Fact Checking and Fake news detection.

A survey of existing methods.

Ayushi Patel
University of Illinois,
Urbana-Champaign
ayuship2@illinois.edu

Sogol Bazargani
University of Illinois,
Urbana-Champaign
sogolb2@illinois.edu

Sushmita Azad
University of Illinois,
Urbana-Champaign
sazad2@illinois.edu

ABSTRACT

With the growing popularity of social media and the increased digitization in all fields, it has become of utmost importance to be able to detect fake news and articles on social media. A large number of people get influenced by false information and rumors which are known to spread faster than facts, owing to them being more riveting. To filter out rumors and fake news articles and to prevent their spread, multiple methods have been proposed that do so in various ways, ranging from neural networks to differentiate between linguistic patterns, to fact checkers which try to find evidence for or against the news article. In this survey, we provide a comprehensive and structured analysis of various fake news detection algorithms proposed in the literature.

KEYWORDS

Fake news, Fact Checking

1 INTRODUCTION

In today's world, information is being both produced, consumed at breakneck speed. One of the pitfalls of the highly open and digitized model of information with the internet is that most conventional fact checks are absent. People make decisions with the information they have at hand, both in the private and public sphere of life. Decisions taken on the basis of factually incorrect information can be detrimental across the private and public sphere of society. Taking for example the popular & widely believed rumour that vaccines cause autism[42]: at the personal level - a child could be exposed to deadly diseases because their parent believes this rumour. At a societal level, having many children sick from preventable diseases (like Measles) could take up invaluable hospital space, leading to a delay in treatment for other patients, thereby making the health care system and setup an issue during the next polls. Exciting and exaggerated claims make for good 'viral' news[3], and the burden of verification inevitably falls upon the reader, who cannot be realistically expected to investigate every claim he encounters. Now, more than ever, is a time when automating the fact checking process will be beneficial not just to individual readers, but democracy as a whole. There are 2 main steps in fact checking. First is 'Claim Detection'. This involves checking if a sentence constitutes a claim worth checking. The second is to actually verify a claim for its truth factor[15]. In this survey we cover papers and research in both of these areas, as detection is a necessary precursor to verification. However, the main focus of this survey is the claim verification problem. Main Features of this survey:

- **Design:** This survey has been designed as mini summaries that show the progress of rumour fake news detection in several categories, and covers all the broad categories of methods for the same.
- **Comparison:** We compare the results of most of the methods within a category provided they have tested on the same datasets and conditions.
- **Coverage:** We attempt to attain good coverage of state of the art rumour detection methods by reviewing most works from top conferences in over the last 4 years.

2 CHALLENGES

The number of fact checking organisations and bodies pales in comparison to the number of news outlets producing both information and misinformation. Attempting to verify this large body of unlabeled data comes with a set of challenges that we enumerate here.

(1) Different methods have been used on different dataset structures. This makes the generalizability of the result more difficult to compute for a fair evaluation between methods.
(2) Different languages require different parsing for both detection and verification.[9]
(3) News producers have lost control over what is published, obscure unpredictable algorithms controlled by social media giants decide what people see or don't[4]. Deciding on a 'source of truth' is a big challenge as decision on what constitutes 'reliable' is oftentimes subjective.[26]
(4) Deciding what constitutes a claim that is worth checking is also subjective, and plays a big part in ClaimDetection.
(5) Given the variety of ways in which a sentence can be expressed, a key task is to automate the process of piecing several disjoint sentences together to form a 'claim', without training the model on the entire article.
(6) Fake news is written to deliberately misguide the reader, making it difficult to detect on the basis of only the article. Secondary information regd. users interactions with fake news is necessary, this data is incomplete, noisy and unstructured.
(7) To verify a user-driven claim, the pre-existing sources in the factchecking database may not contain the necessary facts, and we may need to scrape the internet for relevant documents. However, the authenticity of these documents will need to be further verified, leading down a potentially infinite rabbit-hole.

, ,

(8) While the majority of research relates to verifying the text of the article/post, other information such as images, audio and video are available to the user, embedded in many forms, and which is challenging to verify, and we don't see as much work being done in this area[29].

3 METHODS

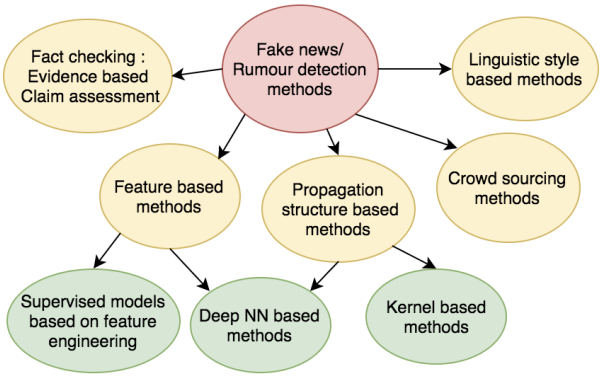

Figure 1: Feature Extraction Can be Performed on the Above Contexts

We broadly divide the methods into 5 main categories and list the existing methods under each category as shown in Figure 1.

3.1 Fact Checking: Claim Assessment by Finding Evidence

Credibility Assessment of Textual Claims on the Web [30]
CATC (Popat et al., 2016) aims to provide an assessment of a user given claim as true or false, along with explanations or evidences. They first automatically find sources in news and social media, extract features based on linguistic patterns in these sources and features based on the reliability of the underlying web sources. The way an article is written- i.e. objectively or subjectively, is an important factor for assigning the confidence of a claim and hence the linguistic features of the stance are used for article classification, which judges whether the article supports or refutes a claim.
They train a distantly supervised classifier for assessing the credibility of a claim (i.e., true or fake) where they attach the available claim labels to all articles reporting on the claim. For inference, they use the joint interaction between linguistic and reliability based features of the web sources. Results in Table 1.

Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media [31]
WTL (Popat et al., 2017) is an improvement over CATC[30] by the same authors. It first extracts stances for a claim by searching for relevant snippets or stances from the web and from fact databases which are created from fact checking websites. It takes into play the stance of the article along with article linguistics and source reliability, which happens to be an important feature, missing in the prior work. These factors are used to train the credibility assessment models: Content-aware Assessment that is based on Distant Supervision and Conditional Random Field (CRF) and Trend-aware Assessment where they exploit the trend of reporting articles over time and revise the assessment with new incoming evidence. The method also handles data imbalance as hoax debunking websites usually have more false claims than true. Results in Table 1.

ClaimVerif: A Real-time Claim Verification System Using the Web and Fact Databases[48]
ClaimVerif (Zhi et al., 2017) is a real-time claim verification system which classifies a user given claim and provides a credibility assessment and justifications for the claim by providing supporting evidences or stances for the given claim. Similar to WTL[31] it extracts stances or evidences by searching the web and finding relevant snippets using an embedding-based method to capture semantic similarity between the claim and candidate snippets. The next step is stance classification. A two-step classifier classifies the opinion of the related articles using high-quality editorial review articles and web documents. First the stance classifier is trained based on the snippets extracted from the review articles in the fact database. The classifier outputs two scores for each snippet, the score supporting the claim, and the score of snippet refuting the claim. These two scores along with the features from the search engine extracted articles are used as input features for the next step of article classification. ClaimVerif also implements a reweighting module to incorporate credibility of the sources. This trained model was used for determining the stance in both Snopes and Wikipedia datasets. Results in Table 1.

	Stance classifier	Claim Credibility
CATC [30]	-	71.96
WTL [31]	76.69	81.39
ClaimVerif [48]	83.62	85.25

Table 1: Accuracy comparison for stance classification and claim credibility assessment on Wiki and Snopes datasets

ClaimBuster [7] (Hassan et al., 2017) is an end to end, fact checking system which uses Machine learning, NLP and database querying to techniques to aid in fact checking given live claims. Steps ClaimBuster follows: a claim monitor which interfaces with various data sources, a claim spotter that finds claims that need to be verified, a claim matcher which check if the claim has already been verified or refuted among a repository of fact checked claims and if not, then a claim checker collects supporting or debunking evidences as done in previous methods and reports results to the user.

CredEye: A Credibility Lens for Analyzing and Explaining Misinformation[32]
Based on the authors' prior work [31], CredEye (Popat et al., 2018) is a system for automatic credibility assessment. It captures an article's linguistic styles, stance towards a claim and the source reliability to classify a claim, with explanations. It is different from the previous methods as it provides an explanation for the claim in

terms of the language style of snippets, which serves as evidence or counter evidence. It also provides feature level explanations for the credibility assessment.

Claim detection: before claim assessment [15] [9]
With the huge influx of real-time data, one of the important steps in determining fake news is first checking if the post or claim is even worthy of being checked. Rumour detection methods cannot be applied on the entire data that comes in as they tend to be more complex. ClaimRank (Jaradat et al., 2018) [9] aims to facilitate rumour detection by first determining checkworthy claims. It supports both Arabic and English sentences and is trained on 9 annotated fact-checking organizations. 'Towards Automated Factchecking' [15] (Konstantinovskiy et al., 2018) develops an annotation schema and then designs a sentence embedding based classification approach which outperforms ClaimBuster and ClaimRank.

3.2 Feature-Based Methods

Feature extraction is a necessary pre-requisite for model construction. All the models covered in this survey extract features from the contexts shown in Figure 2.

3.2.1 *Supervised Models based on feature engineering.*

Information Credibility on Twitter: DTC [1]
(Castillo et al., 2011) is one of the early works on credibility assessment of news propagated through twitter. They use supervised machine learning based methods to analyse and classify user posts as credible or not credible. For this purpose they extract multiple features from the post such as: Message-based features like content, sentiment, hashtags. User-based features like user activity, interests, followers. Propagation-based features such as depth of the re-tweet tree and tweet impact. And Topic-based features which are aggregates computed from the previous feature sets. After using ML models on these features they achieved the best classification results using Decision trees (DTC) and then pruned the feature set to extract the best features. Results in Table 2.

Rumor has it: Identifying Misinformation in Microblogs [34]
One of the early studies of fake news detection on twitter is done by (Qazvinian et al., 2011). The authors analyzed three different categories of characteristics and proved their effectiveness on detecting misinformation on twitter: 1- content based features like lexical patterns and part of speech patterns, 2- Network-based Features which focus on user behaviour on twitter, and 3- Twitter Specific Memes which are hash-tags and URLs extracted from twitter memes. Later the log-likelihood ratio for each feature is calculated against positive and negative training models, and finally the combined ratio is used for classification. Their retrieval model achieved the Mean Average Precision of nearly 95%.

Automatic Detection of Rumor on Sina Weibo: SVM-RBF [45]
(Yang et al., 2012) [45] examines an extended set of features that can be extracted from the microblogs and then used to train a true/false classifier. Along with the previously proposed features they consider new features such as client based features- the client program used to post the microblog and location based features. They then train an SVM-RBF classifier for detecting rumours on Sina Weibo. Results in Table 2.

