2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

An Implicit Crowdsourcing Approach to Rumor Identification in Online Social Networks

Abiola Osho
Dept. of Computer Science
Kansas State University
aaarise@ksu.edu

Caden Waters
Dept. of Computer Science
Kansas State University
crwaters@ksu.edu

George Amariucai
Dept. of Computer Science
Kansas State University
amariucai@ksu.edu

***Abstract*—With the increasing use of online social networks as a source of news and information, the propensity for a rumor to disseminate widely and quickly poses a great concern, especially in disaster situations where users do not have enough time to fact-check posts before making the informed decision to react to a post that appears to be credible. At the same time, we know that misinformation is easily detectable by a certain few, very skeptical, or very informed users. In this study, we demonstrate how blending artificial intelligence and human skills can create a new paradigm for credibility prediction. The crowdsourcing part of the detection mechanism is implemented implicitly, by simply observing the natural interaction between users encountering the messages. Specifically, we explore the spread of information on Twitter at the microscopic (user-to-user propagation) level and propose a model that predicts if a message is *True* or *False* by observing the latent attributes of the message, along with those of the users interacting with it, and their reactions to the message. We demonstrate the application of this model to the detection of misinformation and rank the relevant message and user features that are most critical in influencing the spread of rumor over the network. Our experiments using real-world data show that the proposed model achieves over 90% accuracy in predicting the credibility of posts on Twitter, a significant boost over state-of-the-art models.**

***Index Terms*—Crowdsourcing, Social Networks, Bayesian Learning, Classification and Regression, Misinformation, Dimensionality Reduction/Feature Selection**

I. Introduction

Online social networks (OSNs) like Twitter have become increasingly popular for dissemination of information, news, and events around the world. Due to their wide reach, oversimplified conversations, and ability to provide quick blasts of information, online social networks have also become an avenue for the spread of rumors. With the current political and economic climate around the world, we continue to witness the spread of falsehood initiated in 280 characters or less. In the absence of verification sources, individuals can use online media to disperse and coordinate information, since the potential spread of information (whether true or false) is impartial to the content or source.

The impartial and unrestrained spread of information in social networks can be of great value as observed in September 2015 where the US geological survey tracked earthquakes by simply following mentions of the term 'earthquake' [1], or the 2012-13 flu epidemic where researchers used tweet data to correlate the spread of the disease with a view to reducing its impact [2], and in stock markets where consumer insights companies use social media data to predict shifts in consumer spending behaviors that translate to shifts in stock prices. However, the same social network features that offer these benefits can quickly become detrimental when the spreading information is false, like during hurricane Sandy where there were false tweets about the NYSE being flooded with up to 3 feet of water, which even got reported by some news outlets [3].

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

According to deflationism [4], assertions that predicate truth of a statement do not attribute a truth property to such a statement. Since there is no real-world truth label to posts (i.e., text, images, memes, etc.), OSN users simply decide to react to a post based on the perceived credibility of the message. A message intended to deceive might have concealed meanings, emotions and sentiments even if it appears otherwise. The search for the truthfulness of a message might be lacking, depending on how accepting or prejudiced the user is towards a topic, especially when they are exposed to contradicting information from diverse sources. Since some rumors never completely die out, persisting with low frequencies with potential for flare-ups from time to time, detecting misinformation posts early on, before a flare-up, is more meaningful than detecting them when 90% of the total related post volume has already been consumed [5] [6].

In this study, we adopt an implicit crowdsourcing model for predicting the credibility of posts in OSNs, which works by simply observing users' interaction with these posts. The proposed model is implicit, in the sense that no undue influence is exerted upon the observed users, and hence guarantees that the users' posting and reaction behavior is completely natural. We introduce a new paradigm for credibility prediction predicated on the interaction between users encountering the messages. Seeing as feature design and selection strongly impact a machine learning model's accuracy much more than the model used [7], we place emphasis on identifying the features that determine the spread of *True* posts, and those that determine the spread of *False* posts. We train a Bayesian Logistic Regression model by incorporating network, interaction and message features to measure the node-to-node influence dynamics to rumor propagation.

Existing research in rumor propagation and identification examine the behavior of misinformation posts over the network based on diffusion speed, depth, concentration, location, and sometimes combining features to differentiate posts. However, with access restrictions to the complete Twitter network graph and posts, it is important that we examine how individual users contribute to the diffusion of rumor posts and what features of the post sharer and receivers influence this paradigm. Since the spread of gossip is a uniform process, spreading from node to node [8], it is essential to note that the diffusion process is influenced not only by the creator of the tweet, but also by the sharer of the tweet.

