A Multiple Feature Category Data Mining and Machine Learning Approach to Characterize and Detect Health Misinformation on Social Media

Lida Safarnejad, *Stanford University, Stanford, CA, 94305, USA*
Qian Xu, *Elon University, Elon, NC, 27244, USA*
Yaorong Ge and Shi Chen, *University of North Carolina at Charlotte, Charlotte, NC, 28223, USA*

In this article, we characterize health misinformation infiltration as a dynamic dissemination process on social media in addition to content-based features. Using Zika discussion on Twitter in 2016 as the study system, we identified 264 most influential tweets with misinformation and matched 455 tweets with real information. We developed an algorithm to infer information dissemination network through retweeting for each tweet, and extracted nine network metrics. We then approximated information dissemination as nonhomogeneous Poisson process (NHPP) signal. We then extracted 40 signal features to characterize each NHPP. For content-based features, we applied both linguistic inquiry and word count and document-to-vector to further extract 63 and 50 features for each tweet, respectively. Finally, we also considered four user features. Based on these extracted feature categories, we trained support vector machine and random forest (RF) classifiers. Using all feature categories combined as input, an RF classifier achieved $>$ 83% accuracy and $>$ 90% AUC to detect misinformation.

Social media platforms have provided public health professionals with valuable resources to study public opinions toward various health-related issues, and to use social media as one of the main outlets to efficiently and effectively spread accurate and timely information, especially during health crisis such as Zika epidemic in 2016 and the current COVID-19 pandemic.[6,16]

Nevertheless, social media have also opened the room for misinformation and facilitated its infiltration and proliferation on the Internet during health emergencies.[20] Health misinformation is generally considered to be misleading, incorrect, not evidence-based, and malicious. These different types of health misinformation cause unnecessary confusion, anxiety, and anger of individuals rupture the society and seriously undermine health professionals' efforts in providing evidence-based knowledge and combating the real epidemic. WHO stated that it needed to fight two epidemics at the same time during the 2014 Ebola epidemic, one in the real world and the other on the Internet, especially on social media.[1] The 2016 Zika epidemic was another example where misinformation proliferated on social media, including Facebook and Twitter.[6] Unfortunately, increasing exposure to misinformation online and on social media can reinforce incorrect beliefs of users,[2] making misinformation difficult to eradicate. Therefore, it is equally important, if not more, to reduce the influence of health misinformation on social media during health emergencies alongside curbing the epidemic itself. The first and foremost critical task is to accurately identify health misinformation on social media in order to neutralize them before they inflict harm to users.

Nevertheless, detecting health misinformation among the large volume of user-generated information on social media is a challenging task. State-of-the-art misinformation detection systems are generally based

1089-7801 © 2021 IEEE
Digital Object Identifier 10.1109/MIC.2021.3063257
Date of publication 5 March 2021; date of current version 29 September 2021.

on fact-checking content of social media posts. There are several challenges associated with this approach. First, our knowledge of health issues increases rapidly, especially during an emerging health crisis. Such a fast pace of knowledge refreshing can make certain "knowledge" incorrect when new evidence emerges. For example, Zika was once assumed to transmit only through mosquito biting and from pregnant mother to fetus, but later evidence confirmed the possibility of sexual transmission as well. Second, fact-checking is only useful for objective contents but not effective against more subjective hate language, extreme bias, and discrimination against certain groups of people. Nevertheless, health, especially during large pandemics, is always confounded by complicated social, political, and economic issues. These convoluted issues make content-based fact-checking less useful against various types of misinformation. Third, content-based misinformation detection systems ignore other important features of health misinformation. Similar to a real epidemic, digital epidemic of misinformation is a multiaspect problem. The real pathogen, host, and environment together form the epidemiological triad, the foundation of an epidemic. We can tackle real epidemics more effectively when combining the power of etiology, epidemiology, immunology, and pathophysiology from this triad. Similarly, health misinformation, general users, and the social media environment also collectively form the infodemiological triad.[8] We suggest that a more holistic and accurate characterization of health misinformation is essential to develop a health misinformation detection system.

In this study, we present a novel perspective in characterizing health misinformation on social media from multiple aspects. We further develop a data mining approach to extract new features that distinguish health misinformation from real ones, and develop more effective misinformation classifiers. First, we extract information dissemination features by constructing each tweet's retweeting networks to compute network metrics. In addition, we model the retweeting process of each tweet as a nonhomogeneous Poisson process (NHPP) signal and extract signal-based features. Then we consider content and linguistic features and apply linguistic inquiry and word count (LIWC) and document-to-vector (Doc2-Vec). Finally, we extract user features (e.g., number of followers, friends, and verification status) as well.

