serves as a feature extractor prior to misinformation detection. For example, Esmaeilzadeh *et al.* [24] use a text summarization model to first summarize an article and then input the summarized sequences into a RNN-based neural network to do misinformation detection. The experimental results are compared against the task using only the original texts, and finally demonstrate higher performance.

Fact checking is the task of assessing the truthfulness of claims especially made by public figures such as politicians [25]. Usually, there is no clear distinction between misinformation detection and fact checking since both of them aim to assess the truthfulness of claims, though misinformation detection usually focuses on certain pieces of information while fact checking is broader [26]. However, fact checking can also be a relevant task of misinformation detection when a piece of information contains claims that need to be verified as true or false.

Rumor detection is often confused with fake news detection, since rumor refers to a statement consisting of unverified information at the posting time. Rumor detection task is then defined as separating personal statements into rumor or nonrumor [27]. Thus, rumor detection can also serve as another relevant task of misinformation detection to first detect worth-checking statements prior to classifying the statement as true or false. This can help mitigate the impact that subjective opinions or feelings have on the selection of statements that need to be further verified.

Sentiment analysis is the task of extracting emotions from texts or user stances. The sentiment in the true and misrepresented information can be different, since publishers of misinformation focus more on the degree to impress the audience and the spreading speed of the information. Thus, misinformation typically either contains intense emotion which could easily resonate with the public, or

controversial statements aiming to evoke intense emotion among receivers. Thus, misinformation detection can also utilize emotion analysis through both the content and user comments. Guo *et al.* [28] propose a Emotion-based misinformation Detection framework to learn content- and comment-emotion representations for publishers and users respectively so as to exploit content and social emotions simultaneously for misinformation detection.

1.3. An Overview of the Survey

This survey aims to present a comprehensive review on studying misinformation in terms of its characteristics and detection methods. It first introduces the related concepts and highlights the significance of misinformation detection. It then uses a two-dimensional model to decompose this task: the internal dimension of descriptive analysis (i.e., the characterization of low-credibility information) and the external dimension of predictive modeling (i.e., the automatic detection of misinformation). In particular, the publicly available datasets and the state-of-the-art technologies are reviewed in terms of the detection approaches, feature representations and model construction. Finally, challenges of misinformation detection are summarized and new prospects are provided for future misinformation detection works.

2. DATASETS

Nowadays, information can be collected from diverse sources, such as social media websites, search engines and news agency homepages. Dataset construction is one of the major challenges for automatic misinformation detection due to the limitations on the availability and the quality of the data as well as the cost for annotations especially for supervised learning. Annotations of datasets for misinformation detection need to specify whether one piece of article, claim or statement is true or false based on the ground truth. Generally, annotations can be made through the following ways: Expert journalists, Fact-checking websites (e.g., PolitiFact, Snopes), Industry detectors and Crowd-sourcing workers. Depending on the content forms, datasets for misinformation detection can be categorized as containing short statements, posts on social network sites (SNSs) and entire articles. Table 1 shows a collection of benchmark datasets for misinformation detection.

2.1. Datasets with Short Statements

LIAR: This dataset is collected from fact-checking website PolitiFact through its API [29]. It is annotated with six fine-grained classes and comprises 12,836 annotated short statements reported during the year of 2007 to 2016 along with various information about the speaker. These short statements are sampled from various contexts, such as news releases, TV or radio interviews, campaign speeches, etc. In the dataset, each row of the data contains a short statement, a label of credibility, the subject, the context of the statement and 10 other columns corresponding to various information about the speaker, such as the speaker's statement history and party affiliation.

FEVER: This dataset provides related evidence to short claims for misinformation detection. It contains 185,445 claims collected from Wikipedia. Each claim is labeled as "Supported," "Refuted" or "Not Enough Info." Thus, based on FEVER, a detection system can predict the truthfulness of a claim with the evidence so as to achieve better performance. However, the type of facts and evidence from Wikipedia may exhibit some stylistic differences from those in real scenarios and cannot be fully applied to real-world data.

2.2. Datasets with Posts on SNSs

BuzzFeedNews: This dataset collects 2,282 posts published in Facebook from 9 news agencies during the 2016 U.S. election. Each claim in every post is fact-checked by 5 BuzzFeed journalists. This dataset is further enriched in [30] by adding the linked articles, attached media and relevant metadata. It contains 1,627 articles: 826 mainstream, 356 left-wing and 545 right-wing.

BuzzFace: Santia and Williams [31] extends the BuzzFeed dataset with the comments related to news articles on Facebook. It contains 2,263 news articles and 1.6 million comments.

PHEME: This dataset [21] is collected from Twitter conversation threads, including 6,425 Twitter threads and covering nine newsworthy events such as the Ferguson unrest, the shooting at Charlie Hebdo, etc. A conversation thread consists of a tweet making a true and false claim, and a series of replies. Thus, the dataset has different levels of annotations including the thread level and the tweet level. The annotation labels are true, false, unverified.

RumourEval: This dataset is similar to PHEME in terms of the data structure, covering content and annotation scheme. Similar to PHEME, the dataset contains Twitter conversation threads associated with different newsworthy events. It also has the same annotation labels of threads and tweets. However, RumourEval only contains 325 Twitter threads discussing rumors.

CREDBANK: This is a large scale crowd-sourced dataset of approximately 60 million tweets covering 96 days starting from

Table 1 Publicly available datasets for misinformation detection.

Dataset	Main Input	Data Size	Label	Annotation	Main Task
LIAR	Short claims	12,836	Six-class	PolitiFact	Fake news detection
FEVER	Short claims	185,445	Three-class	Trained annotators	Fact checking
BuzzFeedNews	Facebook post	2,282	Four-class	Journalists	Fake news detection
BuzzFace	Facebook post	2,263	Four-class	Journalists	Fake news detection
PHEME	Tweet	6,425 (threads)	Three-class	Journalists	Rumor detection
RumourEval	Tweet	325 (treads)	Three-class	Journalists	Rumor detection
CREDBANK	Tweet	60 million	Five-class	Crowd-sourcing workers	Fake news detection
BS Detector	Web Post	12,999	Three-class	BS Detector	Reliable detection
FakeNewsNet	News Articles	23,921	Fake or Real	Editors	Fake news detection
Fake or Real News	News Articles	7,800	Fake or Real	Media	Fake news detection

October 2015. The tweets in the dataset cover over 1,000 events, with each event assessed for credibility by 30 annotators from Amazon Mechanical Turk [32].