In this paper, we describe two research problems and adopt an implicit crowdsourcing approach to addressing them:

1) We investigate credibility prediction by exploring rumor propagation founded on microscopic-level misinformation spread. By observing the spreading behavior of rumors in online social networks, we propose a model that predicts if a message is *True* or *False* by observing the latent attributes of the message, along with users and their reactions over the network.
2) We examine the contribution of individual users to rumor propagation in OSNs, by investigating features of users (both the post sharer and receiver) and how these features influence the propagation of rumor.

Previous crowdsourcing-based approaches in rumor detection focus on conversation annotation for credibility detection. We introduce a novel approach that explores crowdsourcing as an automated tool for identifying rumor in online social networks. We classify users based on the types of posts they generally react to: (i) reacts to only *True* posts, (ii) reacts to only *False* posts and (iii) reacts to a mix of *True* and *False* posts. Users in class i and ii are good discriminators for both credibility detection and feature identification, while users in class iii are treated as outliers that do not contribute much in the prediction model. The contributions of this paper are as follows:

- We introduce a new paradigm to rumor identification that applies implicit crowdsourcing for predicting credibility of a post without user annotation.
- We present a model to predict the truth-status of a tweet using the propagation pattern, and show that this model performs better than other state-of-the-art models.
- We demonstrate the abilities of the crowdsourced model by presenting a ranking of features relevant to rumor propagation.

The paper is organized as follows: Section II reviews the related work on misinformation diffusion, and features that aid misinformation spread in social networks. Section III describes our general approach, feature selection, and classification algorithms. Section IV elaborates on the experiment, data used, prediction model and evaluation metrics. Section V presents experimental results and observations, and finally, Section VI gives conclusions and insights into possible future works.

II. Related Work

The spread, detection and control of true and false news online continues to be a topic of interests to researchers in humanities, social sciences and engineering. In this section, we briefly discuss some of the studies and methods relating to rumor propagation and detection.

A. Crowdsourcing Techniques for Misinformation

CrowdFlower is a popular tool among researchers for labelling data for misinformation research. The authors of [9] used CrowdFlower to get a team of journalists to manually label tweets, with the annotators identifying only one of their specified features to support the truth status of the post. They used a feature scheme labeled as: support, response-type, certainty, and evidentiality. Their experiment showed that around 65% of the replies to original tweets were in the form of comments, which added little to the veracity of stories, while around 85% of tweets annotated had no evidence about the content being a rumor. In [10], the authors used CrowdFlower to label tweets as belonging to unsubstantiated information, disputed information, misinformation, reporting, linked disputes, or opinionated posts. Their analysis showed substantial disagreement in regard to posts that provide opinions, with a minority of assessors often describing them as containing disputed information, or being ambiguous.

A tool designed to allow journalists to identify and understand rumours quickly after they begin spreading on social media, using flags like ”Is this true?” is presented in [11]. These rumors are then displayed on a community website where users can up-vote them if they think they’re worth investigating further. The authors of [12] sort to automatically limit the spread of fake news by leveraging flagging tools added by Facebook. They proposed a model that uses Bayesian inference for detecting fake news and jointly learn about users’ flagging accuracy over time. They worked to determine posts that will impact potentially fake news and hand them off to experts to review and remove. The authors of [13] applied a combination of machine learning and crowdsourcing techniques to identify rumor spread on Zika virus, and proposed a model that combined sentiment analysis, linguistic, readability and unique medical domain features to distinguish between rumor and non-rumor tweets.

One of the challenges to crowdsourcing is to ensure workers provide objective and truthful reporting. To account for this, [14] proposed a bidding and incentive mechanism for mobile crowdsourcing. To guarantee trustworthy submissions, the authors applied Evolutionary Game Theory to ensure that the best strategy for workers was to submit trustworthy data. Each worker is assigned a reputation score, which begins at a maximum but is decreased if a worker submits untrustworthy data, and increased if the worker submits trustworthy data. Different tasks on the platform have different reputation thresholds, which workers must exceed to work on the task. This makes reporting trustworthy data the most stable strategy for workers.