After extracting these different categories of features representing various aspects of health (mis) information, we develop classifiers using the most retweeted tweets containing real and misinformation about Zika in 2016. Different categories of features and their combinations are investigated as inputs to develop support vector machine (SVM) and random forest (RF) classifiers. We compare the performance of these models and evaluate the influence of input feature categories on classifiers' performance.

RELATED WORKS

Recent studies have been exploring new features of health misinformation. Monitoring user activity on social media can detect suspicious online behavior and identify potential bots. While these content-based and user-based features are the key aspects to develop misinformation detection systems,[9] other types of features of health misinformation have also been explored.[11,12,17] Two main limitations of the previously proposed approaches are heavily relying on historical users' activities on social media, which might not be readily available, and moreover, considering only the content-based features while ignoring other aspects of the multidimensional misinform detection problem. For instance, Qazvinian *et al.*[14] constructed a probability distribution over retweeters of rumors and used it in addition to content-based features to predict rumors. Another work investigated rumors on Sina Weibo, the largest social media platform in China, and proposed a combination of propagation-based, location-based, and client-based features.[21] Other studies investigate topic network features in addition to content-based features.[5] More recent studies also apply an array of machine learning (ML) and AI methods, time series analysis, and network analysis to build health misinformation detectors,[9] track misinformation dissemination dynamics,[7,10,19] and evaluate role of users (e.g., bots, trollers, opinion leaders) who relay misinformation.[3,4] These studies provide alternative perspectives and technical approaches to identify health misinformation on social media.

DATA RETRIEVAL AND ANNOTATION

Data Retrieval

We used the Gnip API through UNC Charlotte School of Data Science to retrieve all English tweets with keyword *Zika* and other related keywords such as *microcephaly* and *PHEIC*. The entire year of 2016 (January 1 to December 31, 2016) was chosen as the sampling period for this study to bracket the entire WHO Public Health Emergency of International Concern (PHEIC) period from February 2 to November 18, 2016. This time period also covered major milestones in the Zika epidemic timeline, including WHO's initial warning of Zika epidemic across the Americas, the official PHEIC declaration of Zika pandemic on February 1, opening of Rio summer Olympics from August 5 to August 21, and the end of the PHEIC on November 18, 2016. A total of 3.7 million English tweets, retweets, and their associated metadata were collected in 2016. This dataset was

the complete dataset covering all English discussions regarding Zika on Twitter in 2016, unlike the common 1% sampled data from Twitter's own API. Therefore, our dataset provided a more comprehensive, complete, and less biased view of Zika-related tweets.

Tweet Annotation

We ranked all original Zika tweets based on the number of received retweets, from highest to lowest. We define a tweet to be highly influential if received retweets from ≥ 50 distinct retweeters. Based on this criterion, we selected the top 5000 most retweeted Zika tweets as the sample pool. We have established an operational definition of misinformation regarding Zika, such that the content was not evidence-based, in accordance with the commonly used term evidence-based medicine in health domain. Peer-reviewed journal articles and conference proceedings, government and health agencies (e.g., CDC and WHO) reports and statistics, fact-checking websites, are all used to evaluate and cross-check the tweets. Two independent researchers established strict intercoder reliability ($>95\%$) on a randomly selected test set of 100 tweets before proceeding to annotate the remaining tweets. A total of 264 tweets were finally included as the misinformation group with complete metadata. Another 455 tweets were identified as real information group, controlling the effect of posting time and number of retweets to make them comparable to misinformation group. A detailed description of this annotation and misinformation identification process is provided in our previous study.[15]

FEATURES EXTRACTION

Information Dissemination-Based Features

We extracted novel features of these tweets and paved the way for further classification. In the first part of feature extraction, we considered (mis)information as a dynamic information dissemination process through retweeting. We explored two different but interrelated angles: 1) structural features of information dissemination network of retweeting, and 2) time series features of information dissemination signal. For the first feature category, we developed an algorithm to reconstruct and infer the retweeting network. For the second category, we constructed a signal of retweeting process and approximated it as a NHPP for each tweet. These two categories of features characterized (mis)information dissemination among users on social media.

