Proceedings of the 2019 IISE Annual Conference
H.E. Romeijn, A. Schaefer, R. Thomas, eds.

IISE 2019 Conference & Expo
A Multi-Algorithm Approach for Classifying Misinformed Twitter Data during Crisis Events

Kyle Hunt, Puneet Agarwal, Jun Zhuang
Department of Industrial and Systems Engineering, University at Buffalo
Buffalo, New York

Abstract

Social media is being increasingly utilized to spread breaking news and updates during disasters of all magnitudes. Unfortunately, due to the unmoderated nature of social media platforms such as Twitter, rumors and misinformation are able to propagate widely. Given this, a surfeit of research has studied rumor diffusion on social media, especially during natural disasters. In many studies, researchers manually code social media data to further analyze the patterns and diffusion dynamics of users and misinformation. This method requires many human hours, and is prone to significant incorrect classifications if the work is not checked over by another individual. In our studies, we fill the research gap by applying seven different machine learning algorithms to automatically classify misinformed Twitter data that is spread during disaster events. Due to the unbalanced nature of the data, three different balancing algorithms are also applied and compared. We collect and drive the classifiers with data from the Manchester Arena bombing (2017), Hurricane Harvey (2017), the Hawaiian incoming missile alert (2018), and the East Coast US tsunami alert (2018). Over 20,000 tweets are classified based on the veracity of their content as either true, false, or neutral, with overall accuracies exceeding 89%.

Keywords
machine learning; Twitter; misinformation; social media; disaster

1. Introduction
During crisis events around the world, such as hurricanes and bombings, people seek information regarding evacuation planning, event updates, and general safety. As social media proves to be an extremely efficient method to offer and transfer news, it is being relied upon by individuals for breaking news stories and updates [1]. It proves to perform better than the traditional mass media given its timely updates of events and interactive two-way communications [2]. Given this, many research works have studied crisis communications through social media. Due to the unmoderated nature of Twitter and other social media platforms, rumors, misinformation and fake news are free to spread [3]. Rumors, which we define as “an item of circulating information whose veracity status is yet to be verified at the time of posting" [4], can be extremely dangerous during crisis events, as Twitter users are more likely to spread false information than the truth [5]. Classifying rumored or misinformed Twitter data for research usually comes in the form of extensive manual coding, needing many human hours to read through all of the tweets and classify the content within them. In this paper, we study misinformation classification during the Manchester Arena bombing (2017), Hurricane Harvey (2017), the Hawaiian incoming missile alert (2018), and the the East Coast United States tsunami alert (2018). Seven different supervised machine learning (ML) algorithms are applied to classify the content of misinformed tweets as either true, false, or neutral (e.g., opinions, comments, questions), and three data sampling algorithms are applied to balance the data before training ML. The results of this work will be beneficial to researchers in their work surrounding misinformation diffusion and social media, and also government and emergency response agencies in their attempts to track and debunk rumors and misinformation. We fill the research gap by applying classification and balancing algorithms to misinformed Twitter data which propagates during disaster events.

1.1 Rumors and Misinformation Spread during Crisis Events
Due to the the continued use of social media for crisis communications, many researchers have studied the propagation of rumors and misinformation on Twitter. After the 2011 Great East Japan Earthquake, a false rumor was spread which

stated that rain in the earthquake's aftermath may include harmful chemicals, and there was a warning to use umbrellas outside. Researchers located the rumor correction tweet which ended the diffusion of the false information, and were able to create a model to estimate the rumor infection rate and the number of people who still believed in the rumor after the correction tweet was posted [6]. In Hurricane Sandy and the Boston Marathon bombing (2012 and 2013, respectively), Twitter users performed poorly in detecting false rumors, and rushed to spread the rumored news [7]. These results lined up with [8], where the authors studied three false rumors during the 2013 Boston Bombing, and found that Twitter users did not do so well in distinguishing truth from hoax. Results of this study also found that although rumor correction tweets were emerging, the propagation of the misinformation seemed to overpower the correction efforts. Following the 2010 earthquake in Chile, tweets related to news were compared to tweets related to rumors, and the results showed that the propagation patterns were much different. The authors concluded that this was mainly because rumors are questioned much more than news [9]. Many research works which study rumors and misinformation propagation utilize content analysis and manual coding schemes to classify and analyze their datasets [7, 10], and such research has continued to grow in recent years. As many researchers have began utilizing supervised learning techniques to classify Twitter data, it becomes clear that more work is needed in order to aid in detecting and tracking misinformation during crisis events.