B. Rumor Propagation

Research in political science explored the differential diffusion of true, false, and mixed (partially true, partially false) news stories on Twitter using the fact-checked rumor cascades that spread on Twitter over a 12-year period. In [15], they observed that falsehood diffused faster, farther, deeper and more broadly than truth in all categories of information, with a more noticeable impact in false political news. The study also observed that false news are often more novel, inspiring fear, disgust and surprise in replies while true stories inspired anticipation, sadness, joy and trust. In like manner, [16] examined the spread of fake news on Twitter during the 2016 U.S. presidential election and observed that the exposure to fake news sources was extremely concentrated with seven fake news sources accounting for more than 50% of fake news exposures. The study showed that political affinity was associated with the sharing of content from fake news sources and that the sharing of content from fake news sources was positively associated with tweeting about politics, and exposure to fake news sources. Computer scientists like the authors of [17] examined the spread of rumors on Facebook and found that rumor cascades run deeper in the social networks. When rumor debunking posts are available, [18] [17] reported that users will either delete a post, if it is confirmed to be rumor, or share otherwise. Additionally, [19] revealed that users spread the messages that they deem important and mostly retweet messages because of the need to retweet interesting tweet content or tweet creators.

To classify conversations within their formative stages, [6] proposed a rumor classification method to leverage implicit links to classify emergent conversations when very little conversation data is available. They used implicit links formed with hashtag and web links to establish similarity between otherwise unlinked conversations. The authors of [20] focused on the diffusion of information by inferring the embedding of social media users with social network structures; and utilize an LSTM-RNN model to represent and classify propagation pathways of a message.

C. Feature-based rumor detection

To demonstrate the importance of features for rumor detection, [21] extracted 68 features from tweets and categorized them as (1) message-based which considers characteristics of the tweet content, such as length of post, presence of exclamation, number of positive/negative sentiment words, (2) user-based which considers characteristics of Twitter users, such as registration age, number of followers, number of friends, and number of user posted tweets, (3) topic-based which aggregates the message-based and user-based features, and (4) propagation-based which considers characteristics related to the propagation tree that can be built from the retweet of the post. Subsequently, [22] explored rumor identification using users' behavior to differentiate between normal authors and rumormongers. Furthermore, [23] introduced the propagation tree, and used a random walk graph-kernel based hybrid SVM classifier to capture the high-order propagation patterns in addition to topic and sentiment features for rumor detection in Sina Weibo. In [24], the authors proposed two new features: (1) a client-based feature referring to mode of access – whether mobile or non-mobile – and (2) a location-based feature referring to the actual place where the event mentioned by the rumor-related microblogs happened – domestic (in China) or foreign. The work in [25] observed from rumor time series that rumors tend to have multiple and periodic spikes, whereas non-rumors typically have a single prominent spike, and proposed an automatic detection mechanism of rumor on Social Networks using Periodic External Shocks model. [26] analyzed the retweet network topology and found the diffusion patterns of rumors different from news. They also found that rumors tend to be questioned more than news by the Twitter community, suggesting that the Twitter community works as a collaborative filter of information. To show the role of emotional signals in fake news detection, [27] proposed a Long Short Term Memory (LSTM) model that incorporates emotional signals extracted from text to differentiate between credible and non-credible posts. Finally, [28] described a fake news detection model based on a dual emotion representations by simultaneously learning emotion representations for both the publishers and users of posts.

We improve on existing models by adopting an implicit crowdsourcing approach where we identify the truth-status of a post by simply observing the propagation style based on the interactions of users within the network and the hidden qualities associated with the message.

III. Features for rumor propagation and identification

Here, we describe a framework that given a tweet will predict (1) whether the tweet is *True* or *False* by observing user interaction with the tweet, (2) whether the followers of the spreader (could be the author or someone sharing) will react to the tweet in the form of a retweet, share, quote, like or favorite. We suggest 3 categories of features: message, interaction, network, and train a random forest classifier to rank the features in order of importance, then we build a Bayesian logistic regression model for classification. We adopt some of the features examined in the literature and suggest new ones, described below.

A. Network-based features

In microblogs such as Twitter, a *friend* is someone a user follows, and a user can see all of his friends' posts. In like manner, a follower is someone that follows and has direct access to all of a user's posts. We consider three features of the user's network: *followers count*, *friends count*, which have been extensively studied by [21] [22] [24], and *followers to friends ratio*, which was used in [23] to establish opinion leaders. These attributes are important because a user's friends impact the kind and volume of messages that end up in his timeline and the higher the number of followers, the farther the possibility of reach. This is also reflected in policies by OSNs like Twitter and Instagram who attach value to the followers count, where users become verified once they cross a certain threshold, even if the account holder is not a celebrity or public figure. Table I describes the network features used in the model.