Dynamic Retweeting Network Construction and Feature Extraction

We considered retweeting occurs in a network of retweeters for a given tweet. Such a retweeting network can be mathematically modeled as a graph G with two sets of elements: a set of nodes (or vertices, representing users), denoted by V, and a set of edges, denoted by E (representing retweeting sequence). $G = \{V, E\}$. Every edge e in E connects a pair of vertices v_i and v_j. In this study, pairs of vertices connected by edges were ordered (i.e., direction of the edge is relevant), and the graph G was a directed graph. To construct an information dissemination network, we considered the followers' relationship among each retweeter and time difference of two consecutive retweets. For every pair of retweeters (v_i, v_j) there was a directed edge e from v_j to v_i if and only if 1) v_j followed v_i, and 2) v_j had retweeted the tweet after v_i. A detailed description of this algorithm is provided in our prior work.[15]

Once the retweeting network was constructed for each tweet, we further extracted critical network metrics for information dissemination in both real and misinformation groups. A concise description of these network metrics is described in Table 1. Network metrics between the two groups were compared by the Kolmogorov–Smirnov test to detect distribution difference between groups. We identified a total of nine network features with significant differences between real and misinformation groups. These nine features comprehensively characterize dissemination network structure from a global network level (density, reach, diameter, virality, influence, Wiener index) to the local cluster level (modularity) and down to the individual vertex level (top degree and top betweenness centrality score). The details and their relevance to information dissemination are in our prior work.[15]

Evaluating the differences of these network features among real and misinformation groups provided insights on network dissemination differences, which is the cornerstone to build the classifier.

Retweeting Signal Construction and Feature Extraction

Network features characterized (mis)information dissemination among *users*. While we used time difference of retweeting to infer dissemination network, temporal aspect was not explicitly considered in network features. Nevertheless, information dissemination, like a real epidemic, is a dynamic process where temporal aspect is a critical factor. To more accurately capture the temporal dynamics of the retweeting process, we further constructed retweeting signals for each tweet, approximated them as NHPP, and extracted signal features thereof.

We constructed a time series of each tweet's retweeting stream, referred to as retweeting signal hereafter. The signal can be mathematically described as $Y = Y_t, t \in T$, where Y_t is the number of retweets received during time t. Our initial exploration revealed a

TABLE 1. descriptions of network features.

Network Feature	Description
Network Reach	The number of unique vertices, i.e., unique IDs
Network Influence	The number of total retweets
Network Diameter	The shortest distance between the two most distance vertices
Network Density	The existing proportion of potential relationships among vertices
Structural Virality	The average distance between all pairs of vertices
Wiener Index	The sum of the shortest paths between all pairs of vertices
Network Modularity	The likelihood of dividing a network into potential clusters
Largest Out-degree	The largest out-degree centrality of all vertices in a network
Largest Betweenness	The largest betweenness centrality of all verticesin a network

substantial temporal variability between real and misinformation groups.[15] It took significantly less amount of time for misinformation to receive 50% of all retweets but much longer time to receive 90%, 95%, and 100% retweets than real information group. We hypothesized that time to receive 50%, 75%, 90%, 95% retweets were useful to detect misinformation retweeting signal.

In addition, we detected peaks in retweeting signals to identify the time periods when a tweet was highly attractive and received a large number of retweets in a short period of time. Retweeting signals usually had more than one peaks, i.e., information relay. This complied with our previous findings that Twitter discussions could be promoted by bots or influential users from time to time.[16] We found that time difference between two consecutive peaks p_i and p_{i+1} approximated exponential distribution. Therefore, we modeled retweeting signal between two consecutive peaks as a specific homogeneous Poisson process (HPP), and the entire retweeting signal as a NHPP approximated by the union of n underlying HPPs:

$$\text{NHPP} = \bigcup_{i=1}^{n} \text{HPP}_i.$$

Each HPP_i started at ith peak p_i and continued until immediately before the next peak, when the next homogeneous Poisson process HPP_{i+1} started. We characterized several important aspects of HPP_i: the peak p_i, the valley v_i, indicating the end of the process, and the HPP_i rate λ_i. We defined the valley v_i to be the time of the minimum number of retweets farthest from the peak p_i and closest to the p_{i+1}, where the next HPP_{i+1} started. According to our exploration, a retweeting signal always had at most five peaks. Therefore, to make it consistent across all retweeting signals, we set n, the maximum of the number of HPPs for a signal, as 5. For each retweeting signal, we extracted the following features of its underlying HPP_i process.

- Peak time: the time when HPP_i started.
- Peak height: number of retweet counts at peak i.
- Process width: the time difference between the peak and valley in the same HPP_i.
- Valley time: the time when HPP_i ended.
- Valley height: number of retweet counts at the ith valley.
- Full width half max (FWHM): the first time point after each peak where the number of retweets was less than or equal to the half of the peak height.
- FWHM height: number of retweet counts at the ith FWHM.
- λ_i: the HPP_i rate.