BS Detector: This dataset is collected from a browser extension called BS detector developed for checking information veracity. The dataset contains text and metadata from 244 websites and represents 12,999 posts in total from the past 30 days. It searches all links on a given webpage for references to unreliable sources by checking against a manually complied list of domains. The labels are the outputs of BS detector, rather than human annotators.

2.3. Datasets with Entire Articles

FakeNewsNet: This is an ongoing data collection project for fake news detection [11,33]. It consists of headlines and body texts of fake news articles from BuzzFeed and PolitiFact. It also collects information about social engagements of these articles from Twitter, such as the user-news relationships, user-user social networks and user profiles, etc.

Fake or Real News: This dataset is developed by George McIntire and the GitHub repository of the dataset includes around 7.8k news articles with equal distribution of fake and true news and half of the news comes from political domain.The fake news portion of this dataset is collected from Kaggle fake news dataset comprising 2016 USA election news. The true news portion is collected from media organizations such as the New York Times, WSJ, Bloomberg, NPR and the Guardian in the duration of 2015 and 2016.

These are representative publicly available datasets for misinformation detection in recent years. The instances collected in these datasets have been verified in terms of the truthfulness. However, there are still some domain limitations of these datasets. For example, the datasets containing posts on SNSs are limited to a small range of topics and are more frequently used for rumor detection. Instead, datasets consisting of news articles from various publishers [34] are good resources for fake news detection against traditional new media articles. It is also noteworthy that a dataset without aggregate labels is simply based on website source, which is more ore less a website classification task [15].

3. A TWO-DIMENSIONAL APPROACH

Researches and surveys about misinformation detection are manifold, ranging from analytical investigations to predictive modelling. These works can be generalized as utilizing a two-dimensional way of studying misinformation: 1) the **internal dimension** highlights the observation process of characterizing the intrinsic properties of misinformation in comparison to true information; 2) the **external dimension** highlights the detection process of predicting the fake types/degrees with the modelling of various information representations. The two dimensions are mutually defined and represent the two main streams of studies in misinformation detection, as illustrated in Figure 1 below.

3.1. Observation and Characterization

The observation process carries an analytic merit of uncovering the unique properties, content features, propagation patterns of misinformation compared with true information. Although misinformation creators may attempt to control what they are saying, "language leakage" occurs with certain verbal aspects that are hard to monitor such as frequencies and patterns of pronouns, conjunctions and negative emotion word usage [35]. The goal in the internal dimension is to look for such instances of leakage or, so called "predictive deception cues" found in the content of a message. Choosing appropriate characterizing properties is an important function in establishing information credibility.

For example, a study published on *Science* reveals the difference in terms of the diffusion patterns between false and true information and found that false information diffuses farther, faster, deeper and more broadly in social networks on media platforms than true information in all categories, especially for political fake news [36]. One explanation is that false information is more novel than true information, appealing users to be more likely to share. In the work of Newman *et al.* [35], they found that misrepresented information employs a higher proportion of negative emotion verbs (e.g., *hate*, *worthless*, *envy*), or cognitive complexity features, such as exclusive words (e.g., *without*, *except*, *but*) or motion words (e.g., *walk*, *move*, *go*) which effectively points to the deceptive behavior even when the liar deliberately avoids being detected. More linguistic observations can be found in Su [9] that low-credibility information tends

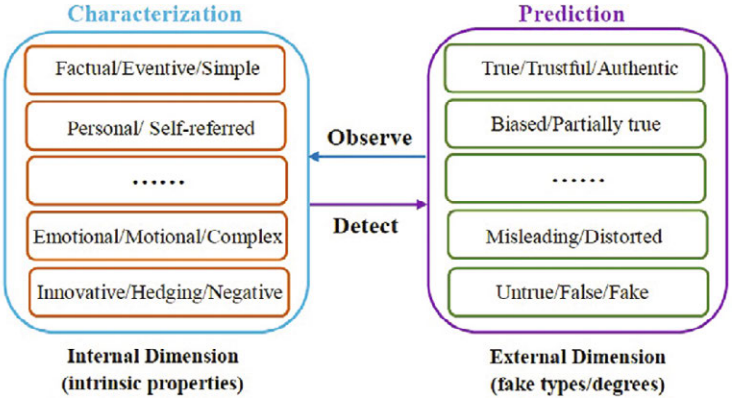

Figure 1 A two-dimensional approach of studying misinformation.

to be flagged by epistemic markers, impersonal views, stance markers, negations or discourse markers which demonstrate the uncertainty of information. These studies are informative in the analysis of misinformation characterizations and have provided important implications for more accurate misinformation detection.

3.1.1. Characterizing features

Real-world deceptive situations are high-stakes, where there is much to be gained or lost if deception succeeds or fails; it is hypothesized that these conditions are more likely to elicit salient cues to deception. Current studies tend to employ more text-based cues instead of nonverbal ones. For example, existing meta-analysis of verbal and nonverbal cues for deception deception shows that verbal cues are more reliable, robust and cost-effective than nonverbal cues [37]. However, it is noted that simple content-related n-grams and shallow part-of-speech (POS) tagging have proven insufficient for the detection task, often failing to account for important context information. Rather, these methods have been proven useful only when combined with more complex methods of analysis. Deep Syntax analysis, e.g., using Probabilistic Context Free Grammars (PCFG) has been proven particularly valuable in combination with n-gram methods [38]. In the following subsection, we will give a general review on the commonly used feature representations for misinformation detection.

TF-IDF: The most frequently adopted feature set is the vectorized Term Frequency-Inverse Document Frequency (TF-IDF) of n-grams. This is a weighted measure of how often a particular n-gram occurs in a document relative to how often n-gram occurs across all documents in a corpus. These patterns will be highly sensitive to the particular news cycle. Usually TF-IDF is calculated for each n-gram within each document and built for a sparse matrix of the resulting features. It can also be applied to any other types of features with the retain of frequency information.

Knowledge Representations: In addition to n-grams, information can be represented by the knowledge extracted from it. Knowledge here is usually defined as a set of triple tuple containing Subject, Predicate and Object (SPO) extracted from texts. For instance, the SPO tuple (DonaldTrump, Profession, President) is the extracted knowledge from the sentence *Donald Trump is the president of the U.S.*, and it can represent the content of the sentence. When detecting misinformation, knowledge within a Knowledge Graph (KG) or a Knowledge base (KB) is regarded as ground truth datasets containing massive manually-processed relational knowledge from the open Web. Specifically, the detection process is to evaluate the truthfulness of information by checking the compatibility of knowledge extracted from the content in comparison to the gold knowledge representations.

