Misinformation in Social Media: Definition, Manipulation, and Detection

Liang Wu*, Fred Morstatter†, Kathleen M. Carley‡, and Huan Liu*
* Arizona State University, Tempe, AZ, USA
† USC Information Sciences Institute, Marina Del Rey, CA, USA
‡ Carnegie Mellon University, Pittsburgh, PA, USA
{wuliang, huan.liu}@asu.edu, fredmors@isi.edu, kathleen.carley@cs.cmu.edu

ABSTRACT

The widespread dissemination of misinformation in social media has recently received a lot of attention in academia. While the problem of misinformation in social media has been intensively studied, there are seemingly different definitions for the same problem, and inconsistent results in different studies. In this survey, we aim to consolidate the observations, and investigate how an optimal method can be selected given specific conditions and contexts. To this end, we first introduce a definition for misinformation in social media and we examine the difference between misinformation detection and classic supervised learning. Second, we describe the diffusion of misinformation and introduce how spreaders propagate misinformation in social networks. Third, we explain characteristics of individual methods of misinformation detection, and provide commentary on their advantages and pitfalls. By reflecting applicability of different methods, we hope to enable the intensive research in this area to be conveniently reused in real-world applications and open up potential directions for future studies.

1. INTRODUCTION

The openness and timeliness of social media have largely facilitated the creation and dissemination of misinformation, such as rumor, spam, and fake news. As witnessed in recent incidents of misinformation, how to detect misinformation in social media has become an important problem. It is reported that over two thirds adults in US read news from social media, with 20% doing so frequently[1]. Though the spread of misinformation has been studied in journalism, the openness of social networking platforms, combined with the potential for automation, facilitates misinformation to rapidly propagate to a large group of people, which brings about unprecedented challenges.

By definition, misinformation is false or inaccurate information that is deliberately created and is intentionally or unintentionally propagated. However, as illustrated in Figure 1, there are several similar terms that may easily get confused with misinformation. For example, disinformation also refers to inaccurate information which is usually distinguished from misinformation by the intention of deception, fake news refers to false information in the form of news (which is not necessarily disinformation since it may be unintentionally shared by innocent users), rumor refers to unverified information that can be either true or false, and spam refers to irrelevant information that is sent to a large number of users. A clear definition is helpful for establishing a scope or boundary of the problem, which is crucial for designing a machine learning algorithm.

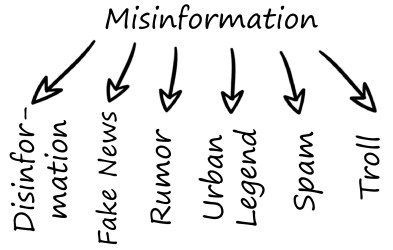

Figure 1: Key terms related to misinformation.

Another challenge is that results on similar problems can often be inconsistent. It is usually caused by the heterogeneity of various misinformation applications, where different features, experimental settings, and evaluation measures may be adopted in different papers. The inconsistency makes it difficult to relate one method to another, which hinders the research results to be applied in real-world applications. To this end, this survey aims to review existing approaches and literature by categorizing them based on the datasets and experimental settings. Through examining these methods from the perspective of machine learning, our goal is to consolidate seemingly different results and observations, and allow for related practitioners and researchers to reuse existing methods and learn from the results.

In this work, we aim to (1) introduce a definition for misinformation in social media that helps establish a clear scope for related research; (2) discuss how misinformation spreaders actively avoid being detected and propagate misinformation; and (3) review existing approaches and consolidate different results, observations and methods from the perspective of machine learning. As we discussed earlier, a definition for misinformation in social media can help a detection method to focus on the specific scope of the problem.

The review article has been partially presented as a tutorial at SBP'16 and ICDM'17

[1]https://www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from-social-media-survey-idUSKCN1BJ2A8

Through studying the diffusion process of misinformation, and how misinformation spreaders manage to avoid being detected, we will introduce methods that are robust to such adversarial attacks. By reviewing existing approaches based on the datasets, features, and experimental settings, it is found that the performance of a method relies on the provided information of a problem, such as the availability of content and network data, and the requirements of a solution, and thus no single method is superior over the rest. We hope these findings will make existing research and results handy for real-world applications.

The rest of the paper is organized as follows. Section 2 presents a definition for misinformation and discusses several related concepts. Section 3 examines misinformation diffusion and several types of adversarial attacks of misinformation spreaders, and introduces countermeasures that make a detection system robust to such attacks. Section 4 introduces misinformation detection methods, which focuses on optimizing both accuracy and earliness. Section 5 discusses feature engineering methods, available datasets, ground truth and evaluation methods. Section 6 concludes the survey, and provide several future directions in this area.

2. MISINFORMATION DEFINITION

There are several related terms similar to misinformation. Rather than the concepts are relatively easier to distinguish, such as spam (a large number of recipients) rumor (verified or unverified) and fake news (in the format of news), the most similar or confusing term is disinformation. Misinformation and disinformation both refer to fake or inaccurate information, and a key distinction between them lies in the intention - whether the information is deliberately created to deceive, and disinformation usually refers to the intentional cases while misinformation the unintentional. Throughout our discussion, *we refer to misinformation as an umbrella term to include all false or inaccurate information that is spread in social media.* We choose to do so since on a platform where any user can publish anything, it is particularly difficult for researchers, practitioners, or even administrators of social network companies, to determine whether a piece of misinformation is deliberately created or not.