TABLE I
NETWORK-BASED FEATURES

Feature	Description
followers count	higher count depict higher reach
friends count	# of accounts user follows
followers-friend	ratio to show influence in the network

B. Interaction-based Features

Since we are exploring rumor propagation as being dependent on the influence being wielded between users and taking propagation depth to be a factor of how messages cascade across the network, we examine the nitty-gritty of the followee-follower relationship to establish the features that influence the spread of rumor over the network. Here, we identify specific attributes of the user's online persona and posting behavior as determinant to being an influencer or influenced in the network. The assumption is that both the follower and followee contribute equally to the diffusion of a post, and an aggregate of network and message attributes tilt the reaction decision. Table II describes the 14 interaction attributes being considered. The last 5 features have been explored by [21], while we introduce 9 new features to the study of rumors in social networks. In addition to these, we describe social homogeneity (common to two users, showing an overlap in the sets of users they relate with, i.e. common friends and followers).

TABLE II
INTERACTION-BASED FEATURES

Feature	Description
shared friends	common nodes they interact with
directed tweets	ratio of tweets directed at someone
dialogue	active interaction from user 1 to 2
retweet-to-tweet	ratio of user's tweets with retweet
tweet wit hashtag	ratio of user's tweets that contain hashtags
tweets with url	ratio of user's posts with URL
tweets with media	ratio of user's posts with media
avg favorite-tweet	ratio of posts that get favorited
avg tweets/day	shows how active the user is
has url	does user's profile have a URL
has description	does user's profile have description
is verified	is the account verified
status count	volume of tweets over account's lifetime
account age	# of days since account was created

C. Message-based Features

Twitter posts are very fluid, taking up various forms as feedback, news, marketing campaigns, etc., so it is expected that rumors in this medium come in all forms. We account for this variation and consider the concealed form and intents of posts. Previous work have focused on count of positive and negative words in a tweet, with some exploring the polarity of the message sentiment but we look to explore the latent attributes of the message by introducing new features encompassing the type of post and emotion it is meant to incite. We adopt paralleldots API to perform content analysis on tweets to reveal the sentiment, intent, emotion and abusive attributes. Table III describe the message attributes - relating to the form, meaning and intent of the message, adopted in our model.

TABLE III
MESSAGE-BASED FEATURES

Feature	Description
quoted status	has post been quoted
is rt	has post been retweet
rt count	# of retweets
rt status	is post a retweet
favorited count	# of favorites
has hashtag	does post contain hashtags
has url	does post contain URL
has mentions	does post mention someone using "@"
has media	does post contain media
avg tweet length	length of tweet / 280 (max length)
positive sentiment	positive polarity of tweet
negative sentiment	negative polarity of tweet
neutral sentiment	neutral polarity of tweet
happy emotion	is post meant to incite happiness
fear emotion	is post meant to incite fear
sad emotion	is post meant to incite sadness
angry emotion	is post meant to incite anger
bored emotion	is post meant to incite boredom
feedback intent	is post meant to be a feedback
news intent	is post meant to be news
query intent	is post meant to be a query
spam intent	is post meant to be spam
marketing intent	is post meant for marketing
abusive	is post abusive

IV. EXPERIMENT SETUP

In this section, we describe the data collection process, prediction models and the metrics for evaluation. The approach is to (1) identify topics labeled as *False* (in other words, rumor) or *True* using *Snopes* [29] – an online fact-checking site – and collect Twitter posts about the topic. (2) To each user, we associate a total of 17 features, to include 3 network and 14 interaction attributes; and to each message, we associate 24 attributes: 10 observable and 14 latent attributes. We then train a Bayesian logistic regression model based on the prediction task. In Figure 1, we present an abstraction of the experiment setup for the credibility prediction task.

A. Data Collection

We used Snopes [29] to identify topics that have been fact-checked and rated as *True* or *False*. Even though Snopes has different categories including those labeled "Mostly True" and "Mostly False", we restrict this research to those that are strictly labelled *True* or *False*. For each topic, we assign a set of keywords and crawl the Twitter search API using queries of the form $(K_1 \vee K_2 \vee K_3)$, similar to that described by [30] but with K_i representing the conjunction of possible keyword combinations. For instance, the topic "In a leaked e-mail, Hillary Clinton said "we must destroy

TABLE IV
TOPICS IDENTIFIED FROM SNOPES, ALONG WITH THE ASSOCIATING KEYWORDS USED IN QUERYING THE TWITTER SEACH API

Category	Topic	Keywords	# tweets
False	Hillary Clinton said "we must destroy Syria for Israel"	Hillary, destroy, syria	96724
	The FBI discovered bones of young children in Jeffrey Epstein's private island	epstein, bones, children	189812
	Odessa shooter was "a Democrat Socialist who had a Beto sticker on his truck."	odessa, shooter, beto, democrat, sticker	19491
True	Blood spots visible in the left eye of Joe Biden during a CNN dxebate in Sept. 2019	Joe, Biden, blood, eye	38983
	Anti-abortion Rep. DesJarlais encouraged some women to have abortions	abortion, Desjarlais, mistress, republican	15591
	Video shows air traffic over the US on 9/11 as thousands of flights were grounded after a terrorist attack	flights, grounded, after, 9/11	10355