However, there are some deficiencies of the knowledge-based approach to detect misinformation. First, knowledge within a KG or a KB is far from complete and thus it demands further post-processing approaches for knowledge inference (e.g., [39,40]). Second, as the information online keeps changing instantly, the knowledge within a KG or a KB should also be timely updated. In this case, problems will reside in the construction of such resources when applying the knowledge-based approach to detect misinformation.

Domain-specific Features: Domain-specific features are specifically aligned to certain domains, such as quoted words, external links, number of graphs and the average length of graphs, etc. Moreover, other features can be specifically designed to capture the deceptive cues in writing styles to differentiate misinformation, such as lying detection features. Inspired by psychological theories such as Undeutsch hypothesis [41] for deception analysis, misinformation is considered as potentially different in writing style from true information. Thus, information content can also be represented by its writing style at multi-levels. Such representation then can be utilized as features to predict misinformation using machine learning methods.

Latent Representation: Via matrix or tensor factorization or neural network models in deep learning (e.g., CNN, RNN, LSTM, attention network, memory network, etc.), latent features of information content can be extracted automatically without hand-crafted features. Such representation methods save time and labor compared to feature engineering, but the selection or extraction of latent features is often driven by prior experience or techniques without sufficient theoretical basis. Therefore, it is challenging to understand and interpret the generated features learned by latent representation methods.

Linguistic Cues: Recently, researchers have become increasingly interested in the linguistic devices by which information quality is encoded. A variety of linguistic cues that might be associated with perceived information quality have been identified. Linguistic-based features are extracted from the text content in terms of document organizations from different levels, such as characters, words, sentences and documents. These cues act as an essential basis for the automatic detection of high-credibility information. It is therefore reasonable to exploit linguistic features that capture the different linguistic properties and sensational headlines to detect misinformation. To capture the intrinsic properties of misinformation, existing works mainly investigate common linguistic features including

- ***Lexical Features:*** Lexical features include character level and word-level features mainly based on the Linguistic Inquiry and Word Count (LIWC) (e.g. [30,42]). For word sequences, pre-trained word embedding vectors such as word2vec [43] and GloVe [44] are commonly used. Among lexical features, the simplest method of representing texts is the "bag-of-words" approach, which regard each word as a single, equally significant unit. For the bag-of-words approach, individual words or "n-grams" (multiword) frequencies are aggregated and analyzed to reveal cues of deception. Further tagging of words into respective lexical cues, e.g., parts of speech or "shallow syntax" [45] affective dimensions [37] or eventuality words [9] are all ways of providing frequency sets to reveal linguistic cues of deception.

  The simplicity of this representation also leads to its biggest shortcoming. In addition to relying exclusively on language, the method relies on isolated n-grams without utilizing context information. In this method, any resolution of ambiguous word sense remains nonexistent. Many deception detection researchers have found this method useful in tandem with in-depth complementary analysis.

- ***Syntactic Features:*** Syntactic features include grammatical and structural features, such as frequency of function words and phrases (i.e., constituents) or punctuations and POS tagging. Syntactic features can be further divided into shallow syntactic features and deep syntactic features [38]. Shallow syntactic features include the frequency of POS tags and punctuations; whereas the deep syntactic features investigates the frequency of productions such as rewritten rules. The rewritten rules of a sentence within an article can be obtained based on Probability Context Free Grammar (PCFG) parsing trees. Then the rewritten-rule statistics can be calculated based on TF-IDF [42]. For example, noun and verb phrases are in turn rewritten by their syntactic constituent parts [38]. Third-party tools, such as the Stanford Parser, AutoSlog-TS syntax analysis, assist in the automation. However, simply syntax analysis might not be sufficiently capable of identifying deception, and studies often combine this approach with other linguistic or network analysis techniques.
- ***Semantic Features:*** Inspired by fundamental theories initially developed in forensic- and social-psychology, semantic-level features investigate some psycho-linguistic attributes by analyzing sentiment, informality, diversity, subjectivity [46], cognitive and perceptual processes in the texts, as well as the quantity information such as the total or average number of characters, words, sentences and paragraphs. These attributes are extracted as high-level features to detect false information within the texts. Tools for the feature extraction can be LIWC for the sentiment, informality, cognitive and perceptual process analysis, and NLTK packages for diversity and quantity analysis.
- ***Discourse Features:*** Rhetorical approach is usually applied to extract features at discourse-level based on Rhetorical Structure Theory (RST), which is an analytic framework to examine the coherence of a story. Combined with Vector Space Model (VSM), RST is often used for misinformation detection [47]. It investigates the relative or standardized frequencies of rhetorical relationships among sentences within a piece of information. Through defining functional relations such as Circumstance, Evidence and Purpose of text units, RST can systematically identify the essential idea and analyze the characteristics of the input text. Misinformation is then identified according to its coherence and discourse structure.

  To explain the results by RST, VSM is used to convert documents/texts into vectors, which are compared to the center of true or fake information in high-dimensional RST space. Each dimension of the vector space indicates the number of rhetorical relations in the text.

  In addition to the rhetorical approach, Karimi and Tang [48] propose a Hierarchical Discourse-level Structure Framework (HDSF) for misinformation detection. HDSF automatically learns and constructs a discourse-level structure for fake or real articles in an end-to-end manner based on the dependency parsing at the sentence level. Specifically, the framework learns inter-discourse dependencies by utilizing BiLSTM network, and then constructs discourse dependency tree and finally learns the structural document-level representation to do misinformation classification. Further structural analysis also suggests that there are substantial differences in the hierarchical discourse-level structures between true and false information.

Complexity Cues: In real-word settings, misinformation creators may control well on strategic verbal cues (e.g., content words and topical keywords) to anti-detect the lying behavior. Therefore, strategic cues may become less effective for misinformation discrimination. By contrast, nonstrategic cues (e.g., some function words, emotional words and cognitive complexity words, syntactic patterns), as a reflection of the liars' emotional and cognitive states, can leak out of the liar's head in a near-unconscious way. This renders a higher chance of success in deception detection by tracing language complexity cues. Regarding the measurement of complexity in indexing lying statements, existing studies have almost been based on the same hypothesis: false statements are more complex than true statements as deception is assumed to be cognitively more demanding than telling the truth.

Pallotti [49] underlines the polysemy of complexity in literature and summarizes the different notions of complexity in this field by referring to three main meanings: 1) structural complexity, a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns; 2) cognitive complexity, having to do with the processing costs associated with linguistic structures and 3) developmental complexity, the order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition. The representation and computation of complexity in a piece of information seems a complex problem, which calls for a multi-facet investigation. For more discussion on computing complexity, works in addressing syntactic complexity can be found in [50,51].