The various concepts that are covered in the umbrella term, such as disinformation, spam, rumor, fake news, all share a characteristic that the inaccurate messages can causes distress and various kinds of destructive effect through social media, especially when timely intervention is absent. There have been examples of widespread misinformation in social media during the 2016 Presidential Election in the US that have been facilitating unnecessary fears through social media. One of them is **PizzaGate**, a conspiracy theory about a pizzeria being a nest of child-trafficking. It started breaking out simultaneously on multiple social media sites including Facebook, Twitter and Reddit[2]. After being promoted by radios and podcasts[3], the tense situation finally motivated someone to fire a rifle inside the restaurant[4]. PizzaGate even circulated for a while after the gunfire and being debunked.

To better understanding misinformation in social media, we organize different types of misinformation below, though the categorization is not exclusive.

- **Unintentionally-Spread Misinformation**:

  Some misinformation is unintentional to deceive its recipients. Regular and benign users may contribute to the propagation merely due to their trust of information sources, such as their friends, family, colleagues or influential users in the social network. Instead of wanting to deceive, they usually try to inform their social network friends of a certain issue or situation. An example is the widespread misinformation about Ebola[5].

- **Intentionally-Spread Misinformation**:

  Some misinformation is intentionally spread to deceive its recipients, which has triggered the intensive discussion about misinformation and fake news recently. There are usually writers and coordinated groups of spreaders behind the popularity, who have a clear goal and agenda to compile and promote the misinformation. Typical examples of intentionally-spread misinformation include those conspiracies, rumors and fake news that were trending during the 2016 Presidential Elections. For example, a fake-news writer, Paul Horner[6], has claimed credits for several pieces of fake news that went viral in 2017.

- **Urban Legend**

  Urban legend is intentionally-spread misinformation that is related to fictional stories about local events. The purpose can often be entertainment.

- **Fake News**

  Fake news is intentionally-spread misinformation that is in the format of news. Recent incidents reveal that fake news can be used as propaganda and get viral through news media and social media [39; 38].

- **Unverified Information**

  Unverified information is also included in our definition, although it can sometimes be true and accurate. A piece of information can be defined as unverified information before it is verified, and those verified to be false or inaccurate obviously belong to misinformation. It may trigger similar effects as other types of misinformation, such as fear, hatred and astonishment.

- **Rumor**

  Rumor is unverified information that can be true (true rumor). An example of true rumor is about deaths of several ducks in Guangxi, China, which were claimed to be caused by avian influenza[7]. It had been a true rumor until it was verified to be true by the government[8]. A similar example of avian influenza, which

[2]https://www.nytimes.com/2016/11/21/technology/fact-check-this-pizzeria-is-not-a-child-trafficking-site.html
[3]https://www.nytimes.com/2017/03/25/business/alex-jones-pizzagate-apology-comet-ping-pong.html
[4]https://www.nytimes.com/2016/12/05/us/pizzagate-comet-ping-pong-edgar-maddison-welch.html
[5]http://time.com/3479254/ebola-social-media/
[6]https://www.nytimes.com/2017/09/27/business/media/paul-horner-dead-fake-news.html
[7]http://www.chinadaily.com.cn/en/doc/2004-01/28/content$_3$01225.*htm*
[8]http://www.people.com.cn/GB/shizheng/1026/2309847.html, in Chinese

turned out to be false, was that some people had been infected through eating well cooked chicken[9].

- **Crowdturfing**

  Crowdturfing is a concept originated from astroturfing, which means the campaign masks its supporters and sponsors to make it appear to be launched by grassroots participants. Crowdturfing is "crowdsourced" astroturfing, where supporters obtain their seemingly grassroots participants through the internet. Similarly as unverified information or rumor, the information promoted by crowdturfing may also be true, but the popularity inflated by crowdsourcing workers is fake and unfair. Some incidents of misinformation that cause negative effects are caused crowdturfing. There are several online platforms where crowdturfing workers can be easily hired, such as Zhubajie, Sandaha, and Fiverr. There have been claims that crowdturfing have been used to target some certain politicians[10].

- **Spam**

  Spam is unsolicited information that unfairly overwhelms its recipients. It has been found on various platforms including instant messaging, email and social media.

- **Troll**

  Another kind of misinformation we focus on is troll. Troll aims to cause disruption and argument towards a certain group of people. Different from other types of misinformation that try to convince its recipients, trolling aims to increase the tension between ideas and ultimately to deepen the hatred and widen the gap. For example, the probability for a median voter to vote for a certain candidate can be aroused by being trolled. In 2016, the troll army that has been claimed to be controlled by the Russian government was accused of trolling at key election moments[11].

- **Hate speech**

  Hate speech refers to abusive content on social media that targets certain groups of people, expressing prejudice and threatening. A dynamic interplay was found between the 2016 presidential election and hate speech against some protected groups, and the peak of hate speech was reached during the election day[12].

- **Cyberbullying**

  Cyberbullying is a form of bullying happening online, usually in social media, that may consist of any form of misinformation, such as rumor and hate speech.