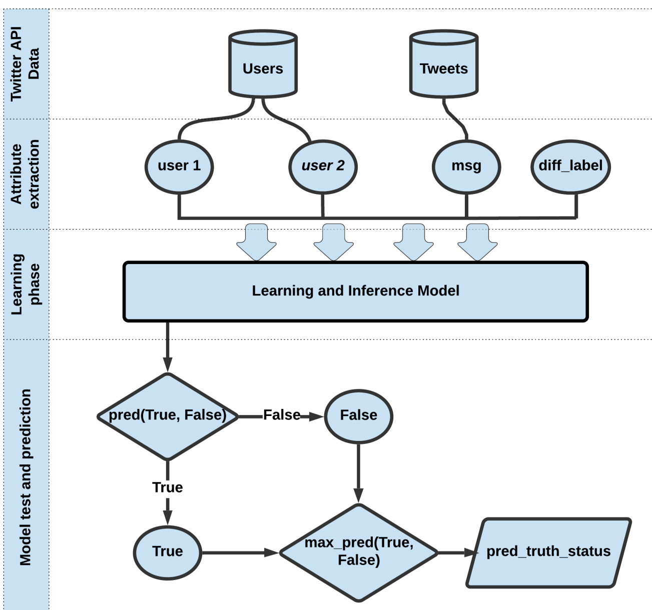

Fig. 1. An illustration of experiment setup for the credibility prediction task

Syria for Israel."" had keywords "hillary, destroy, syria" and query $((hillary \wedge destroy \wedge syria) \vee (hillary \wedge destroy) \vee (hillary \wedge syria) \vee (destroy \wedge syria))$. Table IV gives a breakdown of our topics, along with the associating keywords and number of tweets (including retweets). For reproducibility and future adoption, we make the crawler publicly available for researchers on GitHub [31].

In the dataset, we found a large variation in the volume of tweets in the *True* and *False* collections, with *False* posts accounting for more than 80% of the entire dataset. Also, we observed that the propagation depth of *False* posts ran deeper, with an average retweet depth of 4, while *True* posts averaged a retweet depth of 2. Lastly, we observed a "diffused"/"not diffused" ratio of 35/65 for the tweets in the collection of *True* topics and 45/55 for tweets in the collection of *False* topics. This difference in diffusion rates reveals that *False* posts tend to have more reaction-to-post than *True* posts.

B. Prediction Models

Given a collection of messages and an associated user, we recreate the Twitter followership graph by connecting all of the user's followers. Based on the assumption that users will interact with their friends' messages uniquely, we assign the diffusion label as a function of the reaction observed per message and show that this microscopic-level information spread based on the latent message and user interaction attributes is sufficient to give insight to the credibility of a message. We perform two supervised learning tasks by adopting two off-the shelf machine learning models: Bayesian Logistic Regression - for the prediction tasks and Random Forests - for feature selection.

1) Predicting credibility of posts: We train a model that predicts if a message is *True* or *False*. We extract the features described in Section III, and additionally include the diffusion property as an independent variable during the training phase. More specifically, an edge is said to be *diffused* if and only if the destination user (in Twitter terms: *follower*) has reacted (reply, retweet, quote, like) to the friend's (*followee's*) post. We examine how users on Twitter relate with posts of their friends by building classifiers to distinguish user interactions based on the credibility of the message. For a message m, where $m \in \{1, \ldots, M\}$, spread over a network with n interactions, we train a model that predicts the truth status of the message based on the diffusion behavior observed along each one of the n links along which the message propagates. The predicted output is the majority truth status observed across the n interactions. For instance, if a *True* message is spread over 5 interactions and the model predicts the post to be *True* 3 out of 5 times, we accept the output to be *True* and evaluate the model over its correct classification of M messages in the test collection.

2) Predicting rumor propagation: To further demonstrate the differences in the propagation of *True* and *False* posts, we perform a node-to-node analysis between a pair of users, the spreader and receiver, examining each user's posting behavior, and their interactions to predict the receiver's reaction. Here, we aim to show that our model performs well in an established environment, in order to compare with previous models for propagation prediction. This task is valuable to strengthening our hypothesis that the propagation behavior is a significant attribute to predicting the credibility of a message based on how users in OSN interact with posts of varying veracity.