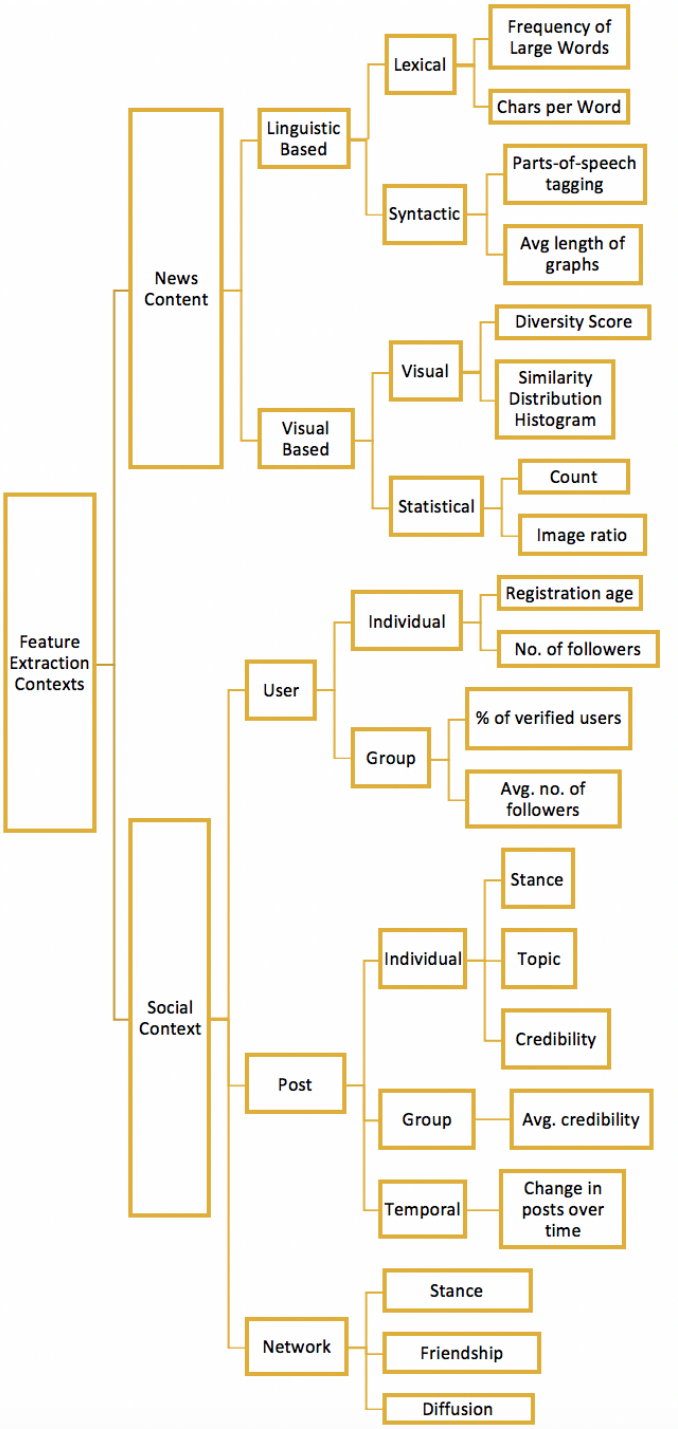

Figure 2: Feature Extraction Can be Performed on the Above Contexts

Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy [6]
(Gupta et al., 2013) [6] studies the fake image detection on twitter during hurricane Sandy. The authors use user-related features like number of friends, number of followers etc. and tweet-based features like length of the tweet, number of words and tweet metadata (number of hashtags, retweets etc.) to train a decision tree classifier and achieved an accuracy of 97% on predicting fake images from real. Furthermore the results showed that tweet based features were more effective in distinguishing fake tweets from real compared to user-based features.

Prominent Features of Rumor Propagation in Online Social Media: RFC [17]
In this paper (Kwon et al., 2013) the features to train the Supervised machine learning model are based on three aspects of rumour propagation or information diffusion: temporal, structural and linguistic. For capturing the temporal characteristics they propose a new method: 'Periodic External Shocks', that describes the periodic bursts in the extracted time series, which observes the number of tweets posted over time. For the structural characteristics, properties of the propagation subgraph are used. For linguistic characteristics, content and sentiment is captured. A classifier based on decision tree, random forest and SVM is trained on these features and the RF classifier (RFC) performs the best. Results demonstrated in Table 2.

	Accuracy	F1 score
DTRank [47]	0.644	0.656
SVM-RBF [45]	0.722	0.769
DTC [1]	0.731	0.740
RFC [17]	0.772	0.801
SVM-TS [22]	0.808	0.834

(a) Twitter dataset

	Accuracy	F1 score
DTRank [47]	0.732	0.737
SVM-RBF [45]	0.818	0.819
DTC [1]	0.831	0.831
RFC [17]	0.849	0.864
SVM-TS [22]	0.857	0.861

(b) Sina Weibo dataset

Table 2: Accuracy and F1 score comparison for top Supervised Machine Learning based methods.

Method	DTC [1]	SVM-RBF[45]	RRD [18]
Accuracy	0.745	0.747	0.897

Table 3: Accuracy comparison between RRD [18], SVM-RBF[45], and DTC [1] on Twitter dataset

Real-time Rumor Debunking on Twitter: RRD [18]
(Liu et al., 2015) studies the problem of real-time rumor detection on Twitter. Since Tweets are short and contain relatively limited information, the authors define the concept of rumor as an event which consists of several microblogs. They continuously monitor the event and dynamically update their model based on newly obtained information. Furthermore, in their feature based model they take advantage of the crowd response along with the post, propagation, content etc., and use these features to train an SVM based classifier. The results are demonstrated in Table 3.

Detect Rumors Using Time Series of Social Context Information on Microblogging Websites: SVM-TS [22]
(Ma et al., 2015)[22] builds upon the previous methods by incorporating the temporal characteristics of the features based on the time series of the rumours' life cycle. The method expands the feature vocabulary by combining the existing content, user and propagation based features with their temporal characteristics, by using time series modelling to incorporate the changes in the social context information over time. They propose a novel time series model-Dynamic Series-Time Structure (DSTS) to capture these variations during the spread of the rumour. Using these DSTS-based features they train an SVM classifier which is also good for early rumour detection and it's performance improves over time as it captures rich variation patterns of features from the time series. They applied the method on events posted on Twitter and Sina Weibo. Results demonstrated in Table 2.

Enquiring minds: Early detection of rumors in social media from enquiry posts: DT-Rank [47]
(Zhao et al., 2015) [47] proposes a real-time detection method to identify trending rumours which is based on the idea that whenever there is a rumour, people will express skepticism about it. Using regular expressions, the system selects tweets that contain skeptical enquiries (signal tweets) which are then clustered together and collapsed into one statement. All other related (non signal) tweets are used to create an extended cluster. Then based on statistical features of the clusters, they train the decision tree classifier using these features and finally rank them in order of likelihood of being rumours. Results demonstrated in Table 2.