[9]http://www.xinhuanet.com/food/2017-02/22/c_1120506534.htm

[10]https://www.fastcompany.com/3067643/how-trumps-opponents-are-crowdsourcing-the-resistance

[11]https://www.nbcnews.com/tech/social-media/russian-trolls-went-attack-during-key-election-moments-n827176

[12]https://www.washingtonpost.com/news/post-nation/wp/2018/03/23/hate-crimes-rose-the-day-after-trump-was-elected-fbi-data-show/

3. MANIPULATION OF MISINFORMATION

In this section, we will investigate solutions to address the challenges brought by adversarial attacks of misinformation spreaders. There are different types of spreaders and we focus on those who spread misinformation in social networks, and our research particularly focuses on those who spread it on purpose. Traditional approaches mainly focus on their excessive suspicious content and network topology, which obviously distance themselves from normal users. However, as indicated by recent incidents of rumor and fake news, misinformation spreaders are not easily discovered with simple metrics like the number of followers and followee/follower ratio. Instead, they will actively disguise themselves, and the performance of a classic supervised learning system would degrade rapidly due to the adversarial attacks. For example, a malicious user may copy content from other legitimate accounts, or even use compromised accounts to camouflage misinformation they are spreading. In order to appear as benign users, they may also establish links with other accounts to manipulate the network topology. To further complicate the problem, there is a lack of availability of label information for the disguised content or behaviors, which makes it difficult to capture the signal of misinformation. In summary, there are mainly two kinds of attacks in social media.

Manipulation of Networks Since many users follow back when they are followed by someone for the sake of courtesy, misinformation spreaders could establish a decent number of links with legitimate users [37]. These noisy links no longer reflects homophily between two nodes, which undermine the performance of existing approaches. In addition, misinformation spreaders may even form a group by connecting with each other, and such coordinated behaviors are particularly challenging for a traditional method.

Manipulation of Content It is easy for a misinformation spreader to copy a significant portion of content from legitimate accounts. The misinformation that they intend to spread is camouflaged by the legitimate messages to avoid being detected. Traditional approaches merge posts of an account altogether as an attribute vector, which would be less distinguishable to capture the signal of misinformation.

3.1 Content-based Manipulation

Social network users are naturally defined by the content they create and spread. Therefore, a direct way to identify misinformation spreaders from social network accounts is to model their content information. Traditional approaches mainly focus on classification methods, trying to decode the coordinated behaviors of misinformation spreaders and learn a binary classifier. For the rest of the subsection, we will introduce traditional methods, adversarial attacks against the models, and possible solutions to tackling the challenges. Figure 2 illustrates an example of misinformation in Twitter. In social media websites, the content information can usually be extracted from posts and user profiles. We summarize several categories of methods based on the content information they rely on.

Content extracted from a user's posts has been studied in early research to directly identify misinformation spreaders [21], and a text classifier can be used to classify malicious users. In order to jointly utilize the network information, previous work also extracts links between social network users, and the classification task is reduced to categorizing attributed vertices in a graph [14; 16]. Content information

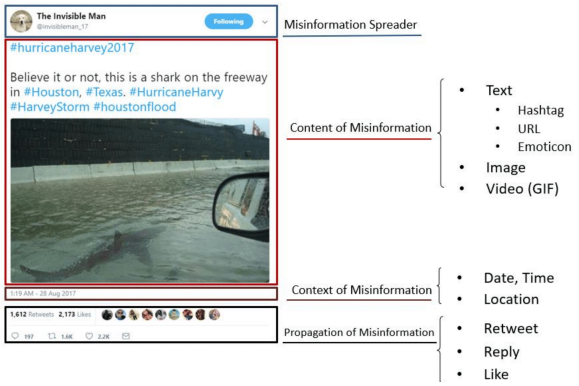

Figure 2: An example of misinformation in social media.

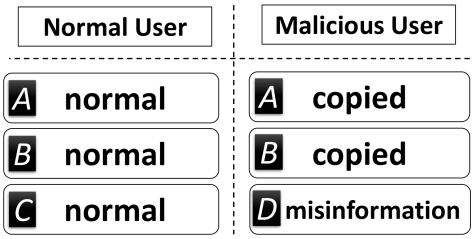

Figure 3: A toy example of camouflaged misinformation spreaders, where a normal user's posts (A, B and C) are copied to camouflage a misinformation post (D).

extracted from profile [20] has also been utilized to compile the attribute vectors with posts.

Profiles can also be directly utilized to identify a misinformation spreader. For example, the length of screen name and account description, and longevity of accounts are jointly used for spreader detection [21]. In addition, a more recent study uses only profile information for the task [22] - by utilizing unigram/bigram, edit distance and other content information, the authors build a classifier that discriminates user-generated from algorithmic generated profiles, where the automatically generated usernames show a distinguishable patterns from the regular manually-generated ones. However, unlike posts that can be essential for spreading misinformation, it is unnecessary that profiles also contain malicious signals. Therefore, profile-based methods are specially designed for certain spreaders on some platforms.

Links to external resources have also been studied, which direct normal users to websites through URLs. For example, while the content of a post varies, researchers find a group of misinformation spreaders may embed URLs in their posts that are directing to the same target [53]. In addition, they also discover a bursty fashion of URL-based misinformation. The main intuition is that a group of accounts can be used for a particular target within a specific period. Based on URLs, a signature can be generated to label all such accounts for detecting the bursts.

Sentiment information embedded in the content has also been used to detect political misinformation and its inception [5]. This stream of research can be further utilized to study the role of misinformation propagation in political campaigns [30], and it is found that centrally-controlled accounts are used to artificially inflate the support for certain political figures and campaigns. In a 2010 US midterm election dataset, based on several traditional classification methods, campaign-related misinformation propagation has been discovered [31].