First, we build separate models for *True* and *False*, performed a supervised learning task using the Bayesian logistic regression by assigning diffusion label "diffused" between a spreader and his follower, if the follower has reacted to an identified tweet (in either case, *True* or *False*) and "not-diffused" otherwise. We adopt an 80-20 train-test split of the data and account for over-fitting by performing 10-fold cross validation. We make predictions on the capability of the model to correctly predict diffusion on the message type and take it a step further by investigating the model's ability to generalize across message type. Then, we build a Random Forests classifier to analyse the importance of the input features and perform selection on the best features for rumor propagation and identification tasks.

C. Evaluation Metrics

The prediction capabilities of the learned model are tested based on its abilities to predict if there is diffusion across an edge given the learned model. We use standard classification evaluation metrics: precision, recall and F score, to assess the efficiency of our model.

Precision describes the ratio of instances correctly classified as "diffused" to the total classified as "diffused", and is estimated as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall is the ratio of instances correctly classified as "diffused" to the total number of instances that "diffused", and is estimated as:

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

Where TP (true positives) is the number of instances correctly classified as "diffused", FP (false positives) is the number of instances incorrectly classified as "diffused", and FN (false negatives) is the number of the instances incorrectly classified as " not diffused".

The **F score** is the harmonic mean of the precision and recall. It is computed as

$$Fscore = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

D. Baseline

We compared the performance of our proposed model to state-of-the-art models in predicting the credibility of posts in social networks.

1) Emotion-based: [28] exploits the emotions of both the publisher and receiver of contents to classify posts as fake or not.

2) Implicit-link: [6] use hashtags and web linkage method to link conversations. We tested using the linking method without pruning.

3) User-behavior: [22] describes a user behavior-based rumor identification scheme, in which the users' behaviors are treated as hidden clues to identify rumor posts in microblogs.

V. Experimental Results

In this section, we report the results obtained from each phase of the experiment.

A. Predicting credibility by implicit crowdsourcing

While some users react to posts of varying credibility, others only react to tweets that are precisely *True* or *False*. So training a model that learns to distinguish this interaction-reaction relationship is useful for identifying the credibility of a tweet by observing the reaction of a user based on the established interaction between the users. By incorporating the diffusion status of a tweet, we train a model to predict the credibility of the message. The objective of the task is to show that collating the implicitly sourced diffusion behavior between users is useful for predicting the credibility of a post. This implicit crowdsourcing approach is important in real-world situations where there is a need for the system to passively interact with the network. A passive interaction is crucial especially in systems requiring real-time and undetectable communication, for example, an automated rumor identification system for social networking websites.

TABLE V
Model performance for predicting credibility of a tweet using crowdsourcing techniques

Model	Precision	Recall	F
Crowdsourced	**0.919**	**0.903**	**0.911**
Not-Crowdsourced	0.838	0.801	0.823
Emotion-based	0.798	0.832	0.815
Implicit-link	0.861	0.713	0.780
User-behavior	0.753	0.873	0.809

The result from our experiment validate our assumption that the difference associated with the message, interaction and diffusion patterns of *True* and *False* posts can be exploited in predicting the credibility of messages. By combining these attributes and using the F-score as a measure of accuracy, we were able to achieve 91% accuracy in identifying whether messages are credible or not, see Table V. It is important to note that the model is tested using labelled data with existing ground-truth. To show the impact of the diffusion attribute to the credibility prediction task, we carried out a parallel credibility identification task without the diffusion label and observed a performance of 82%. We also show that a comprehensive model exploiting the attributes of the network, interaction and message will perform better than those that use one or the other.

B. Features analysis for rumor propagation

Establishing a difference in the diffusion prediction models for *True* and *False* posts is amply dependent on showing that there exists a difference between these types of messages and the attributes that steer user reactions. For us to efficiently apply a crowdsourcing approach to the detection of misinformation, we need to differentiate the attributes of *False* posts from those of *True* posts, before we can demonstrate that they diffuse differently. Differentiating between this diffusion pattern is beneficial for the early detection of rumor to mitigate its spread and effect within the network. One justification for using multivariate methods is that they take into account feature redundancy and yield more compact subsets of features, as features that are individually irrelevant may become relevant when used in combination, which also shows that correlation between sets of features does not necessarily imply redundancy. Considering that the goal of the feature analysis task of this study is to identify the optimal set of features necessary to maximize diffusion prediction irrespective of credibility-status, we train a random forests model and then select the top 20 features for the rumor propagation tasks.