Leveraging Joint Interactions for Credibility Analysis in News Communities [26]
The authors(Mukherjee et al., 2015) have developed a probabilistic graphical model, CCRF (Continuous Conditional Random Field) to perform credibility analysis of news. They choose to combine the influence of various factors involved in the determination of veracity of a news article like format of news, political stance, writing style etc, using them all as moderate signals to arrive at a strong signal for information believability. Factors captured by the model are: 1) Language, objectivity and credibility 2) Expertise, format and viewpoint. 3) User expertise and review scores. An inference method for the CCRF leverages interplay between the above factors to jointly both predict veracity scores and rank articles sources by trustworthiness. Cliques are formed based on an article, source, users and reviews, with no. of cliques equal to the no. of news articles, which are then weighted, with features inside a clique being local interactions, while global weights indicate shared information between cliques. MeanSquaredError (MSE) comparison of the CCRF

model against baseline methods provides a 19.5% MSE reduction over the best of other aggregated models.

Hawkes Processes for Continuous Time Sequence Classification: an Application to Rumour Stance Classification in Twitter [20]
This paper(Lukasik et al., 2016) proposes using Hawkes Processes (Hawkes, 1971), commonly used for modelling information diffusion in social media, for the task of rumour stance classification. The model assumes that the occurrence of a tweet will influence the rate at which future tweets will arrive. The frequency of tweets generated by them is determined by an underlying intensity function which considers the influence from past tweets. The difference in the pattern of user activity over time with respect to a rumor and a non-rumor is captured using Hawkes Processes based on which a classifier is built.

Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes [16]
(Kumar et al., 2016) focuses on hoax detection on Wikipedia. First the authors study how a fake article impacts Wikipedia by considering 1- how long the article survives before being detected as hoax 2- how many page views it receives and 3- how many pages over the web refer to the fake article. Then they identify the characteristics of a successful hoax on wikipedia by comparing it to both failed hoaxes and legitimate articles. The captured features fall into four main categories 1- appearance based, which are directly visible to a reader, containing wiki-link density, article length, web link density etc, 2- network features which consider high order network linkage by computing the Ego-network clustering coefficient for each article 3- Support features which pays attention to the prior references to the article, and the time and the creator of the first reference 4- Edit features which consider the edit histories of article creators. Then the obtained features are used to train a random forest classifier which classifies wiki articles with an AUC/ROC of 98% while humans achieve the accuracy of 63% in discerning hoaxes.

Information Credibility Evaluation on Social Media[44]
(Wu et al., 2016) establishes a system: Network Information Credibility Evaluation (NICE), which serves as a repository for verified rumours on Sina Weibo and can automatically classify unverified real-time information generated by users on social media. Their algorithm learns a representation for posts, based on content, time, user and comment information. Then a logistic regression classifier is trained on this representation and classified rumours are stored into the existing repository of verified rumours.

Fake vs. Real vs. Satire [8, 36]
Another line of research in fake news detection focuses on identifying fake news from satire. In (Rubin et al., 2016), authors study the satirical news detection problem by developing 5 predictive features namely Absurdity, Humor, Grammar, Negative Affect, and Punctuation. They used an SVM based classifier and achieved a precision of 90% in their experiments. In (Horne et al., 2017) experiments on three datasets showed that occasionally fake news is a lot more similar to satire than real news. This makes it harder to distinguish fake news from satire. The paper extracts linguistic and content based features from fake, real and satirical news articles and trains a classifier to distinguish between the three. Findings show that certain features like long titles and repetitive content in the article body are the most useful in detecting fake news from real news and satire.

A Stylometric Inquiry into Hyperpartisan and Fake News [33]
(Potthast et al., 2017) shows how we can detect the likelihood of a news-article being fake by analysis of its writing style, and reveals a high correlation (97%) between highly politically one sided (hyperpartisan) articles and fake news. The portions of their research that are of interest to us are seeing if style based fake news detection is a reasonable aim, and whether we can differentiate it from satirical articles. Features are extracted based on characters, stop words, parts of speech, ratio of words to links (both internal and external), count of paragraphs. They use a method called UnMasking, which has been traditionally used to detect if articles are written by the same author, to bucket articles into leftwing vs rightwing vs mainstream. Classification error curves show steeper decrease for hyperpartisan and fake news articles,both forms of extremism seem to share similar writing styles. Importantly, it also finds that satire style is very distinct from hyperpartisan news, ensuring that humour wont be weeded out by a style based detector.

Exploiting Tri-Relationship for Fake News Detection [38]
This paper(Shu et al., 2017) suggests that a users social media data and his or her social engagements towards a news article provide valuable data for the detection of fake news on social media. Users prefer to believe a news that conforms to their personal stances and is approced by like minded users. The paper proposes a Tri-Relationship Fake News detection framework (TriFN), a semi-supervised detection framework that exploits the correlations between publisher bias, news stance, and related user engagements (both user-user and user-news engagements) simultaneously.

3.2.2 *Neural Network Based Methods.*

Detecting Rumors from Microblogs with Recurrent Neural Networks [21]
The above supervised ML based methods use handcrafted features obtained from the tedious and labor intensive task of feature engineering. Variations of contextual information of relevant posts over time are captured by extracting features sets over different timesteps. In this paper (Ma et al., 2016) the authors propose a model based on Recurrent neural networks (RNNs) for learning the hidden states by using the variation of aggregated information over time. For every event we wish to classify as a rumour or not, the input to the RNN (GRU or LSTM) is the sequence of all posts relevant to the event over a time series. The hidden states capture the latent features and the output is the classification label. For further improvement, more hidden and embedding layers are added. The method is compared to the 5 previous methods mentioned above on a new Twitter and Sina Weibo dataset. Results demonstrated in Table 5 and 6

CSI: A Hybrid Deep Model for Fake News Detection [37]

The authors(Ruchansky et al., 2017) propose a model called CSI which stands for Capture, Score and Integrate, each of which is a module in the network. It is based on the article text, it's response and the users promoting the article. The Capture module uses an RNN to capture temporal properties of user activity-text and response. The Source module learns source characteristics based on user behavior and finally the Integrate module integrates the two modules to learn a classifier for fake news as shown in Figure 3.