To compare the classification methods that have been used, it is difficult to say that a single method outperforms the rest consistently. By studying early literature that directly applies classic classification methods, various models have been reported to produce the optimal results. For example, Naive Bayes, a generative model that has been widely applied in various tasks because of its simple structure, consistent performance and robustness to missing data, has been found for classifying social media users accurately [8]. The performance of a NB model heavily relies on the Bayes' Theorem with conditional independence assumptions between the features. In the work, the proposed features are manually compiled to avoid such problems. Similarly, SVM, another popular classification algorithm that minimizes the generalization error, has also been found to achieve the best performance for the task [23]. A drawback of SVM is its sensitivity to missing data, and the proposed methods mainly rely on generating pattern-based features for each user in the dataset, which avoids any missing entry. Other methods like decision trees and AdaBoost have been reported to produce the best results [31].

Therefore, if considering misinformation spreader as a classic classification task, these binomial classification algorithms perform very similarly with each other. The superiority highly depends on a certain dataset and what/how features are used (feature engineering). We will talk more about feature engineering in Section 5.

Content of misinformation can, however, be highly manipulated. For example, as illustrated in Figure 3, where a normal user posts A, B, and C, and the adversarial rival copies them to camouflage the polluting post D. Misinformation, *i.e.* post D, can be very difficult to detect since traditional methods will merge all content together for a user. A key challenge here is data scarcity - since camouflage can take up most of the content from a misinformation spreader, it is not easy to identify the scarce polluting evidence. Given enough labels, the problem can naturally boil down to a post classification task.

In order to fight against such manipulation, researchers propose to adaptively model user content. In particular, they aim to select a group of posts in a user's content that are less likely to be camouflage [48]. It recursively models content and network information to find groups of posts that distinguish a user from others. The results from real-world data prove such adaptive modeling helps a classifier better identify suspicious accounts.

Besides supervised methods, unsupervised approaches have also been investigated in solving the problem. Since some misinformation spreaders' accounts are generated in a batch, they may distance themselves from an organic account. For example, Webb *et al.* try to discover highly similar accounts based on user profiles [45], to be generated with the same template. Chu *et al.* propose to focus on the posting behaviors - since misinformation spreaders are often employed to post information related to a specific topic, their posting behavior often contains long hibernation and bursty peaks. Thus the proposed methods leverage the temporal features

and discover malicious behaviors [8]. User behavioral patterns, such as online rating [25] and locations [24], have also been studied.

3.2 Network-based Manipulation

In this subsection, we will discuss network-based attacks of misinformation spreaders and how to deal with them. Since many users follow back when they are followed by someone for the sake of courtesy, misinformation spreaders could establish a decent number of links with legitimate users [46; 37]. These noisy links no longer reflect homophily between two nodes, which undermine the performance of existing approaches. In addition, misinformation spreaders may even form a group by connecting with each other, and such coordinated behaviors are particularly challenging for a traditional method. To this end, existing research focuses on how social interactions could be used to differentiate malicious users from the legitimate ones. This is inherently a graph classification problem that manifests in social networks. Therefore, the source of network information can be leveraged to identify misinformation spreaders.

Misinformation spreaders may behave very differently from regular users, and previous research aims to identify distinguishing characteristics. A classic assumption is that misinformation spreaders seldom make friends, and thus a small number of links versus a relatively long account age may indicate being fake [26]. It is obvious that such detection methods are prone to be tricked through harvesting friends on social networks. The hidden assumption here is that friendship on a social media platform can only be established with regular users. Therefore, the methods relying on the absolute number of followers are effective only for a certain type of misinformation spreaders. There are methods focusing on follower/followee ratio [20], however, it is still vulnerable as long as enough number of followers can be harvested - the ratio can be easily manipulated by unfollowing the followees.

In order to cope with attacks of follower harvesting, a relatively recent research direction focuses on homophily [27], *i.e.,* assuming a pair of friends are likely to be of the same label. The corresponding research can be categorized as neighbor-based methods. Based on the assumption, a supervised learning algorithm can clamp prediction results of a pair of friends [60; 43; 44; 32]. Another common approach is to use the links to derive a group of users that are densely connected with each other. Since a group of malicious users usually focus on specific topics, selected features that are better reflecting the group can be used to discover misinformation spreaders [14]. An attack against neighbor-based methods is that misinformation spreaders can harvest links with regular users. The hidden assumption is that social media users are careful about connections. However, many people would simply follow back after being followed.

In order to fight against attacks of regular-user friend harvesting, group-based approaches have been proposed. First, researchers have been focusing on finding coordinated groups of misinformation spreaders. Given labels of some known social media users, the task can be seen as propagating labels using the links. The coordinated misinformation spreaders are expected to be grouped together by the dense connections between them [55]. Second, the task can also be regarded as an unsupervised problem, where misinformation spreaders are expected to be outliers in the results [18; 1]. The underlying assumption here is misinformation spreaders do not behave normally and cannot associate with any social community [10].

However, it is still challenging to apply group-based methods in real-world applications. First, both kinds of methods focus only on specific misinformation spreaders and will suffer from a large number of false negatives. The first category of methods aim to achieve a group structure where misinformation spreaders can be grouped together, while the second category of methods aims to achieve a group structure where they can be detached from groups. Second, a hidden assumption of these approaches is that misinformation spreaders are homogeneous and behave similarly. However, misinformation spreaders may emerge from different sources and the optimal parameters, such as the size of a cluster and number of clusters, are very difficult to find. Adaptively acquiring the parameters have been discussed in recent work [47].