Current studies of loaning complexity device mainly focus on a few lexical semantic categories, such as negative emotion verbs, or cognitive complexity features, or motion words [35,52]. A vital problem is that these are largely sparse and lexically biased for representing the full picture of a cognitively complicated event–language. Although many studies utilized n-grams or syntactic features for measuring structural complexity, the investigated complexity cues are far from representing language complexity, which is complicated and should not be represented from any single perspective.

3.2. Detection and Prediction

Another dimension of studying misinformation is the predicton of various degrees/types of misinformation with machine learning algorithms based on feature representations. This has become one of the major research interests in NLP. Existing techniques employed for misinformation detection are varied, from supervised or unsupervised learning settings. In recent decades, the main framework for misinformation detection takes it as a binary (or multi-) classification issue with a supervised method based on a set of pre-labelled corpus. This usually includes two phases: (i) feature conversion and (ii) model construction. The feature conversion phase aims to represent the information content and related auxiliary information in a formal mathematical structure, and the model construction phase builds machine learning models to differentiate misinformation from high-credibility information based on the feature representations [11].

Misinformation detection originally depends on the exploration of the information content, whereas in recent years there are also researches exploring the propagation patterns of information on social media platforms. Thus, current approaches to misinformation detection can be generally divided into two categories: **content-based** and **propagation-based** methods. While propagation-based approaches mainly rely on social contexts and social engagements of users such as the publisher-news relationship, news-user relationships, user social networks and user profiles, NLP techniques and methods are mainly applied to explore the content of misinformation. As such, this review focuses on content-based approaches to misinformation detection in recent advancement of the NLP community. Details of detection methods will be given in the following subsections.

3.2.1. Problem formulation

In this subsection, we present three basic categories of problem formulation for misinformation detection from the NLP perspective, which are detailed as follows:

Classification: A simplest case of misinformation detection can be defined as a binary classification problem to predict a piece of information as true or false. However, the binary classification formulation is not efficient in the cases that the information is partially real and partially fake. To address this problem, misinformation detection can also be formulated as a fine-grained multi-classification problem by adding additional classes to datasets. For example, LIAR consists of short political statements classified as pants-fire, false, barely-true, half-true, mostly-true and true. When using the datasets with multi-class labels, the expected outputs are multi-class labels and the labels are learned independently [29,34].

Regression: Misinformation detection can also be formulated as a regression task. For example, Nakashole and Mitchell [53] formulate the task by outputting a numeric score of truthfulness. With license to regression, the evaluation can be done by calculating the difference between the predicted scores and the ground truth scores, or using Pearson/Spearman Correlation tests. However, since the available datasets have discrete ground truth labels, the regression formulation becomes problematic because converting the discrete labels to numeric scores seems a challenging task.

Clustering: Misinformation detection is usually formulated as a supervised learning problem given an annotated dataset with labels. However, real-world data is more often without any labels. Semi-supervised and unsupervised methods are hence proposed to develop misinformation detection systems by formulating it as a clustering problem [4]. For instance, Guacho *et al.* [54] propose a semi-supervised method for content-based detection of misinformation via tensor embeddings. They initiate the work in representing collections of articles as multi-dimensional tensors, leveraging tensor decomposition to derive concise article embeddings that capture spatial/contextual information about each article. Yang *et al.* [55] propose an unsupervised method to infer the quality of information based on the users' credibility. They consider users' credibility as latent random variables which somehow reflect their opinions toward the authenticity of information.

Among the above three formulations, most existing approaches to misinformation detection rely on the supervised learning, which requires an annotated gold standard dataset to train a model. However, a reliable annotated dataset with labels of misinformation degrees is usually time-consuming and human-laboring. It often requires expert annotators to do careful analysis with the provision of additional evidence and contexts from authoritative sources. Thus, in real scenarios, it is more practical to learn semi-supervised or unsupervised models given limited or no labeled dataset.

3.2.2. Algorithms and models

Many studies in misinformation detection have employed machine learning algorithms. The major technique can be generalized as a binary or multiple classification task with the use of predicative modelling on features at multiple levels (see Section 3.1.1). In general, commonly adopted models of misinformation detection in NLP include two main categories as follows:

Statistical Models: The most commonly used statistical models in misinformation detection are Support Vector Machine (SVM) [56] and Naïve Bayes Classifier (NBC) [57]. The construction of the two models is quite easy and fast, yet with outstanding performance in most cases. When a mathematical model is sufficiently trained from pre-labelled examples in one of two categories, it can predict instances of future deception on the basis of numeric clustering and distances. The use of different clustering methods and distance functions between data points shape the accuracy of SVM, which invites new experimentation on the net effect of these variables. Naïve Bayes algorithms make classifications based on accumulated evidence of the correlation between a given variable (e.g., n-gram) and the other variables present in the model. In addition to SVM and NBC, there are many other frequently adopted statistical models for misinformation detection, including, e.g., Logistic Regression (LR), K-Nearest-Neighbourhood (KNN), Decision Tree and Random Forest Classifier (RFC). They have shown different strengths in predicting misinformation in relation to various feature representations.

For example, Gilda [58] adopted several multiple classification algorithms, i.e., SVMs, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees and Random Forests and found that TF-IDF of bi-grams fed into a Stochastic Gradient Descent model achieves the best performance (77.2% accuracy). Feng, Banerjee and Choi [38] utilized syntactic cues and achieved 85%-91% accuracy in deception related classification tasks by testing on online review corpora. On the basis of this work, Feng and Hirst [59] conducted a semantic analysis with "object:descriptor" pairs for locating semantic contradictions and got further improvement of detection accuracy. Rubin, Lukoianova and Tatiana [60] analyzed rhetorical structure using a VSM with similar success. Ciampaglia *et al.* [61] employed language pattern similarity networks with a pre-existing KB and achieved a superior performance over the state-of-the-art methods. These works in applying statistical models to misinformation detection tend to suggest the effectiveness of modelling complex and deep features compared to superficial features, such as bag-of-words. The comparison of the features, statistical models and performances in these studies is summarized in Table 2.

Table 2 Studies on using statistical models.

Study	Dataset	Feature	Model	Result
Feng *et al.* [38]	Online hotel review	Syntactic feature	SVM	91.0 Acc.
Feng and Hirst [59]	opspamv1.3	"object:descriptor" pairs	SVM^{perf}	91.3 Acc.
Rubin *et al.* [47]	36 personal stories	Rhetorical structure	SVC	67.0 Acc.
Ciampaglia *et al.* [61]	Three RDF datasets	Knowledge graph	KNN	65.0 AUC
Gilda [58]	Data from Signal Media	bi-grams TF-IDF	SGD	72.0 Acc.