By extracting and comparing signal features, we obtained a more quantitative characterization of the temporal aspect of (mis)information dissemination on social media.

Content-Based Features

In addition to information dissemination feature categories, we also explored and extracted various content-based features. Contents in misinformation, similar to the actual pathogens in a real transmission chain, interacted with human users (hosts) and exert the influence psychologically. In this study, we focused on textual content and exclude other contents such as pictures, memes, and GIFs. Two major categories of content features, LIWC and Doc2Vec, were extracted to more comprehensively characterize content's topic, semantic and linguistic aspects.

LIWC: Linguistic Inquiry and Word Count Feature Extraction

To extract socially and psychologically relevant features from tweet content, we applied the widely used LIWC tool.[18] LIWC summarized the tweet in various categories of emotional and cognitive patterns. Categories were represented and quantified by linguistic features of word counts and statistics. Examples of such features included positive or negative emotions, social relationships, social coordination, and individual differences. These features were more subjective for

characterizing tweet content. According to our exploration of most retweeted Zika tweets, emotions (e.g., anger, confusion, fear, and mistrust) and social issues (e.g., bias and discrimination towards certain demographic groups) were profound along with this emerging disease and its discussion on social media.

After applying LIWC, each tweet received a numeric vector of 63 features using LIWC software. These features were further compared between real and misinformation groups. Such insights would be useful for classifier to differentiate real and misinformation groups.

Doc2Vec: Document-to-Vector Feature Extraction

LIWC provided a quantitative way to characterize, extract, and quantify linguistic features. While LIWC vastly expanded our understanding of computational linguistics and paved the way for feature extraction, it is more subjective (e.g., related to social and psychological aspects), relied on a human-developed dictionary, and allowed for little flexibility of incorporating new insights, especially on emerging health issues.

Word embedding techniques, on the other hand, provided a more direct and sophisticated representation, including semantics and relationships among words within the content. Word embedding used neural networks (NN) to learn and represent features of content. Each word is mapped into a predefined vector space where the values in the vector were learned from NN. This ML technique allowed for more natural capturing and representation of the content, comparing to LIWC. While word embedding also used a corpus (dictionary) to generate vector space, the corpus size was much larger than the one in LIWC and did not have human-assigned labels, thus, allowing much more detailed and objective characterization of the content.

However, word embedding techniques had some shortcomings as well. They were not effective when analyzing short texts with different lengths, as in the case of tweets.[13] In this study, we applied an extension of the original word embedding technique, Doc2Vec proposed by Le *et al.*[13] In this new method, documents regardless of length were embedded into a predefined vector space. We treated each tweet as a document and applied Doc2Vec. After initial evaluation and preprocessing (e.g., dropping nontext components such as figures, memes, and videos), we converted each input tweet to a document vector with a size of 50, i.e., 50 features in Doc2Vec vector space. These features could be explicitly compared between real and misinformation groups for further classification task.

User Feature Extraction

We extracted user features from their profiles, including number of followers, number of friends, verification status, and percentage of retweeters with verified status. These features were not associated with a specific tweet but to the user ID. Studies have used user features to detect potential malicious IDs, which frequently sent out misinformation. However, we suggested that this approach could have both false negative and false positive issues. Our more intensive exploration had identified some IDs which frequently sent out misinformation (e.g., @naturalnews, a far-right conspiracy theorist account); however, not all tweets it generated were misinformation. On the other side, some seemingly credible sources (e.g., @theEconomist) also sent out inaccurate information due to a lack of understanding of the evolving health issue.[16] We used user features in conjunction with tweet-specific features to provide a more comprehensive characterization of health misinformation.

Summary of Feature Categories

In summary, we have identified and extracted five (5) categories of features: 9 dynamic network features and 40 signal features related to information dissemination process through retweeting; 63 LIWC and 50 Doc2Vec features based on actual content of each tweet. In addition, the fifth group of four user-based features is also extracted. Eventually, we extract 162 features for each tweet, both real and misinformation in this study, and perform further ML classification based on these categories of features.

CLASSIFICATION BASED ON MULTIPLE FEATURE CATEGORIES

We further built two types of supervised ML classifiers, RF and SVM. We defined confirmed misinformation as *positive* while real information as *negative* in this study. Though there are many other classifiers available, we suggested the merit of this study was to provide a more comprehensive characterization of health misinformation challenge on social media, and a data mining approach to extract new features. Therefore, the main goal of this study was not to develop new classifiers nor to systematically explore different classifiers' performance, but to develop effective classifier based on commonly used and proven robust algorithms such as RF and SVM.