4. MISINFORMATION DETECTION

Misinformation detection seems to be a classification problem, which has the same setting as text categorization. Traditional text categorization tasks, where the content is mostly organic and written/compiled to be distinguishable, *e.g.,* sports news articles are meant to be different from political news. By contrast, misinformation posts are deliberately made seemingly real and accurate. Therefore, directly and merely focusing on the text content will be of little help in detecting misinformation. As illustrated in Figure 2, based on the information that a method mainly utilizes, we categorize the detection methods as,

- Content-based misinformation detection: directly detecting misinformation based on its content, such as text, images and video.
- Context-based misinformation detection: detecting misinformation based on the contextual information available in social media, such as locations and time.
- Propagation-based misinformation detection: detecting misinformation based on the propagation patterns, *i.e.,* how misinformation circulates among users.
- Early detection of misinformation: detecting misinformation in an early stage before it becomes viral, usually without adequate data or accurate labels.

4.1 Content-based Approaches

Although it is very difficult to obtain useful features from content information, there have been research directly utilizing text data for different purposes. For example, some studies focus on retrieving all posts related to a known piece of misinformation [40; 15]. This stream of research is more of a text matching problem, where the targeted posts are those very similar or duplicate ones of an original misinformation post. These methods can be very helpful in the later phase of misinformation propagation. When a certain piece of information has been proved to be inaccurate or fake, text-matching methods can be used to find all related posts. However, it is challenging for the methods to capture misinformation that has been intentionally rewritten.

In order to extend the limits of text matching methods, supervised learning methods have been studied to identify misinformation. They usually collect posts and their labels from

microblogging websites such as Twitter and Sina Weibo, and then train a text classifier based on the collected content and labels[56; 54; 58; 11]. The underlying assumption of these methods is that misinformation may consist of certain keywords and/or combinations of keywords, so a single post with enough misinformation signals can be classified. In addition, other contextual information like network structures has also been incorporated [42; 12; 34]. However, post-based methods can be overly sensitive to misinformation content. There can be a large number of posts in real applications, and there may be posts containing certain keywords that lead to false positives.

Message-cluster based methods have been proposed to control the model sensitivity. Instead of focusing on individual posts, these algorithms first cluster messages based on the content, posting time and authors. Then the data instances to classify become the clusters of messages. The methods either aim to find those suspicious instances [29; 59], or find credible clusters of discussions [6; 7; 52]. A practical issue is that these methods can only be trained on popular topics, and a large number of posts need to be collected to support the clustering. Therefore, these methods are better at detecting popular misinformation, which is usually of greater importance. Those misinformation incidents that are not as popular - the ones on the long tail, can often be neglected.

4.2 Context-based Approaches

On social media platforms, contextual information consists of posting time and geolocations. It is usually jointly used with other information to facilitate the detection, or directly vectorized and being used as additional features. In addition, there are also studies that use just contextual information. For example, Kwon *et al.* propose to model the bursty patterns of misinformation [19]. Unlike legitimate posts that are scatterly posted over the time, the authors argue that misinformation posts are usually posted in a burst. The underlying assumption is that misinformation is intentionally promoted by certain groups of accounts and thus has a different posting patterns. Similar observations have been found in an earlier study that rumors are periodically popular in bursts [9].

4.3 Propagation-based Approaches

Information diffusion describes how information spreads in social networks, and related research usually focuses on the users who post and forward information, such as predicting the ultimate influence of a message. In this subsection, we will introduce how propagation information can be used to detect misinformation. A more detailed introduction about utilizing information diffusion research for misinformation studies can be found in a previous review [51].

Since intentional spreaders of misinformation may manipulate the content to make it seem very real, it is very challenging to obtain useful features from content for these emerging applications. To address this problem, some recent work concentrates on modeling the propagation of messages in a social network. For example, a framework called TraceMiner classifies propagation pathways of a message based on network embeddings of social media users [50]. Experimental results on real-world datasets indicate that the proposed method can provide a high degree of classification accuracy compared with content-based approaches. This is natural since content information can be very sparse and noisy on social media.

A key intuition of utilizing propagation information is that the direction of information exchange can reveal community structures and personal characteristics. For example, non-parametric methods have been proposed to infer topical interests of social media users based on content of news data [17]. Therefore, the homogeneity of topical interests of each user can be calculated based on collections of content information, and further they find the corresponding homogeneity can be leveraged to improve supervised misinformation detection systems as an additional feature. In addition to modeling user behaviors, propagation information also allows for understanding characteristics of news being spread. For example, a previous work finds that false rumors have a "bursty" pattern of being popular [19] - rather than being popular only once as a piece of regular news, fake and unverified information can have hibernation and multiple bursts on social media. Therefore, such temporal patterns help characterize information propagation.

4.4 Early Detection of Misinformation

Compared with traditional classification tasks, which mainly focus on optimizing performance metrics like accuracy and F-measure, misinformation detection approaches further take into account the earliness of a method. The earliness, or timeliness of a method describes how fast a misinformation detection method can be ready to classify misinformation. In this context, two challenges immerse as key issues of solving the problem, *i.e.,* **lack of data**, and **lack of labels**. Therefore, recent research introduces related techniques.

In order to cope with the lack of data, a critical issue is to allow for aggregation of data at the early stage. Most existing methods focus on learning from conversations between users, since the content, sentiments, discussions and debates manifested in the interactions provide contextual information useful for characterizing a certain topic. However, in the early stage of misinformation propagation, discussions are usually scattered and it takes time for them to develop into long conversations. In order to shorten the waiting period, researchers propose to analyze whether structural information can be captured in the early stage to help link the scattered discussions [36]. In particular, three types of structural information are discussed, including hashtags, web links and content similarity. The main intuition is to leverage such links to merge individual discussions into a cluster of "conversations".