Neural Networks: Early studies on misinformation detection rely on hand-crafted feature extraction to train machine learning models. However, the process of feature engineering can be time-consuming. With the advancement of neural networks, recent studies of misinformation detection have witnessed the critical role of deep learning methods in place of traditional machine learning models. For example, Rashkin *et al.* [34] added LIWC to a LSTM model which, however, performed unexpectedly worse than the baseline, while the case for the Naive Bayes model is improved. In addition to LSTM, recent research also proves the efficiency of RNN-based models to represent sequential posts and user engagements [62–64]. Many studies also adopt CNN-based models to capture local features of texts and images [65,66]. Furthermore, Generative Adversarial Networks (GANs) are often used to obtain fundamental features for texts across different topics or domains, which can be applied to misinformation detection in future research. However, one major problem of deep learning models is that they often require massive training data and substantial training time, also for parameter tuning, and the performance of deep learning models is usually difficult to interpret. As a summary to research in this field, we compare the features, deep learning models and top performances in some representative works in Table 3.

3.3. Evaluation Metrics

To evaluate the performance of algorithms for misinformation detection, various metrics have been used. In the NLP community, the most widely used metrics for misinformation detection include *Precision*, *Recall*, *F1*, and *Accuracy*. These metrics enable us to evaluate the performance of a classifier from different perspectives. Specifically, accuracy measures the similarity between predicted labels and the gold labels. Precision measures the fraction of all detected misinformation that are annotated as fake, addressing the important problem of identifying which information is fake. Recall is used to measure the sensitivity, or the fraction of annotated misrepresented articles that are predicted to be misinformation. F1 is used to combine precision and recall, which can provide an overall prediction evaluation. Note that for these metrics, with the range of 0 to 1, the higher the value, the better the performance.

In addition to the above metrics, the Receiver Operating Characteristics (ROC) curve provides a way of comparing the performance of classifiers by looking at the trade-off in the *False Positive Rate* (FPR) and the *True Positive Rate* (TPR). The ROC curve compares the performance of different classifiers by changing class distributions via a threshold.

Based on the ROC curve, the Area Under the Curve (AUC) value can be computed, which measures the overall performance of how likely the classifier is to rank the information higher than any true news, as shown below.

$$AUC = \frac{\sum(n_0 + n_1 + 1 - r_i) - n_0(n_0 + 1)/2}{n_0 n_1} \quad (1)$$

where r_i is the rank of ith false information piece and n_0 (n_1) is the number of false (true) information pieces. It is worth mentioning that *AUC* is more statistically consistent and more discriminating than accuracy, and it is usually applied in an imbalanced classification problem, where the number of ground truth fake article's and true articles have a very imbalanced distribution.

For an evaluation demonstration, Table 4 [1] summarizes some recent experiments using different datasets and models for misinformation detection with the measurement of Accuracy.

4. CHALLENGES AND PROSPECTS

Misinformation detection (esp. fake news detection) has attracted a lot of research interests in NLP with promising results in recent decades. However, it is still a challenging problem for models to automatically detect the authenticity of information given the diverse and dynamic nature of all sorts of misinformation online, and this makes existing detection algorithms ineffective or not applicable. First, misinformation is usually intentionally created to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on the verbal content. In this regard, some studies resort to auxiliary information, such as user social engagements and profiles on social media, to help make an inference. Second, exploiting auxiliary information is also problematic as users' social engagements with false information will produce data that is big, incomplete, unstructured and noisy. In this section, we bring together some open issues in misinformation detection with implications for some directions in future research.

[1] The differences on the Acc. between Yang *et al.* [55], Karimi *et al.* (2018) and Roy *et al.* (2018) is mainly due to the different proposed methods, model architecture designs and the experimental settings. For instance, Yang *et al.* [55] use an unsupervised method with a Bayesian network model and an efficient collapsed Gibbs sampling approach to infer the truths of news and the users' credibility without any labelled data. Karimi *et al.* (2018) propose a Multi-source Multi-class Fake news Detection framework MMFD, which combines automated feature extraction, multi-source fusion and automated degrees of fakeness detection into a coherent and interpretable model. Roy *et al.* (2018) develop various deep learning models for detecting fake news and classifying them into the pre-defined fine-grained categories. The representations are fed into a Multi-layer Perceptron Model (MLP) for the final classification.

Table 3 Studies on using deep learning models.

Study	Dataset	Feature	Model	Result
Rashkin *et al.* [34]	PolitiFact	Lexicons in LIWC	LSTM	57.0 F1 (2-class) 22.0 F1 (6-class)
Zhang *et al.* [64]	PolitiFact	Latent features	Deep diffusive network	63.0 F1 (2-class) 28.0 F1 (6-class)
Ma *et al.* [62]	Two Twitter datasets	Structural and textural properties	RNN	83.5 F1
Yang *et al.* [65]	A news dataset	Text and image information	CNN	93.0 F1
Liu and Wu [66]	Weibo	User characteristics	RNN	92.0 Acc.
	Twitter 15-16		CNN	85.0 Acc.

Table 4 Recent experiments on misinformation detection.

Paper	Dataset	Method	Acc.
Karimi and Tang [48]	FakeNewsNet	N-grams	72.37
	Fake or Real News	LIWC	70.26
		RST	67.68
		BiGRNN-CNN	77.06
		LSTM [w + s]	80.54
		LSTM [s]	73.63
		HDSF	**82.19**
Wu *et al.* [5]	PHEME	SVM	72.18
		CNN	59.23
		TE	65.22
		DeClarE	67.87
		MTL-LSTM	74.94
		TRNN	78.65
		Bayesian-DL	80.33
		sifted MTL	**81.27**
Qian *et al.* [67]	Weibo	LIWC	66.06
		POS-gram	74.77
		1-gram	84.76
		CNN	86.23
		TCNN	88.08
		TCNN-URG	**89.84**
Yang *et al.* [55]	LIAR	Major voting	58.6
		TruthFinder	63.4
		LTM	64.1
		CRH	63.9
		UFD	**75.9**
Karimi *et al.* [68]	LIAR	SVM	29.98
		RandomForests	27.01
		NN	29.12
		MMFD	**38.81**
Roy *et al.* [69]	LIAR	hybrid CNN	27.4
		hybrid LSTM	41.5
		Bi-LSTM	42.65
		CNN	42.89
		RNN-CNN	**44.87**

4.1. Address to the Content Issue

4.1.1. Identify check-worthy features

As content-only cues are insufficient and sometimes unreliable for misinformation detection, identifying check-worthy features may serve as an alternative way for improving the efficiency of misinformation detection. For instance, to identify whether a given topic or content is worth checking, analysis to the topic, website, domain, language, culture, etc., can be a potential research direction for enhancing misinformation detection. Additionally, other related tasks as mentioned in Section 1.2 can serve as facilitating mechanisms to help detect misinformation, such as applying summarization models to check the main idea of the information and using stance detection to find out the argument of the information creator, which are important cues for the classification of misinformation degrees.