TABLE VI
TOP 20 FEATURES FOR EFFICIENT DIFFUSION PREDICTION OF *True* AND *False* POSTS SELECTED USING RANDOM FOREST CLASSIFIERS

Rank	False	True
1	MSG is RT	MSG is RT
2	MSG favorited count	social homogeneity
3	MSG has mentions	MSG favorited count
4	dest tweet with hashtag	src tweets with URL
5	src retweet-to-tweet	MSG feedback intent
6	MSG news intent	MSG positive sentiment
7	src followers count	src directed tweet
8	MSG has URL	MSG has URL
9	src followers-friends	src avg favorite-tweet
10	src account age	src avg tweet/day
11	src tweets with URL	src followers count
12	MSG fear emotion	MSG has mentions
13	dest directed tweet	src account age
14	src status count	dest retweet-to-tweet
15	src friends count	src retweet-to-tweet
16	social homogeneity	src has URL
17	MSG RT count	dest follower-friends
18	dest friends count	src status count
19	MSG positive sentiment	MSG has hashtag
20	MSG negative sentiment	MSG RT status

From the ranked features in VI, we see that in tweets with *False* status (Rumor), the attributes of the message account for 45% of the ranked features with the combination of network and interaction accounting for 55%, while message attributes account for 40% of top ranked features for *True* posts. As anticipated, the latent attributes of the message rank in the top features for both *True* and *False* models, confirming that the meaning, intention and emotions of messages influence users' decisions in the diffusion process. From the ranked features, we can infer that rumor posts masked as news, meant to incite fear will diffuse better than others. However, it is surprising that the diffusion of rumor posts cannot be strictly tied to their sentiment as we observed that both negative and positive sentiments contribute equally to the performance of the model. Even though it ranks differently in both models, social homogeneity ranking well in both models shows that a user will most likely respond to the post of someone with interests similar to his own.

C. Predicting Rumor Propagation

We focus on the problem of predicting the diffusion decision (to react or not) of a user based on his perception of the message and interaction with the spreader of the information. In this model, we do not take into account the effect of previous exposure to similar posts, or the popularity of the message, we simply make an inference on whether a user will retweet, share, quote or favorite a tweet by estimating the probability of diffusion.

In Table VII, we show the performance of the model across message type, using the performance metrics previously highlighted. The model achieved 91.6% and 89.9% prediction accuracy for message with *True* and *False* status respectively.

TABLE VII
MODEL PERFORMANCE FOR PREDICTING DIFFUSION OF *True* AND *False* POSTS OF A POST

Model	Precision	Recall	F
False	0.897	0.902	0.899
True	0.908	0.925	0.916

To show that the proposed model can be effectively transferred across topics and credibility status, we tested our model's performance over topics outside the training list. The results for inter-topic and inter-credibility prediction tasks are reported in Table VIII. For inter-topic test, we observed performance of similar magnitude in diffusion prediction capabilities when the models are exposed to topics outside the training list. As observed from the table, there is a difference for inter-credibility test and we believe this is due to the difference in the features that influence diffusion for the message types. This result piques our interest because it shows that the properties of *True* and *False* posts are distinct enough that either model can discriminate significantly between each type of post.

TABLE VIII
MODEL PERFORMANCE FOR INTER-TOPIC, INTER-CREDIBILITY DIFFUSION PREDICTION

Model	Precision	Recall	F
False	0.887	0.889	0.882
True	0.899	0.919	0.908
False model-True test	0.856	0.821	0.838
True model-False test	0.849	0.921	0.884

VI. CONCLUSIONS AND FUTURE WORK

In this study, we hypothesize that there is a difference in the spreading behavior of rumor and truthful information in OSN. We presented a model based on Bayesian logistic regression to predict the credibility status of a message by simply crowdsourcing the interaction and propagation behaviors of similar messages. The crowdsourcing detection model integrates information diffusion by using the diffusion label ("diffused" or "not diffused") associated with the node-to-node interaction between a pair of users. This diffusion label is then combined with the user and message attributes to predict the credibility status for that edge. The credibility status of a particular message is aggregated over all the edges in the network. We showed that rumor is mostly masked as news content, meant to incite fear emotions in the reader with mixed sentiments, and that the diffusion attribute is significant to predicting the credibility of a tweet.

To identify the credibility of a post especially in the conversation emergent stage where there are not enough posts on the topic or a veracity source, users who interact with specific types of posts serve as good discriminators of credibility. A system looking to efficiently identify the truth status associated with a message will benefit from a comprehensive model exploiting the attributes of the network, interaction and message rather than focusing on just the content of the post. In the future, we hope to adapt this model to topics that have mixed content to observe how this paradigm affects the model. Our goal is to estimate just *how much* of the message is true.