For each type of classifiers (RF and SVM), we first included features from a single feature category. Then we built another set of classifiers by merging different feature categories. We ran k-fold cross-validation for each classifier we built. The data were split into 10-folds ($k = 10$) and randomly picked nine slices to train the model while using the remaining unseen data to cross-validate the model. This process was repeated 10 times such that each slice was used exactly once

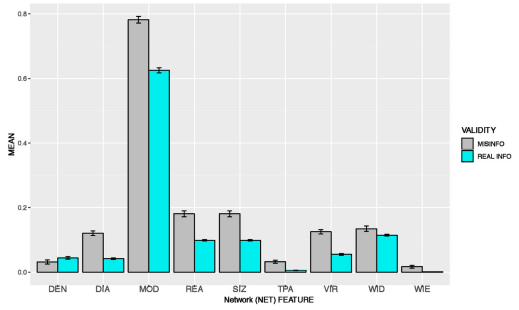

FIGURE 1. Network features comparison between real and misinformation groups.

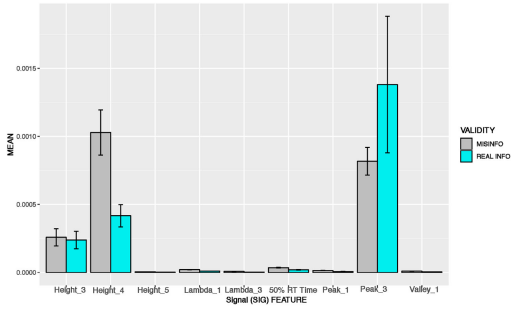

FIGURE 2. Top signal features comparison between real and misinformation groups.

for cross-validation. Classification metrics such as accuracy, F1 score, and AUC are calculated and averaged from this 10-fold cross-validation to evaluate RF and SVM classifiers' performance.

In this study, the dataset was not balanced, and there were more tweets with real information than misinformation. Therefore, F1-score is a more representative performance metric than accuracy. ROC curve and AUC (the area under ROC curve) also helped visually evaluate and compare model performance.

Data processing, mining, and exploratory analyses were carried in *R* 3.5.0 with various supporting packages. ML classifiers were developed in *Python* 3.7 with *SciKitLearn*. All annotated data and accompanying codes will be freely available on open data and code sharing repository (GitHub).

RESULTS

Feature Comparison Between Zika Real and Misinformation Tweets

First, we show how network features differ between Zika real and misinformation tweet groups in Figure 1. Note that these features are scaled between 0 and 1, so they do not represent actual values in Figure 1 through Figure 4. All nine network features, characterizing information dissemination network from global network to individual vertex level, differ significantly ($p < 0.05$) between the two groups according to the Kolmogrov–Smirnov test. Based on RF classification results in the later section, the most influential network feature, as measured by Gini impurity, is total path (TPA), followed by Wiener index (WIE) and virality (VIR). These results reveal information dissemination network structures differ substantially between real and misinformation groups.

For signal features, we identify the top 9 most influential features using Gini impurity out of a total of 40 features, making it consistent with network features. A comparison of these 9 important signal features is shown in Figure 2. Interestingly, the height of third peak, not the first two peaks, is the strongest signal to differentiate misinformation from real information. Our explanation is that most real information signals may not have more than two peaks, thus, having a third peak could be a sign of misinformation. The next important signal feature is time to receive 50% retweets, also known as the "half-life" of a tweet. The third important feature is the valley width of the first peak, i.e., time duration between first and second peaks. This feature quantifies the rate of information relay and can be used to distinguish misinformation. Kolmogrov–Smirnov tests further show that these nine most influential signal features' distributions also differed significantly ($p < 0.05$) between real and misinformation groups.

In addition to information dissemination feature categories, we characterized Zika real and misinformation content features on Twitter. First, we identified a list of top differentiating LIWC features (see Figure 3), which were significantly different between the two groups ($p < 0.05$). "Money" was the most important feature, and misinformation was related to the monetary aspect of Zika much more frequently than real information. The next was "Analytic," where real

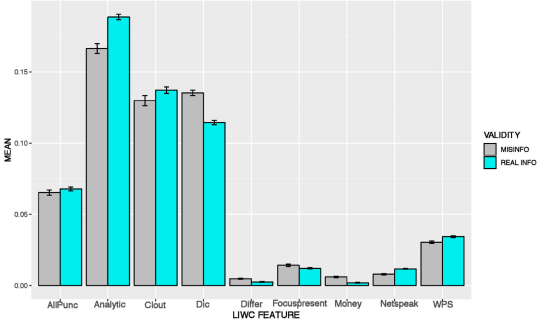

FIGURE 3. Top LIWC features comparison between real and misinformation groups.