4.1.2. Detect early stage propagation patterns

One of the challenging tasks for misinformation detection is to identify misinformation at early stages before its fast propagation. Therefore, detecting misinformation at an early stage of propagation can be a significant step to mitigate and intervene misinformation. For example, Ramezani *et al.* [70] propose a model considering earliness both in modeling and prediction. The proposed method utilizes RNN with a new loss function and a new stopping rule. First, the context of news is embedded with a class-specific text representation. Then, the model utilizes available public profile of users and speed of news diffusion for early labeling of misinformation. Their experimental results have demonstrated the effectiveness of their model which outperforms the competitive methods in term of accuracy while detecting in an earlier stage.

4.2. Address to the Auxiliary Issue

4.2.1. Incorporate multi-source data

Most previous research on misinformation detection mainly rely on the content as input, but recent studies have shown that incorporating additional information, such as speaker profiles [71] or social engagement data can further improve the accuracy of detection systems. A case study in Kirilin and Strube [72] shows that the attention model pays more attention to speaker's credibility than a statement of claim. Long *et al.* [71] added speaker profiles such as party affiliation, speaker title, location and credibility history into LSTM model and outperformed the state-of-the-art method by 14.5% in accuracy using a benchmark dataset. Moreover, social engagements data also proves to be effective for misinformation detection [11]. With the advancement of graph neural networks, recent studies also integrate multi-source data into a graph network [73,74]. By constructing a heterogeneous graph, the relationships among multi-source data about misinformation, such as the creator, the content and the corresponding subject/theme can be learned and updated by graph neural networks (i.e., Graph Convolutional Network [75], Graph Attention Network [76]). However, one problem is that relying judgments on speakers, creators, publishers or social networks may cause some risks. As Vlachos puts it, the most

dangerous misinformation comes from the sources we trust, and upgrading or downgrading specific sources will silence minorities' voice [77].

4.2.2. Multi-modal representation

Social media information often contains both text and visual content (e.g., image and videos), and each has both focused and complementary information. Therefore, for misinformation detection, it is necessary to use the multi-modal detection approach by integrating the text and visual information to assess the truthfulness of the information. Among the existing works, representative methods include attRNN [78], EANN [79] and MVAE [80].

For example, Jin *et al.* [78] introduced multi-modal information into fake news detection for the first time through deep neural networks. They proposed a RNN-based model (attRNN) with attention mechanism to integrate text and visual information. Experiments showed that this method is able to recognize misinformation that is difficult to discriminate by using a single modal information. Wang *et al.* [79] propose an end-to-end model based on adversarial networks. The motivation is that many current models learn event-related features that are difficult to migrate to newly emerged events. In this method, TextCNN is applied to extract semantic features of the text, and VGG-19 is used to extract semantic features of visual content. The multi-modal features are then concatenated to represent the content of misinformation and also achieved promising results. Dhruv *et al.* [80] argues that the simple concatenation of text and visual features is insufficient to fully express the interaction and relation between the two modal information. Therefore, they propose an encoding-decoding method to construct multi-modal feature representation. In this model, the concatenated features of texts and visual content are encoded as an intermediate expression, and a reconstructed loss is used to ensure that the encoded intermediate expression can be decoded back to the original state, and then the intermediate expression vector is used for misinformation detection.

Although many multi-modal resolutions have been sought, the current multi-modal approach to misinformation detection faces two major challenges. The first challenge is that high-quality annotated multi-modal misinformation datasets are in a scarce state. The other challenge is that despite building larger datasets, unsupervised or semi-supervised methods for misinformation detection should be developed when dealing with unlabeled data.

5. CONCLUSIONS

Misinformation has been created and propagated explosively with the popularity of Internet and social media platforms, which has been a major issue concerning the public and individuals. This survey has provided a comprehensive review, summarization and evaluation of recent research on misinformation detection from the perspective of NLP. Existing datasets, features, models, performances, challenges and prospects for misinformation detection are described in a two-dimensional way. The internal dimension highlights the analytical virtue of research on uncovering the salient characteristics of misinformation, e.g., fake news, rumor, deception and spam opinion, which facilitates the external dimension of studies of detecting misinformation with more discriminative features. Although efforts have been pursued for decades from various perspectives, misinformation detection remains a difficult task for the NLP community. The dynamic, evolving and muti-facet aspects of misinformation determines the challenges for its definition and identification; the existing datasets are usually domain-specific, mainly targeting the political news, and also show limitations in terms of data size and pre-labelling, which constrains the accuracy and reliability of misinformation detection systems. It is reasonable to believe that the future of misinformation detection may be directed to the construction of a more diversified dataset which accommodates multiple indications of source + modality + feature in one system.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORS' CONTRIBUTIONS

Qi Su shaped the framework of this article and provided sufficient information and guidance to the development of the main content. Mingyu Wan facilitated the writing of the manuscript, reorganized the structure and updated the major body of literature review. Xiaoqian Liu assisted the collection of literature and the draft of some previous works. Chu-Ren Huang provided critical feedback and suggestive support to this work. All authors contributed to the final manuscript.

Funding Statement

This work is supported by National Key R&D Program of China (no. 2019YFC1521200), the GRF grant (PolyU 156086/18H) and the Post-doctoral project (no. 4-ZZKE) at the Hong Kong Polytechnic University.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their valuable and constructional advices on the previous versions of this article; all remaining errors are our own.

REFERENCES

[1] M. Zuckerman, B.M. DePaulo, R. Rosenthal, Verbal and nonverbal communication of deception, Adv. Exp. Soc. Psychol. 14 (1981), 1–59.

[2] G. An, Literature Review for Deception Detection, PhD Thesis, The City University of New York, New York, NY, USA, 2015.

[3] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, A. Vlachos, Fake news stance detection using stacked ensemble of classifiers, in Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, Copenhagen, Denmark, 2017, pp. 80–83.