More generally, implicit crowdsourcing of certain detection and estimation tasks is a severely under-explored area of artificial intelligence, and one that is bound to become a very important paradigm in a near future in which human skill and AI blend together to facilitate data interpretation and discovery at increasingly higher speeds and complexities.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation under grants No. 1527579 and 1619201.

References

[1] Twitter, "How the usgs uses twitter data to track earthquakes," 2015, last accessed 21 January 2020. [Online]. Available: https://blog.twitter.com/en_us/a/2015/usgs-twitter-data-earthquake-detection.html

[2] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic," *PloS one*, vol. 8, no. 12, 2013.

[3] CNN-Business, "Man faces fallout for spreading false sandy reports on twitter," 2012, last accessed 21 January 2020. [Online]. Available: https://www.cnn.com/2012/10/31/tech/social-media/sandy-twitter-hoax/index.html

[4] S. Soames, "The truth about deflationism," *Philosophical Issues*, vol. 8, pp. 1–44, 1997.

[5] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 1589–1599. [Online]. Available: https://www.aclweb.org/anthology/D11-1147

[6] 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*, ser. CIKM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 2377–2382. [Online]. Available: https://doi.org/10.1145/2983323.2983697

[7] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato Hamilton, 1999.

[8] B. Pittel, "On spreading a rumor," *SIAM Journal on Applied Mathematics*, vol. 47, no. 1, pp. 213–223, 1987.

[9] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie, "Crowdsourcing the annotation of rumourous conversations in social media," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery, 2015, p. 347–353. [Online]. Available: https://doi.org/10.1145/2740908.2743052

[10] R. McCreadie, C. Macdonald, and I. Ounis, "Crowdsourced rumour identification during emergencies," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery, 2015, p. 965–970. [Online]. Available: https://doi.org/10.1145/2740908.2742573

[11] P. Resnick, S. Carton, S. Park, Y. Shen, and N. Zeffer, "Rumorlens: A system for analyzing the impact of rumors and corrections in social media," in *Proc. Computational Journalism Conference*, vol. 5, no. 7, 2014.

[12] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, "Detecting fake news in social networks via crowdsourcing," *arXiv preprint arXiv:1711.09025*, 2017.

[13] A. Ghenai and Y. Mejova, "Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter," *CoRR*, vol. abs/1707.03778, 2017. [Online]. Available: http://arxiv.org/abs/1707.03778

[14] Y. Wang, Z. Cai, G. Yin, Y. Gao, X. Tong, and G. Wu, "An incentive mechanism with privacy protection in mobile crowdsourcing systems," *Computer Networks*, vol. 102, pp. 157–171, 2016.

[15] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[16] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 us presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[17] A. Friggeri, L. Adamic, D. Eckles, and J. Cheng, "Rumor cascades," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[18] T. Takahashi and N. Igata, "Rumor detection on twitter," in *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*. IEEE, 2012, pp. 452–457.

[19] N. A. Abdullah, D. Nishioka, Y. Tanaka, and Y. Murayama, "User's action and decision making of retweet messages towards reducing misinformation spread during disaster," *Journal of Information Processing*, vol. 23, no. 1, pp. 31–40, 2015.

[20] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, ser. WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, Inc, Feb. 2018, pp. 637–645, 11th ACM International Conference on Web Search and Data Mining, WSDM 2018 ; Conference date: 05-02-2018 Through 09-02-2018.

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

[22] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, 2015.

[23] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st international conference on data engineering*. IEEE, 2015, pp. 651–662.

[24] 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*. ACM, 2012, p. 13.

[25] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 1103–1108.

[26] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 71–79.

[27] A. Giachanou, P. Rosso, and F. Crestani, "Leveraging emotional signals for credibility detection," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 877–880. [Online]. Available: https://doi-org.er.lib.k-state.edu/10.1145/3331184.3331285

[28] C. Guo, J. Cao, X. Zhang, K. Shu, and H. Liu, "Dean: Learning dual emotion for fake news detection on social media," *arXiv preprint arXiv:1903.01728*, 2019.

[29] snopes.com, "Snopes.com," 1994, last accessed November 2019. [Online]. Available: https://www.snopes.com/

[30] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 1155–1158.

[31] A. Osho, “Github ariseabiola/misinfo_diffusion: Explore the dynamics of microscopic-level misinformation topology in online social networks.” 2019. [Online]. Available: https://github.com/ariseabiola/misinfo_diffusion