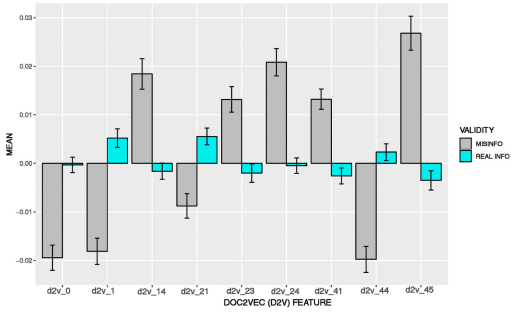

FIGURE 4. Top Doc2Vec features comparison between real and misinformation groups.

information had much higher likelihood to involve analytical aspect than misinformation. All these results showed content topic differences between Zika real and misinformation groups. It reinforced our suggestion that the health is not an isolated issue but is confounded with various social, political, and economic issues, which may lead to potential misinformation.

Next, we showed Doc2Vec feature differences between Zika real and misinformation groups (see Figure 4). According to the later RF model, feature 24 was the single most influential feature, accounting for more than 75% of decision tree split based on Gini impurity and dwarfing all other Doc2Vec features. However, unlike LIWC features, Doc2Vec features were learned by NN directly and did not have a clear interpretation.

ML Classification and Detection of Zika Misinformation

Tables 2 and 3 described classifier performance of RF and SVM, respectively. In general, a single feature group as input is less effective to detect Zika misinformation on Twitter. Among the five feature categories we have explored, content-based Doc2Vec (D2V) features show the best performance with the RF model, followed by user-related (USER), LIWC, and network (NET) features. Signal (SIG) features alone have the least power to differentiate misinformation from real ones. Using SVM, the same results still hold where Doc2Vec is the most differentiating group of features.

Combining different groups of features increase classifier's performance. For RF, combining NET and SIG as dissemination features slightly increase model accuracy, F1 score, and AUC (73%, 79%, 78%) from NET (71%, 77%, 77%) and SIG (70%, 77%, 75%) group alone. Interestingly, combining LIWC and D2V as content features does not increase model performance. The highest model performance is achieved when combining all five categories of features, with accuracy, F1 score, and AUC at 82%, 86%, and 90%, respectively (Table 2, last row). These results demonstrate that

TABLE 2. Performance with different feature categories: RF.

Feature Categories	Accuracy	F1-score	AUC	TPR	FPR
USER	0.770	0.823	0.839	0.87	0.39
LIWC	0.731	0.791	0.776	0.83	0.43
Doc2Vec (D2V)	0.771	0.830	0.821	0.91	0.45
NETWORK (NET)	0.711	0.772	0.765	0.8	0.44
SIGNAL (SIG)	0.697	0.769	0.75	0.82	0.51
LIWC+D2V (Content)	0.750	0.807	0.825	0.85	0.41
NET+SIG	0.733	0.787	0.782	0.81	0.44
USER+LIWC +D2V	0.796	0.841	0.877	0.88	0.34
All Feature Categories	**0.822**	**0.859**	**0.901**	0.89	0.23

health misinformation on social media is a multiaspect problem, and we need to characterize different aspects of health misinformation more comprehensively. Comparison of ROC curve and AUC values across combinations of features in RF is shown in Figure 5.

Comparing between the two classifiers, RF always outperforms SVM in this study. Combining all feature categories still yield the best performance in SVM (Table 3, last row). Nevertheless, the best-performed SVM classifier trails behind RF with accuracy, F1 score,

TABLE 3. Performance with different feature categories: SVM.