[4] V.L. Rubin, T. Vashchilko, Identification of truth and deception in text: application of vector space model to rhetorical structure

theory, in Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection, Avignon, France, 2012, pp. 97–106.
[5] L. Wu, F. Morstatter, K.M. Carley, H. Liu, Misinformation in social media: definition, manipulation, and detection, ACM SIGKDD Explor. Newslett. 21 (2019), 80–90.
[6] B. Guo, Y. Ding, L. Yao, Y. Liang, Z. Yu, The future of misinformation detection: new perspectives and trends, arXiv preprint arXiv:1909.03654, 2019. https://arxiv.org/abs/1909.03654
[7] F. Torabi Asr, M. Taboada, Big data and quality data for fake news and misinformation detection, Big Data Soc. 6 (2019), 1–14.
[8] P.S. Kulkarni, R.T. Aghayan, L. Huang, S. Gupta, Misinformation Detection in Online Content (US20200004882A1), US Patent App. 16/019,898, 2020. https://arxiv.org/abs/1909.03654
[9] Q. Su, Information quality: linguistic cues and automatic judgments, in: Z. J.-S. Chu-Ren Huang, B. Meisterernst (Eds.), The Routledge Handbook of Chinese Applied Linguistics, chap. 32, Taylor & Francis: Routledge, London, UK, 2018.
[10] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on twitter during the 2016 us presidential election, Science. 363 (2019), 374–378.
[11] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: a data mining perspective, ACM SIGKDD Exp. Newslett. 19 (2017), 22–36.
[12] S.B. Parikh, P.K. Atrey, Media-rich fake news detection: a survey, in Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, USA, 2018, pp. 436–441.
[13] C. Shao, G.L. Ciampaglia, O. Varol, K. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nat. Commun. 9 (2018), 1–9.
[14] T.O. Meservy, M.L. Jensen, J. Kruse, J.K. Burgoon, D.P. Nunamaker, J.F. Twitchell, G. Tsechpenakis, D.N. Metaxas, Deception detection through automatic, unobtrusive analysis of nonverbal behavior, IEEE Intell. Syst. 20 (2005), 36–43.
[15] R. Oshikawa, J. Qian, W.Y. Wang, A survey on natural language processing for fake news detection, arXiv preprint arXiv:1811.00770, 2018. https://arxiv.org/abs/1811.00770
[16] D. Galasinski, The Language of Deception: a Discourse Analytical Study, SAGE Publications, Thousand Oaks, CA, USA, 2018. https://www.jstor.org/stable/24047513
[17] P. Ekman, K.G. Heider, The universality of a contempt expression: areplication, Motiv. Emot. 12 (1988), 303–308.
[18] D.M.J. Lazer, M.A. Baum, Y. Benkler, A.J. Berinsky, K.M. Greenhill, F. Menczer, *et al.*, The science of fake news, Science. 359 (2018), 1094–1096.
[19] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 2016.
[20] M. Mendoza, B. Poblete, C. Castillo, Twitter under crisis: can we trust what we RT?, in Social Media Analytics, SOMA, KDD Workshop, Washington, DC, USA, 2010.
[21] A. Zubiaga, M. Liakata, R. Procter, G.W.S. Hoi, P. Tolmie, Analysing how people orient to and spread rumours in social media by looking at conversational threads, PLoS ONE. 11 (2016), 1–33.
[22] S. Dungs, A. Aker, N. Fuhr, K. Bontcheva, Can rumour stance alone predict veracity?, in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 2018, pp. 3360–3370. https://www.aclweb.org/anthology/C18-1284
[23] L. Wu, Y. Rao, H. Jin, A. Nazir, L. Sun, Different absorption from the same sharing: sifted multi-task learning for fake news detection, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019.
[24] S. Esmaeilzadeh, G.X. Peh, A. Xu, Neural abstractive text summarization and fake news detection, arXivpreprintarXiv:1904.0078, 2019. https://arxiv.org/abs/1904.00788
[25] A. Vlachos, S. Riedel, Fact checking: task definition and dataset construction, in Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, 2014, pp. 18–22.
[26] J. Thorne, A. Vlachos, Automated fact checking: task formulations, methods and future directions, 2018. arXiv preprint arXiv:1806.07687. https://arxiv.org/abs/1806.07687
[27] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, Detection and resolution of rumours in social media: a survey, ACM Comput. Surv. 51 (2018), 32.1–32.36.
[28] C. Guo, J. Cao, X. Zhang, K. Shu, M. Yu, Exploiting emotions for fake news detection on social media, arXiv preprint arXiv: 1903.01728, 2019. https://www.researchgate.net/publication/331543936_Exploiting_Emotions_for_Fake_News_Detection_on_Social_Media
[29] W.Y. Wang, 'Liar, Liar Pants on Fire': a new benchmark dataset for fake news detection, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017.
[30] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, arXiv preprint arXiv:1702.05638, 2017.
[31] G.C. Santia, J.R. Williams, Buzzface: a news veracity dataset with facebook user commentary and egos, in Proceedings of the Twelfth International Conference on Web and Social Media, Palo Alto, CA, USA, 2018, pp. 531–540. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/viewPaper/17825
[32] T. Mitra, E. Gilbert, CREDBANK: alarge-scale social media corpus with associated credibility annotations, in Proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, England, 2015, pp. 258–267. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10582
[33] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, 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. https://www.liebertpub.com/doi/abs/10.1089/big.2020.0062
[34] H. Rashkin, E. Choi, J.Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: analyzing language in fake news and political fact-checking, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017, pp. 2931–2937.
[35] M.L. Newman, J.W. Pennebaker, D.S. Berry, J.M. Richards, Lying words: predicting deception from linguistic styles, Pers. Soc. Psychol. Bull. 29 (2003), 665–675.