Experimental results on real-world datasets show that hashtags are the most helpful linkage for misinformation detection, which even enables computationally cheap classification model to significantly outperform competitive baselines. Instead of waiting for a conversation to grow, the proposed method rapidly learns from individual tweets in the formative process, which allows for misinformation to be detected when the very first batch of content are posted. In addition, though the performance of the proposed method and traditional approaches tends to converge when more data is available, the links do not erode its performance in a late stage. A critical issue of utilizing such links is to control the threshold of merging two posts. Sampson *et al.* provide detailed descriptions about how consistent accuracy can be achieved using an engineered number of links [36].

However, even the data sparsity problem can be alleviated, collecting label information still takes time that hinders the

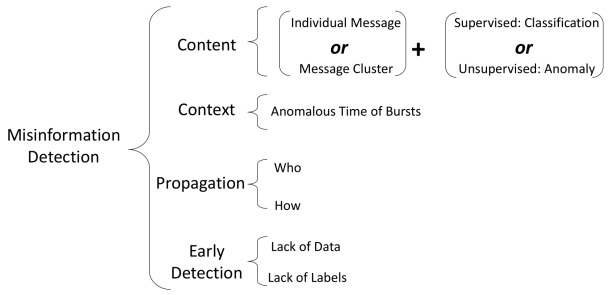

Figure 4: An overview of misinformation detection methods.

earliness of a method. In the context of a traditional classification task, label information can be collected beforehand. For example, a labeled news dataset can be used to classify sports and political news for the next month. However, it is particularly difficult for misinformation problems to directly reuse label information. Misinformation data is usually very topic-sensitive, and the corresponding vocabulary and word choice may vary substantially between different pieces of fake news or rumors. Therefore, directly reusing existing label information unavoidably brings in noise. A common practice is to build a labeled dataset for each misinformation incident. For example, in order to detect a certain piece of fake news, current systems usually need to build a specific training dataset for it, and the task is very similar to text search/matching.

Collecting label information can be time-consuming and labor-intensive, and researchers have proposed to find ways to allow for reusing training data from previous misinformation datasets. Though different rumors and fake news may be different from each other, two different pieces of misinformation may trigger similar reactions. For example, a wedge-driving rumor can cause hatred, and atrocious fake news arouses astonishment among its readers. Therefore, researchers propose to discover useful patterns from reactions of readers [49]. The proposed framework focuses on user reactions and jointly clusters data with similar reactions and selects topic-invariant features from the comments and reactions. By clustering misinformation by reactions instead of the topic, previous training data can be readily used to identify emerging misinformation incidents, which breaks the detection bottleneck of earliness.

To summarize, we depict the area in Figure 4. Misinformation detection methods include content-based methods, context-based methods, and propagation-based methods, which are categorized based on the information they rely on, and early detection methods, which focuses on detection earliness in addition to accuracy.

5. DISCUSSION

In this section, we will discuss some related topics about misinformation, including common feature engineering methods for dealing with social media data, and available resources for conducting experiments.

5.1 Feature Engineering

Content features are extracted from an account's profile and posts. Profile features usually include a username, a brief biography and a profile photo. A relatively larger source of content information are the posts a user publishes or forwards. Early research in this field regards social media users as a special type of documents [3; 26]. There are various ways of engineering features of social media content, we list several commonly used methods here.

- **Words** Words in the content can be directly used. They can either be used as binary bit that indicates whether a word has appeared or a weight that indicates how important the word is to the user. The weight can be derived from word frequency, which is the same as bag of words (BoW), or derived from other weighting schemes like TF-IDF [35].
- **Latent Features** Since social media content is usually very short, directly using words leads to very sparse representations that are hard to process. Therefore, in contrast to using raw words, various methods test using latent semantic features that are learned from words, such as Language Modeling (LM) [28], word embeddings [33] and Topic Modeling [4].
- **Keywords** A challenge of dealing with social media content is the sparsity due to short content and a large vocabulary. In order to tackle the challenge, some methods rely on extracting and learning keywords. The keywords can be a compiled list of specific words, or words taken from *URL*, *hashtags* and mentions, such as domain names and account names. Sentiment words and some emoticons (emojis) have also been used in related studies.

Based on the extracted features, the task of misinformation spreader detection can naturally boil down to a classic supervised learning problem, where each data instance is a user and a classification model is applied to categorize the users based on the content and network features.

5.2 Ground Truth and Evaluation

Evaluation of misinformation spreader detection is very similar to that of traditional classification tasks. To reduce the variance of evaluations, as well as improve the performance of classification method, the size of training data plays a crucial role. However, it is particularly challenging to obtain ground truth for this task. Considering that misinformation detection is a time-critical task, labeling data manually that may be time-consuming and labor-intensive is impractical in real-world applications. Therefore, existing research mainly relies on automatic ways of obtaining ground truth. There are two major ways that ground truth data can be collected automatically - utilizing the lists of suspended users, and deploying honeypot accounts on social network websites.

Suspended Users

Social media websites, such as Twitter and Facebook, constantly suspend users that violate their community rules. Many researchers leverage these suspended users to obtain ground-truth data. It is straightforward when the list is publicly available and the suspended accounts can be directly used. In order to leverage the list when it is not publicly available, researchers usually select a large number of active accounts and monitor them on social media websites. If any monitored account is found to have been suspended, the account together with its content and behaviors can be used to establish a labeled dataset. In particular, the active users are labeled as regular users while the suspended

as misinformation spreaders. Though the method enables ground-truth data to be easily collected on a large scale, recent studies have also realized that the quality may be heavily influenced by community rules that are used to filter accounts [22; 41].