Feature Categories	Accuracy	F1-score	AUC	TPR	FPR
USER	0.679	0.765	0.792	0.85	0.59
LIWC	0.678	0.756	0.698	0.81	0.53
Doc2Vec (D2V)	0.753	0.810	0.807	0.87	0.43
NETWORK (NET)	0.737	0.808	0.774	0.90	0.52
SIGNAL (SIG)	0.669	0.761	0.645	0.86	0.68
LIWC+D2V (Content)	0.729	0.787	0.782	0.81	0.39
NET+SIG	0.707	0.787	0.782	0.81	0.52
USER+LIWC +D2V	0.768	0.814	0.843	0.83	0.32
All Feature Categories	**0.771**	**0.817**	**0.847**	0.83	0.30

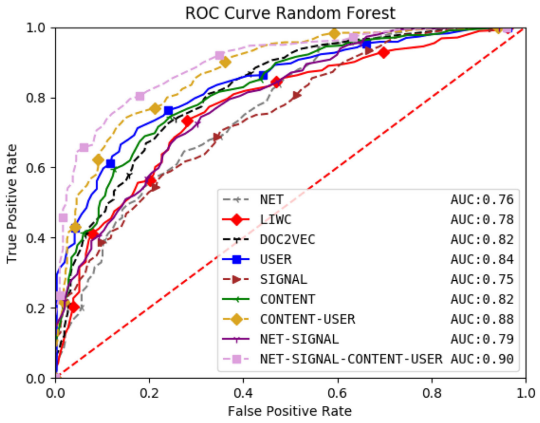

FIGURE 5. ROC and AUC of RF classifier with different feature categories.

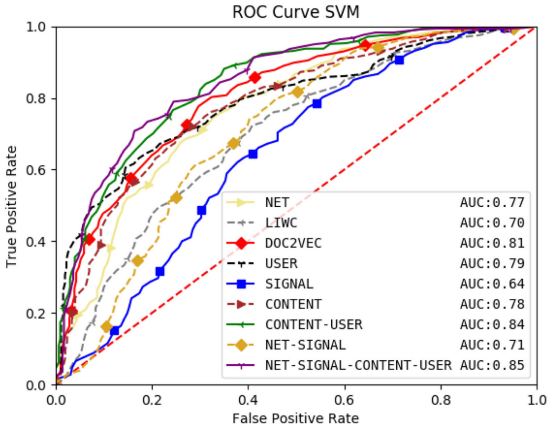

FIGURE 6. ROC and AUC of SVM classifier with different feature categories.

and AUC at 77%, 82%, and 95%, respectively, with a −5%, −4%, and −5% difference from the best RF model. Comparison of ROC curve and AUC values across combinations of features in SVM is shown in Figure 6.

DISCUSSION

This study delivers an accurate health misinformation classifier that takes multiple aspects of misinformation into consideration. More importantly, this study provides a more holistic view of the health misinformation challenge on social media by exploring different aspects of misinformation, not just the content itself. Health misinformation is like a pathogen in the real world, and no single discipline can tackle pandemic alone. Therefore, combining power from different aspects of health misinformation will enable better understanding and response to health misinformation challenge on social media.

Practically, this study utilizes data mining techniques to extract more features, especially information dissemination features. The retweeting network $G = \{V, E\}$ in this study is constructed at the end of information dissemination when the last retweet is received. Future work can be done to further construct and characterize temporal network $G_t = \{V, E, t\}$ at given time t and merge insights with signal-based features that explicitly tackle the temporal aspect of (mis)information dissemination.

Findings from this study not only deliver an accurate Zika misinformation classifier on social media, but also shed new insights on characterizing and understanding general misinformation, including health misinformation more comprehensively on social media. The new perspective and approach can be readily transferred to tackle other emerging issues such as the current COVID-19 pandemic, and to develop more generic classifiers.

However, we need to point out some limitations of our work. First, the feature categories are not an exhaustive list, and new features are yet to be discovered. Like pathogens in the real world, misinformation can also adapt to the changing environment, mimic behavior of real information, and become more difficult to detect. Second, this study emphasizes more on feature extraction, and future work will continue identifying the most effective set of input features across feature categories. We have applied RF to rank feature importance based on Gini impurity, and this will guide further fine-tuning of ML classifiers with fewer inputs. In addition, we will evaluate potential overfitting issue in delivering ML classifiers. Third, this classifier is developed in the context of 2016 Zika discussion, and its effectiveness needs to be re-evaluated in more recent health issues such as current COVID-19, where social media landscape has changed since 2016. We will apply transfer learning techniques to adopt this work in new health emergencies.

CONCLUSION

In this article, we transferred existing knowledge of real-world epidemics to comprehensively characterize Zika misinformation infiltration on social media in 2016. We developed a novel data mining technique to construct (mis)information dissemination networks and signals, extract dissemination features, and further combine content-based features based on LIWC and Doc2-Vec and user features. Based on various combinations of different feature categories, we developed an accurate Zika misinformation classifier using RF that can detect misinformation with $>85\%$ accuracy and $>90\%$ AUC. The novel perspective and analytical framework in this article can be transferred to respond to misinformation during current COVID-19 and future pandemics.