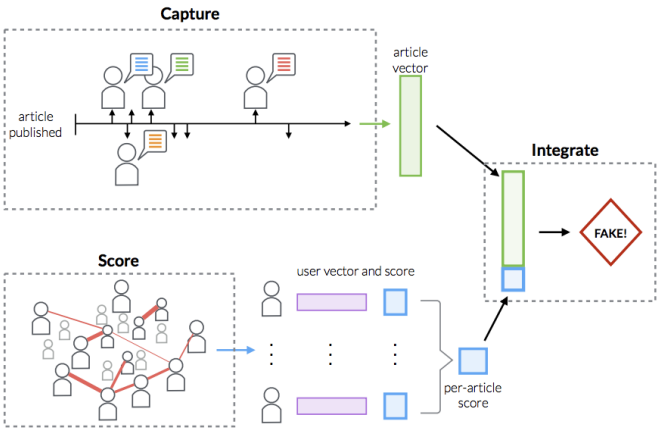

Figure 3: Csi: intuition [37]

Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection [41]
In this paper the authors(Wang et al., 2017) create LIAR: a new, publicly available dataset for fake news detection. LIAR is a 12.8K statement large manually labeled short statement dataset which provides a detailed analysis report and links to source documents for each case. It is an order of magnitude larger than previous public fake news datasets and contains useful metadata. They investigate automatic fake news detection methods based on surface-level linguistic patterns. They also propose a novel, hybrid convolutional neural network(CNN) to integrate speaker related metadata with text to enhance the performance of fake news detection. The method does better than one with the neural network trained without using just the text data on their LIAR dataset. Results demonstrated in Table 4.

Fake News Detection Through Multi-Perspective Speaker Profiles [19]
(Long et al., 2017) proposes an attention based hybrid LSTM model for fake news detection, which incorporates speaker profiles such as party affiliation, speaker title, location and credit history in two ways. The first LSTM includes speaker profiles and news article topic information as attention factors while the second LSTM inputs profiles to obtain vector representations of speakers. The two representations are then concatenated in the soft-max function for classification. Evaluations are performed using the LIAR dataset by Wang (2017). Results demonstrated in Table 4.

	With Attn	Without Attn
CNN-Wang [41]	0.247	N/A
CNN-WangP [41]	0.274	N/A
Base LSTM [19]	0.245	0.255
LSTM+profiles [19]	0.407	0.415

Table 4: Comparing [41] and [19] on classification using the Liar dataset by Wang

Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection: CallAtRumors [2]
(Chen et al., 2017) study the problem of debunking fake news diffusion at early stages. To deal with this problem, authors propose a novel deep attention based RNN, which automatically learns temporal embeddings for sequential posts and eliminates the need for manual feature engineering. They experimented their model on Twitter and Sina Weibo datasets. According to Table 5 CallAtRumors outperforms the aforementioned baselines and detects rumors substantially earlier than its competitors.

Method	Precision	Recall	F-measure
DT-Rank [47]	71.50%	63.41%	0.6721
SVM-TS [22]	76.33%	77.92%	0.7712
GRU-NN [21]	80.87%	82.97%	0.8191
CallAtRumors [2]	88.63%	85.71%	0.8715

Table 5: Performance Comparision on the Twitter dataset

Fake News Detection with Deep Diffusive Network Model [46]
(Zhang et al., 2018) introduces FakeDetector, which attempts to infer credibility labels for news articles as well as both authors and topics. Identifying trustworthiness of authors and topics is of great significance, as it can assist us to detect a large number of fake news originating from a malicious creator. This framework uses bag-of-words and RNN prediction models to learn latent embeddings of news article, news author and news topic. Furthermore, FakeDetector proposes a deep diffusive network in order to model the fusion of different heterogeneous information in a social network. The aforementioned deep diffusive model consists of 2 primary elements - representation feature learning and credibility label inference. Authors have conducted extensive experiments on the Politifact dataset which contains tweets from the official PolitiFact twitter account along with the fact check articles about the mentioned tweet posts from the PolitiFact website. Their obtained results show that this framework can achieve a bi-class accuracy of 0.65 and F1 score of 0.8.

Multi-Source Multi-Class Fake News Detection [13]
(Karimi et al., 2018) propose a MMFD (Multi-source Multi-class Fake News Detection) framework which deals with the problem of fake news detection from three perspectives:

(1) Since fake news is naturally diverse in content, topic and style, the authors use a deep neural network as their automatic feature extraction framework.

(2) As fake news can be mixed with legitimate news to mislead the readers, this paper considers degrees of fakeness which also makes the problem harder.
(3) Since large amount of information related to fake articles is spreaded over different sources with different degrees of reliability, authors propose an attention-based model to integrate information from multiple sources.

Conducted experiments on LIAR dataset against 5 baselines namely Basic-SVM, Basic Random Forest, Basic NN and Wand [41] show that the MMFD framework outperforms other baselines achieving highest accuracy in the presence of 4 sources.

Neural User Response Generator: Fake News Detection with Collective User Intelligence [35]
One of the major challenges in fake news detection area is the exponential propagation of fake news in early stages. This paper(Qian et al., 2018) addresses this challenge by focusing on early detection of misinformation. Due to the fact that user feedback and spreading patterns are not available at early propagation stages, this research only considers the fake news article text.
The proposed method called TCNN-URG consists of 2 parts namely "Two level Convolutional Neural Network"and "User Response generator". TCNN learns representations for sentences and words, consequently, it captures the semantic information of the article text. In addition authors leverage the historical user feedback to previous articles and propose a generative model of user feedback or URG which can be used to simulate user responses for new articles. More specifically URG uses a Conditional Variational Autoencoder (CVAE) underneath, which is trained to generate responses for a given article text. Authors conducted experiments on Weibo and a self-collected Twitter dataset, and compared TCNN-URG's performance against 4 baselines namely LIWC, POS-gram, 1-gram and CNN. Obtained results show that TCNN-URG achieves an accuracy of 88.83% on Twitter dataset and 89.84% on Weibo dataset, outperforming all other baselines.