- **Honeypots** Another approach to discovering misinformation spreaders on social media is to create "honeypots" which attract misinformation spreaders to attack. A honeypot in the context of this discussion is an automated account controlled by the researcher in order to attract real misinformation spreaders to follow it. The way each honeypot network ("honeynet") behaves is determined by the researcher. Honeypots have been used in recent work. In [21], the authors use a honeynet containing 60 honeypots to tempt bots to follow them. Each of these honeypots focuses on gaining attraction by tweeting trending topics and links as well as regular tweets and tweets mentioning other honeypots. Throughout the course of their 7-month study, they attracted 22,223 bots into their honeynet. The dataset used in this work is publicly available[13]. While the honeypot process is a promising way to collect misinformation spreaders, there are two major drawbacks. First, Twitter may ban some of the honeypot accounts, causing the researcher to have to recreate API keys to continue the data collection process. Furthermore, while bots follow honeypots, real users do not. Thus, a researcher will need to collect a series of accounts from real users to obtain negative labels.
- **Sina Weibo** is a popular microblogging platform in China, which is very similar to Twitter. A nice property of Weibo is that they have launched a fact-checking platform, which is based on crowdsourcing, to enable normal users to report and label suspicious content. On the platform[14], any normal user can report a post and specify a certain community rule that the content has violated. The final judgement will be made by a committee based on a majority vote. The committee members are also regular users of Weibo, and they are all volunteers and do not receive any stipend. This process is like a crowdsourced version of Twitter's user suspension, which is based on Twiiter employees and their community rules[15].
- **Fact-checking Websites** In order to obtain labels for misinformation posts and post clusters, fact-checking data is extracted and experimented. Popular websites include Snopes[16], TruthorFiction[17], and PolitiFact[18]. These websites mainly focus on popular events, and they maintain a team of professional editors to manually check the truthfulness of the incidents.

Besides relying on ground truth, unsupervised methods [13], and evaluation methods without the underlying ground truth can also be used for social media data. More details can be found in a related review [57]

[13]http://infolab.tamu.edu/data/
[14]http://service.account.weibo.com/
[15]https://help.twitter.com/en/rules-and-policies/twitter-rules
[16]https://www.snopes.com/
[17]https://www.truthorfiction.com/
[18]https://www.politifact.com/

6. CONCLUSION AND FUTURE RESEARCH

As witnessed in recent incidents of misinformation, social media have allowed for rumors and fake news to spread to a large group of people rapidly. While researchers have intensively focused on the problem of misinformation detection, seemingly different observations and experimental results are reported from different research works. In this survey, we aim to consolidate related observations and results. In particular, we try to answer the following questions in this survey.

- **How is misinformation detection different from text classification?** We introduce the definition of misinformation, misinformation spreaders in social media, explaining several similar terms such as spam, rumor, disinformation and reason why we use misinformation as an umbrella term. We also discuss how misinformation detection is computationally different from classic classification problem, and how researchers and practitioners tackle different challenges toward detecting it.

- **How to identify misinformation spreaders?** Misinformation is usually spread by certain accounts that distance from regular social media users. We introduce how these spreaders can be detected by discussing the feature engineering methods, available sources of label information. Since misinformation spreaders actively manipulate social media platforms to avoid being detected, we introduce several state-of-the-art approaches that are robust to such attacks with networks and content.

- **Beyond text, what other information can we utilize to characterize misinformation and its spreaders?** As we realize text information can provide limited help in identifying misinformation and misinformation spreaders, we talk about additional information sources that can help expose malicious behaviors and information, such as temporal patterns, posting frequency and propagation paths. We introduce how the additional information can be individually utilized to complement text information.

In addition, benchmark datasets and evaluation metrics are also introduced for misinformation identification and intervention. Since mining misinformation in a social network is an emergent field of study, we also list a number of interesting potential problems for future exploration:

- **How to predict the potential influence of misinformation in social media?** As an instance of classification, existing misinformation detection methods focus on optimizing classification accuracy. In real-world applications, however, detecting an influential spreader is may be more useful than ten unimportant ones that can hardly spread misinformation to regular users. It will be interesting to define influence of misinformation spreaders and formulate a computational problem to cope with it.

- **How are misinformation spreaders spreading misinformation and attracting attention?** Existing research mostly focuses on the spreaders - or the accounts that post misinformation in social media. In the real world, a spreader would more than that to "spread" misinformation, such as commenting under certain topics, making friends with similar communities, and even privately messaging interested accounts. In addition to detecting them, it would be interesting to discover and understand such spreading behaviors, which may ultimately facilitate building a robust detection system.

- **How to make detection methods robust to adversarial attacks, or how to exploit adversarial learning to enhance a detection method?** Adversarial machine learning aims to enable machine learning methods to be robust and effective in the presence of adversarial attacks. Current research focuses on adversarial attacks of misinformation spreaders, however, if there is a malicious adversary that has partial or full knowledge of the misinformation detection algorithm, existing methods can be vulnerable. It will be interesting to discover robust methods in the presence of adversarial attacks.

7. ACKNOWLEDGEMENTS

This material is based upon work supported by, or in part by, the National Science Foundation (NSF) grant 1614576, and the Office of Naval Research (ONR) grant N00014-16-1-2257. Some content has been presented as a tutorial in SBP'16 and ICDM'17. We would like to acknowledge the insightful feedback from the audience.