[36] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science. 359 (2018), 1146–1151.
[37] A. Vrij, Detecting Lies and Deceit: Pitfalls and Opportunities, John Wiley and Sons, New York, NY, USA, 2008. https://psycnet.apa.org/record/2008-01237-000
[38] S. Feng, R. Banerjee, Y. Choi, Syntactic stylometry for deception detection, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Association for Computational Linguistics, Jeju Island, Korea, 2012, vol. 2, pp. 171–175.
[39] B. Shi, T. Weninger, Discriminative predicate path mining for fact checking in knowledge graphs, Knowl. Based Syst. 104 (2016), 123–133.
[40] R. Trivedi, B. Sisman, J. Ma, C. Faloutsos, H. Zha, X.L. Dong, Linknbed: multi-graph representation learning with entity linkage, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018.
[41] U. Undeutsch, Beurteilung der glaubhaftigkeit von aussagen, Handbuch der psychologie. 11 (1967), 26–181.
[42] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, arXiv preprint arXiv:1708.07104, 2017. https://arxiv.org/abs/1708.07104
[43] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013. https://arxiv.org/abs/1301.3781
[44] J. Pennington, R. Socher, C. Manning, GloVe: global vectors for word representation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.
[45] J. Hancock, D. Markowitz, Linguistic traces of a scientific fraud: the case of diederik stapel, PLoS ONE. 9 (2014), e105937.
[46] M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky, Linguistic models for analyzing and detecting biased language, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 2013, pp. 1650–1659. https://www.aclweb.org/anthology/P13-1162
[47] V.L. Rubin, N.J. Conroy, Y. Chen, Towards news verification: deception detection methods for news discourse, in Hawaii International Conference on System Sciences, Kauai, HI, USA, 2015. https://0-works.bepress.com.library.simmons.edu/victoriarubin/6/
[48] H. Karimi, J. Tang, Learning hierarchical discourse-level structure for fake news detection, arXiv preprint arXiv:1903.07389, 2019.
[49] G. Pallotti, A simple view of linguistic complexity, Second Lang. Res. 31 (2015), 117–134.
[50] M. Wan, C. Fang, A re-examination of syntactic complexity by investigating the inter-structural variations of adverbial clauses across speech and writing, in Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong, China, 2018. https://www.aclweb.org/anthology/Y18-1082
[51] M. Wan, A.C. Fang, C.-R. Huang, The discriminativeness of internal syntactic representations in automatic genre classification, J. Quant. Linguist. 23 (2019), 1–34.
[52] Y. Zhou, Q. Su, Exploring distinction between deception and truth in chinese: an analysis based on linguistic features, Int. J. Knowl. Lang. Process. 9 (2018), 1–16. http://www.ijklp.org/archives/vol9no1/index.html
[53] N. Nakashole, T.M. Mitchell, Language-aware truth assessment of fact candidates, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 2014, pp. 1009–1019.
[54] G.B. Guacho, S. Abdali, N. Shah, E.E. Papalexakis, Semi-supervised content-based detection of misinformation via tensor embeddings, in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, Barcelona, Spain, 2018, pp. 322–325.
[55] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, H. Liu, Unsupervised fake news detection on social media: a generative approach, in Proceedings of 33rd AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 2019.
[56] H. Zhang, Z. Fan, J. Zeng, Q. Liu, An improving deception detection method in computer-mediated communication, J. Netw. 7 (2012), 1811.
[57] S. Oraby, L. Reed, R. Compton, E. Riloff, M. Walker, S. Whittaker, And that's a fact: distinguishing factual and emotional argumentation in online dialogue, in Proceedings of the 2nd Workshop on Argumentation Mining, Denver, CO, USA, 2017, pp. 116–126. https://arxiv.org/abs/1709.05295
[58] S. Gilda, Evaluating machine learning algorithms for fake news detection, in Proceedings of the 2017 IEEE 15th Student Conference on Research and Development (SCOReD), Putrajaya, Malaysia, 2017, pp. 110–115.
[59] V.W. Feng, G. Hirst, Detecting deceptive opinions with profile compatibility, in Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, 2013, pp. 338–346. https://www.aclweb.org/anthology/I13-1039
[60] V.L. Rubin, T. Lukoianova, Truth and deception at the rhetorical structure level, J. Assoc. Inf. Sci. Technol. 66 (2015), 905–917.
[61] G.L. Ciampaglia, P. Shiralkar, L.M. Rocha, J. Bollen, F. Menczer, A. Flammini, Computational fact checking from knowledge networks, PLoS ONE. 10 (2015), e0128193.
[62] J. Ma, W. Gao, K.-F. Wong, Rumor detection on twitter with tree-structured recursive neural networks, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 1980–1989.
[63] N. Ruchansky, S. Seo, Y. Liu, CSI: ahybrid deep model for fake news detection, in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, Singapore, 2017, pp. 797–806.
[64] J. Zhang, L. Cui, Y. Fu, F.B. Gouza, Fake news detection with deep diffusive network model, arXiv preprint arXiv:1805.08751, 2018. https://arxiv.org/abs/1805.08751
[65] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, P.S. Yu, TI-CNN: convolutional neural networks for fake news detection, arXiv preprint arXiv:1806.00749, 2018. https://arxiv.org/abs/1806.00749
[66] Y. Liu, Y.-F.B. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA 2018, pp. 354–361. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16826
[67] F. Qian, C. Gong, K. Sharma, Y. Liu, Neural user response generator: fake news detection with collective user intelligence, in International Joint Conferences on Artificial Intelligence (IJCAI), 2018, pp. 3834–3840.

[68] H. Karimi, P. Roy, S. Saba-Sadiya, J. Tang, Multi-source multi-class fake news detection,in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1546–1557.

[69] A. Roy, K. Basak, A. Ekbal, P. Bhattacharyya, A deep ensemble framework for fake news detection and classification, 2018. arXiv preprint arXiv:1811.04670.

[70] M. Ramezani, M. Rafiei, S. Omranpour, H.R. Rabiee, News labeling as early as possible: real or fake?, in International Conference on Advances in Social Networks Analysis and Mining, Vancouver, Canada, 2019.

[71] Y. Long, Q. Lu, R. Xiang, M. Li, C.-R. Huang, Fake news detection through multi-perspective speaker profiles, in Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 2017, pp. 252–256. https://www.aclweb.org/anthology/I17-2043

[72] A. Kirilin, M. Strube, Exploiting a speakers credibility to detect fake news, in Proceedings of Data Science, Journalism and Media workshop at KDD (DSJM18), London, UK, 2018. https://drive.google.com/file/d/1zEbxEfMfZn-2frsU2bVScvZM8dsDS3hZ/view

[73] V. Vaibhav, R.M. Annasamy, E.H. Hovy, Do sentence interactions matter? Leveraging sentence level representations for fake news classification, in Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), Hong Kong, China, 2019.

[74] Y. Ren, J. Zhang, HGAT: hierarchical graph attention network for fake news detection, arXiv preprint arXiv:2002.04397, 2020. https://arxiv.org/abs/2002.04397

[75] T. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907, 2016. https://arxiv.org/abs/1609.02907

[76] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903, 2017. https://arxiv.org/abs/1710.10903

[77] L. Graves, Understanding the promise and limits of automated fact-checking, Factsheet. 2 (2018), 1–8. https://ora.ox.ac.uk/objects/uuid:f321ff43-05f0-4430-b978-f5f517b73b9b

[78] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in Proceedings of the 25th ACM International Conference on Multimedia, ACM, Mountain View, CA, USA, 2017, pp. 795–816,

[79] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, 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, London, UK, 2018, pp. 849–857.

[80] D. Khattar, J.S. Goud, M. Gupta, V. Varma, MVAE: multimodal variational autoencoder for fake news detection, in Proceedings of the 2019 World Wide Web Conference, ACM, 2019, pp. 2915–2921.