Detect Rumor and Stance Jointly by Neural Multi-task Learning [24]
Rumour detection tries to determine the veracity of a given claim. One of the methods used for the task is stance classification on the opinions expressed towards the possible rumour. Previous works treat the task of stance classification and rumor detection as separate tasks. This paper (Ma et al., 2018) uses multi-task learning to unify the dependent tasks and trains both jointly, using weight sharing, to extract features common to both tasks. Each task can separately learn its task-specific features. Owing to the strong interconnection between the rumor detection and stance classification, both are improved by this multi-task learning approach.

Rumor Detection with Hierarchical Social Attention Network [5]
(Guo et al., 2018) propose a hierarchical neural network to detect rumors by leveraging hierarchical representations at multiple levels and social contexts. The network is a hierarchical bidirectional LSTM for representation learning where the social contexts are introduced into the network through attention, such that important semantic information is incorporated for a more robust rumor detection.

3.3 Propagation-Based Methods

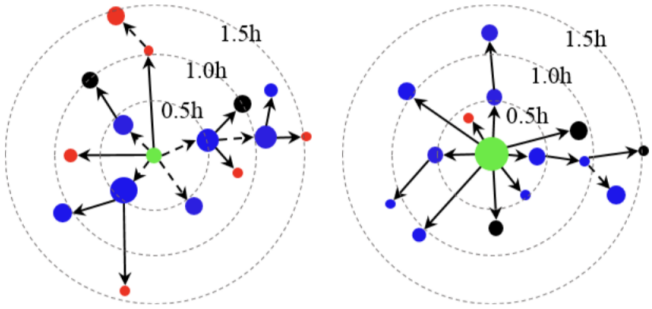

Figure 4: Propagation structure of rumor vs non-rumor via kernel learning [23]

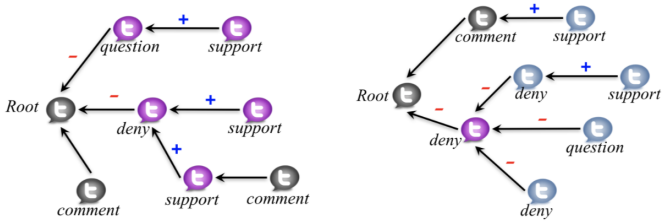

Figure 5: Propagation structure of rumor vs non-rumor via tree-structured recursive neural networks [25]

One of the most important features for rumour detection is the propagation structure of the rumour diffusion process which is key in differentiating rumors and non-rumors. The feature based methods described above: Information Credibility on Twitter, Automatic Detection of Rumor on Sina Weibo, Prominent Features of Rumor Propagation in Online Social Media, Detect Rumors Using Time Series of Social Context Information on Microblogging Websites, and Enquiring minds, also fall under this category. Other methods based on the propagation structure include but are not limited to the following:

Epidemiological modeling of news and rumors on twitter [11]
This paper (Jin et al., 2013) models the propagation of false information over twitter by using an epidemiologically-based population model called SEIZ. According to SEIZ, nodes in a network fall into 4 categories: susceptible(S), exposed(E), infected(I), and skeptical(Z). Individuals can transition between these categories with probabilities pre-estimated from the data. Conceptually it is easily comprehended that initially all nodes are in S state, and whenever each node receives a piece of information, they can move to Infected,

skeptical or exposed states. Real true and false Twitter examples were used to train the SEIZ model. Authors later define a ratio as the sum of the effective transition rates entering state E (from S) to the sum of the transition rates exiting E (to I). This ratio can be used for rumor detection. Their experiments showed that the obtained ratio is much less for rumors compared to legitimate news.

3.3.1 *Kernel Based Methods.*

False Rumors Detection on Sina Weibo by Propagation Structures: SVM-HK [43]
(Wu et al.,2015) study the problem of fake news detection on Sina Weibo using a hybrid kernel SVM classifier which combines a random walk graph kernel with a normal RBF kernel. They extracted three categories of features for their classifiers namely message-based features , user-based features (extracted from the message and its creator), and repost-based features. Results demonstrated in Table 6.

Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning: SVM:TK [23]
(Ma et al., 2017) tries to make use of the news propagation structure for differentiating rumors from non-rumors. The method captures observations such as that rumors are first posted by low-impact users and then popular users join, while non-rumours are posted by popular users and spread by general users as in Figure 4. They propose a kernel based method: Propagation Tree Kernel (PTK), where the post diffusion is first represented as a propagation tree and the kernel function is defined as the similarity between two trees- based on common subtrees, linguistic and temporal patterns. Finally the tree kernel function incorporated into a supervised Machine learning framework which uses a kernel based SVM classifier. Results demonstrated in Table 6.

3.3.2 *Neural Network Based Methods.*

Rumor Detection on Twitter with Tree-structured Recursive Neural Networks [25]
In (Ma et al., 2018), the authors propose two tree-based recursive neural network models: bottom-up and top-down RvNNs for the purpose of classifying rumours. The tree structure in the network conforms to the propagation pattern of tweets or claims. The model captures both content and structural semantics while learning the property that when a post denies a false rumour, it gains tremendous support by it's children and when a post denies a true rumour, it is denied by it's children as in Figure 5. These rumor indicative signals are enhanced by recursively aggregating signals from different branches. In bottom-up RvNN, after aggregating from bottom to up, the root node's state represents the entire tree and this representation is used in classification. In top-down RvNN, the learnt representations are embedded into the leaf nodes' states and then aggregated to use for classification. They test on the Twitter datasets released by Ma et al. (2017), namely Twitter15 and Twitter16, which respectively contain 1,381 and 1,181 propagation trees. Results demonstrated in Table 6.