8. REFERENCES

[1] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

[2] M. Alfifi, P. Kaghazgaran, J. Caverlee, and F. Morstatter. Measuring the impact of isis social media strategy, 2018.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[4] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] J. Bollen, H. Mao, and A. Pepe. Determining the public mood state by analysis of microblogging posts. In *ALIFE*, pages 667–668, 2010.

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

[7] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.

[8] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.

[9] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.

[10] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On community outliers and their efficient detection in information networks. In *SIGKDD*, pages 813–822. ACM, 2010.

[11] G. B. Guacho, S. Abdali, N. Shah, and 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)*, pages 322–325. IEEE, 2018.

[12] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st workshop on privacy and security in online social media*, page 2. ACM, 2012.

[13] S. Hosseinimotlagh and E. E. Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. 2018.

[14] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *IJCAI*, volume 13, pages 2633–2639, 2013.

[15] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 14–24. Springer, 2017.

[16] P. Kaghazgaran, J. Caverlee, and A. Squicciarini. Combating crowdsourced review manipulators: A neighborhood-based approach. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 306–314. ACM, 2018.

[17] J. Kim, D. Kim, and A. Oh. Homogeneity-based transmissive process to model true and false news in social networks. *arXiv preprint arXiv:1811.09702*, 2018.

[18] E. M. Knorr and R. T. Ng. A unified notion of outliers: Properties and computation. In *KDD*, pages 219–222, 1997.

[19] S. Kwon and M. Cha. Modeling bursty temporal pattern of rumors. In *ICWSM*, 2014.

[20] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[21] K. Lee, B. D. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*. Citeseer, 2011.

[22] S. Lee and J. Kim. Early filtering of ephemeral malicious accounts on twitter. *Computer Communications*, 54:48–57, 2014.

[23] C. Li, W. Jiang, and X. Zou. Botnet: Survey and case study. In *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, pages 1184–1187. IEEE, 2009.

[24] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[25] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM, 2010.

[26] M. Mccord and M. Chuah. Spam detection on twitter using traditional classifiers. In *international conference on Autonomic and trusted computing*, pages 175–186. Springer, 2011.

[27] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[28] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[29] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[30] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[31] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonalves, A. Flammini, , and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, 2011.

[32] S. Rayana and L. Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 985–994. ACM, 2015.

[33] X. Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[34] N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM, 2017.

[35] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.

[36] J. Sampson, F. Morstatter, L. Wu, and H. Liu. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2377–2382. ACM, 2016.

[37] S. Sedhai and A. Sun. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 223–232. ACM, 2015.

[38] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu. Combating fake news: A survey on identification and mitigation techniques. *arXiv preprint arXiv:1901.06437*, 2019.

[39] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[40] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings*, 2014.

[41] K. Thomas, C. Grier, and V. Paxson. Adapting social spam infrastructure for political censorship. In *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats*, pages 13–13. USENIX Association, 2012.

[42] N. Vo and K. Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284. ACM, 2018.

[43] G. Wang, S. Xie, B. Liu, and P. S. Yu. Review graph based online store review spammer detection. In *Data mining (icdm), 2011 ieee 11th international conference on*, pages 1242–1247. IEEE, 2011.

[44] G. Wang, S. Xie, B. Liu, and P. S. Yu. Identify online store review spammers via social review graph. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):61, 2012.

[45] S. Webb, J. Caverlee, and C. Pu. *Social Honeypots: Making Friends With A Spammer Near You*. Conference on Email and Anti-Spam, 2008.

[46] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM conference on WSDM*, pages 261–270. ACM, 2010.

[47] L. Wu, X. Hu, F. Morstatter, and H. Liu. Adaptive spammer detection with sparse group modeling. In *ICWSM*, pages 319–326, 2017.

[48] L. Wu, X. Hu, F. Morstatter, and H. Liu. Detecting camouflaged content polluters. In *ICWSM*, pages 696–699, 2017.

[49] L. Wu, J. Li, X. Hu, and H. Liu. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 99–107. SIAM, 2017.

[50] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 637–645. ACM, 2018.

[51] L. Wu, F. Morstatter, X. Hu, and H. Liu. Mining misinformation in social media. *Big Data in Complex and Social Networks*, pages 123–152, 2016.

[52] S. Wu, Q. Liu, Y. Liu, L. Wang, and T. Tan. Information credibility evaluation on social media. In *AAAI*, pages 4403–4404, 2016.

[53] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: signatures and characteristics. *ACM SIGCOMM Computer Communication Review*, 38(4):171–182, 2008.

[54] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.

[55] J. Ye and L. Akoglu. Discovering opinion spammer groups by network footprints. In *Machine Learning and Knowledge Discovery in Databases*, pages 267–282. Springer, 2015.

[56] S. Yu, M. Li, and F. Liu. Rumor identification with maximum entropy in micronet. *Complexity*, 2017, 2017.

[57] R. Zafarani and H. Liu. Evaluation without ground truth in social media research. *Commun. ACM*, 58(6):54–60, 2015.

[58] Q. Zhang, S. Zhang, J. Dong, J. Xiong, and X. Cheng. Automatic detection of rumor on social network. In *Natural Language Processing and Chinese Computing*, pages 113–122. Springer, 2015.

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

[60] Y. Zhu, X. Wang, E. Zhong, N. N. Liu, H. Li, and Q. Yang. *Discovering Spammers in Social Networks*. AAAI, 2012.