REFERENCES

1. S. R. Bedrosian *et al.*, "Lessons of risk communication and health promotion—West Africa and United States," *Morbidity Mortality Weekly Rep.*, vol. 65, no. 3, pp. 68–74, 2016. [Online]. Available: http://dx.doi.org/10.15585/mmwr.su6503a10external icon
2. A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Science vs conspiracy: Collective narratives in the age of misinformation," *PLoS One*, vol. 10, no. 2, 2015, Art. no. e0118093. [Online]. Available: https://doi.org/10.1371/journal.pone.0118093
3. A. Bovet and H. A Makse, "Influence of fake news in twitter during the 2016 US presidential election," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, 2019, doi: 10.1038/s41467-018-07761-2.
4. D. A Broniatowski *et al.*, "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate," *Amer. J. Public Health*, vol. 108, no. 10, pp. 1378–1384, 2018, doi: 10.2105/AJPH.2018.304567.
5. C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684, doi: 10.1145/1963405.1963500.
6. S. Chen *et al.*, "Dynamics of health agency response and public engagement during public health emergency: A case study of CDC tweeting pattern during 2016 Zika epidemic," *JMIR Public Health Surveillance*, 2018, vol. 4, Art. no. e10827, doi: 10.2196/10827.
7. M. Del Vicario *et al.*, "The spreading of misinformation online," *Proc. Nat. Acad. Sci.*, vol. 113, no. 3, pp. 554–559, 2016, doi: 10.1073/pnas.1517441113.
8. G. Eysenbach, "Infodemiology and infoveillance tracking online health information and cyberbehavior for public health," *Amer. J. Prev. Med.*, vol. 40, no. 5, pp. S154–S158, 2011, doi: 10.1016/j.amepre.2011.02.006.
9. A. E. Fard, M. Mohammadi, Y. Chen, and B. Van de Walle, "Computational rumor detection without non-rumor: A one-class classification approach," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 830–846, Aug. 2019, doi: 10.1109/TCSS.2019.2931186.
10. A. Friggeri, L. Adamic, D. Eckles, and J. Cheng, "Rumor cascades," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8122/8110
11. Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2972–2978.
12. S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining*, 2013, pp. 1103–1108, doi: 10.1109/ICDM.2013.61.
13. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196, doi: 10.5555/3044805.3045025.
14. V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1589–1599, doi: 10.5555/2145432.2145602.
15. L. Safarnejad, Q. Xu, Y. Ge, A. Bagavathi, S. Krishnan, and S. Chen, "Contrasting real and misinformation dissemination network structures on social media during the 2016 Zika epidemic," *Amer. J. Public Health*, vol. 110, no. S3, pp. S340–S347, 2020, doi: 10.2105/AJPH.2020.305854.
16. L. Safarnejad, Q. Xu, Y. Ge, A. Bagavathi, S. Krishnan, and S. Chen, "Identifying influential factors on discussion dynamics of emerging health issues on social media: A computational study," *JMIR Public Health Surveillance*, vol. 6, no. 3, 2020, Art. no. e17175, doi: 10.2196/17175.
17. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.
18. Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, 2010, doi: 10.1177/0261927X09351676.
19. A. Tong, D.-Z. Du, and W. Wu, "On misinformation containment in online social networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 341–351. [Online]. Available: https://papers.nips.cc/paper/2018/file/9b04d152845ec0a378394003c96da594-Paper.pdf
20. H. M. Webb and M. Jirotka, "Nuance, societal dynamics, and responsibility in addressing misinformation in the post-truth era: Commentary on lewandowsky, ecker, and cook," *J. Appl. Res. Memory Cogn.*, vol. 6, no. 4, pp. 414–417, 2017, doi: 10.1016/j.jarmac.2017.10.001.
21. F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proc. ACM SIGKDD Workshop Mining Data Semantics*, 2012, pp. 1–7, doi: 10.1145/2350190.2350203.

LIDA SAFARNEJAD is with Stanford University, Stanford, CA, USA. Contact her at lsafarne@uncc.edu.

QIAN XU is with Elon University, Elon, NC, USA. Contact her at qxu@elon.edu.

YAORONG GE is with the University of North Carolina at Charlotte, Charlotte, NC, USA. Contact him at yge@uncc.edu.

SHI CHEN is with the University of North Carolina at Charlotte, Charlotte, NC, USA. He is the corresponding author of this article. Contact him at schen56@uncc.edu.