		Accuracy	FR F1	TR F1
[47]	DTRank	0.409	0.311	0.364
[1]	DTC	0.454	0.355	0.317
[17]	RFC	0.565	0.422	0.401
[22]	SVM-TS	0.544	0.472	0.404
[21]	GRU-RNN	0.641	0.634	0.688
[43]	SVM-HK	0.493	0.439	0.342
[23]	SVM-TK	0.667	0.669	0.772
[25]	BU-RvNN	0.708	0.728	0.759
	TD-RvNN	0.723	0.758	0.821

(a) Twitter15 dataset

		Accuracy	FR F1	TR F1
[47]	DTRank	0.414	0.273	0.630
[1]	DTC	0.465	0.393	0.419
[17]	RFC	0.585	0.415	0.547
[22]	SVM-TS	0.574	0.420	0.571
[21]	GRU-RNN	0.633	0.715	0.577
[43]	SVM-HK	0.511	0.434	0.473
[23]	SVM-TK	0.662	0.623	0.783
[25]	BU-RvNN	0.718	0.712	0.779
	TD-RvNN	0.737	0.743	0.835

(b) Twitter16 dataset

Table 6: Accuracy and F1 score comparison for [21], [43], [23] and [25] against previous methods. FR- false rumours, TR- true rumours

News Credibility Evaluation on Microblog with a Hierarchical Propagation Model [12]
This paper focuses on veracity within microblogs, which have come to be one of the primary mediums for news consumption nowadays. (Jin et al., 2014) present a multi-layer hierarchical propagation network, capturing the event, subevents, messages and social connections within news events. It can represent the event from different scales and disclose a new set of features.

3.4 Methods Using Crowd Sourcing

These methods try to use the power of the crowds to aid them in automating the fact checking process. Some do it by leveraging the fact that a percentage of people who encounter suspicious news will try to verify it, and tracking those inquiries can help choose the claims or article to investigate. Other involve people in decision of subjective matters such as integrity of a source, using their trust in a source as weightage in the model.

Enquiring minds: Early detection of rumors in social media from enquiry posts [47]
The paper proposes a real-time detection method to identify trending rumours which is based on the idea that whenever there is a rumour, people will express skepticism about it. Using regular expressions, the system selects tweets that contain skeptical enquiries (signal tweets) which are then clustered together and collapsed into one statement. All other related (non signal) tweets are used to create an extended cluster. Then based on statistical features of the

clusters, they train classifiers using these features and finally rank them in order of likelihood of being rumours.

Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation [14]
Recently Facebook provided a flagging tool for users which enables them to mark a posts as fake. By leveraging these signals, highly voted posts can be sent to a third-party fact-checking organization for fake news identification. Since the number of posts flagged by users is relatively high, several approaches emerged to select a much smaller subset of posts with the goal of sending to a fact-checking organizations. This paper is one of the most recent works in this area, which approaches the problem by developing an online algorithm, called Curb which "Curb" s the number of posts sent to an expert, by solving a stochastic optimal control problem. This research captures users flagging activity by leveraging a framework of marked temporal point processes. They experimented their model on both real and fake posts and re-shares gathered from Twitter and Weibo, and compared the performance of their algorithm with an unrealistic variant of CURB- Oracle and 3 baselines namely "Flag Ratio", "Flag Sum", and "Exposure". Their results demonstrated that CURB and Oracle achieve comparable results and outperform other approaches in most cases.

Fake News Detection in Social Networks via Crowd Signals [39] Similar to the previous work this paper proposes a hybrid human-AI approach which engages users of online social networks in reporting fake news. Their algorithm called Detective learns about users flagging accuracy overtime and performs bayesian inference to detect fake news. As an improvement on the previous work, Detective learns about flagging accuracy of users, while CURB assumes that users are equally reliable. The authors'objective in this research is to minimize the spreading of fake news meaning they aim to limit the number of users who end up seeing the fake news before it is blocked. Experiments are conducted against 3 baselines namely fixed-CM, NoLearn, an Random. Furthermore, 2 additional fictitious variations of Detective- Oracle (has access to true labels of posts) and OPT(knows about the users' flagging behaviour) has been considered as well. Their results demonstrated that Detective's performance converges to OPT algorithm as it progresses with learning users' flagging activity.

An Interpretable Joint Graphical Model for Fact-Checking from Crowds [28] In this paper, they use the combination of crowd annotations machine learning to create whats called a PGM (Probabilistic Graphical Method). The advantages of using a PGM are: 1)User knowledge can be added to continuously strengthen the model 2)Transparency 3)Uncertainty of predictions can be quantified. The transparency ensures that the model can be deeply inspected, understood and criticized by people. The model assumed that all annotators are fallible and make mistakes, a good stance to take in a subjective domain like fact checking. They use 'veracity' and 'stance' as hidden variables, and send the respective confusion matrices, source reliability and stance prediction parameters as further parameters to the EM Algorithm. What this translates to is that a claim is deemed likely to be untrue if credible sources have opposing stance, which their model explicitly catches. Comparing against a baseline of 2 LR models for stance prediction (Ferreira and Vlachos 2016) and claim veracity (Dawidand Skene's 1979), trained on the same Emergent dataset, we find that Gibbs Sampling shows their method to be more accurate, at a time tradeoff against baseline standards.

Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking [27]
Note: This paper is more of a collaborative process between an individual user and the model than crowdsourcing of predictions. This paper focuses on making the model that arrives at fake-news predictions transparent and understandable, recognizing that currently many people do not trust popular fact-checking websites. It also postulates that models should be open about the confidence in prediction, and allow users to provides their own opinions to enable the model to formulate an integrated prediction of trustworthiness. The predictions are based on two machine learning classifications, 1) detecting an article's position on an issue, and 2) for truthfulness of a claim. Their stance classifier is first fed with an article's heading and a claim. It predicts the headlines position wrt to the claim, and this prediction is then fed into the veracity classifier, that uses all the stance outputs for all related articles regarding a claim to finally predict the truthfulness of a claim, with reputed sources being more highly weighted in the training dataset. They conducted multiple experiments with their open model of prediction, using mechanical turk users as subjects. However, multiple users found the system confusing to use, and we must account for the fact that people who are not being paid to interact with such fact checking systems may not feel incentivized to interact with these models, with the learning curve outweighing the openness of the model.

3.5 Methods Focused on Linguistics and Writing Style

Considerable work has been done in the realm of investigating how language and writing style can predict truthfulness of an article or post.At a high level, the papers below note significant differences in use of biased, inflammatory, subjective language between suspicious and trustworthy news posts and user comments, and use this information for lexicon and model generation.

Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter [40]
This paper uses a linguistically weighted neural network (RNN and CNN) to analyse tweets, and performs 2 tasks. I) A binary classification between verified and suspicious tweets. II) A multi-class classification to categorize them as one of: 1) Propaganda 2) Hoax 3) Satire and 4) Click-bait. In order to add linguistic weightage, they use a 'late fusion' approach. Fusion lets neural networks learn from a combined channel of multiple input sources. They use fusion to train the models to combine data representations from different methods (network and text features) to increase accuracy of predictions. Linguistic Inquiry Word Count (LIWC) features are used (Pennebaker et al., 2001) to capture additional signals of persuasive and biased language in tweets. Experiments are conducted against several baselines: (a)TFIDF features, (b)Doc2Vec vectors and (c) Doc2Vec or TFIDF features concatenated with linguistic or network

		Twitter		Weibo	
		Accuracy	F score	Accuracy	F score
Previous	DTRank [47]	0.624	0.636	0.732	0.726
	DTC [1]	0.711	0.702	0.831	0.831
	SVM-TS [22]	0.767	0.773	0.857	0.861
	LSTM-1 [19]	0.814	0.808	0.896	0.913
	GRU-NN [21]	0.835	0.830	0.910	0.914
	CI [37]	0.847	0.846	0.928	0.927
	CIt [37]	0.854	0.848	0.939	0.940
	CSI [37]	0.892	0.894	0.953	0.954

feature, and results (Shown in table 1 below) reveal that both the Linguistic RNN and CNN models outperform logistic regression baselines on both the tasks.

	Word2Vec	TFIDF	L.RNN	L.CNN
Binary(A)	0.75	0.79	0.95	0.95
Multiclass(F1)	0.88	0.89	0.91	0.91

Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media [10] This paper analyses comments left on articles on social media and finds that as veracity of an article decreases, the quality of comments on the post decreases too, with increased usage of emojis and swear words, off topic comments and lack of objectivity. They then use this information to automatically predict the truthfulness of posts using a new lexicon called ComLex. They find that hyperpartisan news publishers employ a sensational and inflammatory style of writing. ComLex is used as its main word cluster and they use LIWC (like the previous paper) to validate and support their findings. ComLex is built as a blend of unsupervised machine learning and learning word embeddings, and performs Word2Vec transformations, keeping emotional cues such as exclamation marks and emjois intact. Depending on the word frequencies in each cluster, a statistic is calculated per user comment. Notable results are given in the table below. Baseline comparison for ComLex is done against LIWC and EmoLex.

4 CONCLUSION

In this survey, we investigated the fake news problem from a data mining point of view, by reviewing existing literature and state of the art models for automating the fact checking process. We have first reviewed the process contexts for feature extraction. Models have been categorized based on methodology, arriving at 5 different broad categories including evidence based assessment, feature based models, propagation based models, and linguistic style based models and crowd sourcing. We explored the most recent and novel approaches in each category, and we have compared between models experimenting on the same datasets.

5 FUTURE WORK

Key work remains to be done, as the fake news epidemic is still in relatively early stages. Implementing multiple methods on a common and suitable dataset to be able to fairly evaluate and compare between models to arrive at a provably best method would be the logical next steps to our research. Retraining current models on new data to verify if their consistency and accuracy remains significant is also of importance, as the language and manifestations of fake news are ever evolving. Finally, in the future we are likely to be subject to fake news in the form of synthesized recordings, doctored videos, and so on. Developing methods that are capable of detecting them will be a major challenge for the entire community.

ACKNOWLEDGMENTS

The authors would like to thank Shi Zhi and Qi Li.

REFERENCES

[1] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.
[2] 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*.
[3] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.
[4] Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science* 121, 817–825.
[5] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor Detection with Hierarchical Social Attention Network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 943–951.
[6] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 729–736.
[7] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12, 1945–1948.
[8] Benjamin D Horne and Sibel Adali. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*.
[9] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. *arXiv preprint arXiv:1804.07587*.
[10] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, 82.
[11] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 8.
[12] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 230–239.
[13] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-Source Multi-Class Fake News Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1546–1557.
[14] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 324–332.
[15] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *arXiv preprint arXiv:1809.08193*.
[16] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 591–602.
[17] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1103–1108.
[18] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1867–1870.
[19] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the*

Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Vol. 2. 252–256.

[20] Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 393–398.

[21] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks.. In *IJCAI*. 3818–3824.

[22] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1751–1754.

[23] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 708–717.

[24] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 585–593.

[25] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. 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)*, Vol. 1. 1980–1989.

[26] Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 353–362.

[27] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 189–199.

[28] An T Nguyen, Aditya Kharosekar, Matthew Lease, and Byron C Wallace. 2018. An Interpretable Joint Graphical Model for Fact-Checking From Crowds.. In *AAAI*.

[29] Shivam B. Parikh and Pradeep K. Atrey. 2018. Media-Rich Fake News Detection: A Survey. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 436–441.

[30] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2173–2178.

[31] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1003–1012.

[32] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A Credibility Lens for Analyzing and Explaining Misinformation. *CRF* 71, 88.74, 80–00.

[33] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

[34] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1589–1599.

[35] Feng Qian, ChengYue Gong, Karishma Sharma, and Yan Liu. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence.. In *IJCAI*, Vol. 3834. 3840.

[36] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. 7–17.

[37] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.

[38] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

[39] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 517–524.

[40] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 647–653.

[41] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

[42] Wikipedia contributors. 2018. MMR vaccine controversy — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=MMR_vaccine_controversy&oldid=867038457 [Online; accessed 29-November-2018].

[43] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 651–662.

[44] Shu Wu, Qiang Liu, Yong Liu, Liang Wang, and Tieniu Tan. 2016. Information Credibility Evaluation on Social Media.. In *AAAI*. 4403–4404.

[45] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on Sina Weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 13.

[46] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. 2018. Fake News Detection with Deep Diffusive Network Model. *arXiv preprint arXiv:1805.08751*.

[47] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.

[48] Shi Zhi, Yicheng Sun, Jiayi Liu, Chao Zhang, and Jiawei Han. 2017. ClaimVerif: A Real-time Claim Verification System Using the Web and Fact Databases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2555–2558.