1.2 Social Media Data Classification

Over the last decade, many works have applied supervised machine learning algorithms to classify Twitter data. In [11], the authors utilize sequential and non-sequential classifiers to conduct stance classification on tweets within rumor threads, and their results indicate low F-1 scores and poor classification performance. In [12], the authors manually code a collection of 4,300 tweets from a diverse set of five rumors, and then utilize supervised learning to classify the tweets as either rumor affirming, rumor denying, or neutral. The features used to train ML were linguistic, sentimental, and Twitter elements, and the random forest classifier performed the best on all of the rumor cases, with accuracy and F-1 scores exceeding 90% in some cases. Research from [13] investigates the determinant features of online rumor spreading from temporal, structural, and linguistic aspects, and tests the determinant features with three different classifiers (random forest, SVM, and decision tree). The algorithms were trained to label topics as either rumor or non-rumor, and the resulting indicators show great classification performance (maximum accuracy of 90% with random forest). To the best of our knowledge, no previous work has applied supervised ML techniques to classify the veracity of misinformed tweets, or addressed the unbalanced nature of Twitter data before classifiers are utilized. In this work, we fill the research gap by applying and comparing k-nearest neighbors (k-NN), support vector machine (SVM), decision tree, random forest, AdaBoost, multinomial naive bayes (MNB), and XGBoost to classify the content of misinformed tweets as either true, false, or neutral. Due to the fact that the classes are very unbalanced before training, we also apply synthetic minority over-sampling technique (SMOTE), adaptive synthetic sampling (ADASYN), and random oversampling (ROS) to balance the data. Given this, our results also show the best combinations of classifiers and sampling algorithms for automatically classifying social media data.

2. Research Methodology

To acquire a diverse dataset for ML, we study four different cases where misinformation was spread, including two false rumors and two false alerts. The Holiday Inn rumor, which was spread following the 2017 Manchester bombing, was identified through major news outlets, such as The New York Times and CBS. For the Hurricane Harvey immigration rumor, we identified the case on FEMA's rumor control page [14]. The 2018 Hawaiian incoming missile false alert, and the 2018 East Coast tsunami false alert, were both identified in major news broadcasts. The news coverage from both of these events was international, and was broadcast online, on the radio, on television, and alerts were sent to citizens phone around the United States. Throughout our work with these four cases, we utilized Python programming language for all machine learning and sampling algorithms, and R programming language for all data handling and data processing.

2.1 Disasters and Misinformation

Manchester Bombing Holiday Inn Rumor On May 22nd, 2017, Ariana Grande was performing in the Manchester Arena in England. When the concert concluded and attendees were beginning to leave the venue, a suicide bomber detonated the explosives attached to his body. The bombing led to 23 deaths and over 139 wounded people, and was the deadliest terrorist attack in England since the 2005 London bombings. After the bombing, a rumor was spread on Twitter and Facebook stating that unaccompanied children were being taken to safety at the local Holiday Inn. A

Holiday Inn representative had to make a statement saying that this rumor was false, and there were no unaccompanied children at the hotel.

Hurricane Harvey Immigration Rumor On August 25th, 2017, Hurricane Harvey made landfall in Texas. During Hurricane Harvey, there was legislation due to be passed in Texas that was aiming to increase anti-immigration policies. As some people began to inquire about identification checks at evacuation shelters, a false rumor began to proliferate throughout social media and Texas that stated shelters were going to be checking IDs. This rumor proved to be very dangerous, as many undocumented immigrants were afraid to go to shelters due to their lack of citizenship and the potential threat of deportation. Many government and news accounts, including Houston's official Twitter account, posted rumor debunking tweets in order to help contain the spread of the misinformation.

Hawaii Missile False Alert On January 13th, 2018, Hawaii's Emergency Management Agency sent out a notification to cell phones, televisions, and radio stations stating that a ballistic missile was inbound. The emergency notification sent to the cell phones across Hawaii read "BALLISTIC MISSILE THREAT INBOUND TO HAWAII. SEEK IMMEDIATE SHELTER. THIS IS NOT A DRILL." A second alert, sent 38 minutes later, notified the public that this was a false alert, and there was no incoming missile. The false alert was blamed on human error, and Governor David Ige and other government authorities posted tweets to clarify the misinformation.

East Coast Tsunami False Alert On February 6th, 2018, weather alerts were sent to cell phones of citizens along the East Coast of the United States which stated that there was a tsunami warning in effect. Soon after the notifications were sent out, many agencies, such as the National Weather Service, posted to Twitter stating that there was no tsunami warning, and this information was not true. A later release stated that the weather advisory, which was sent to millions of cell phones, was intended to be a test of the emergency systems.

2.2 Data Collection

Twitter's REST Application Programming Interface (https://dev.twitter.com/rest/public) and Python programming were used for collecting all of the tweets across the four cases studied. The standard Twitter Search API does not return a complete set of tweets; therefore our datasets contains a sample of the tweets related to the misinformation cases based on the search criteria we used. In an attempt to counter this problem and retrieve more complete datasets, data collection took place over a 28 day window for every event, with collection done every three days using the same respective search criteria for every collection. Using this method, we are able to collect twice for every day in the 28 day window, excluding the last three days. Although this method still does not supply every related tweet and is computationally expensive, it gives us a less limited dataset. The different search dates for the four cases, along with the total tweets collected, can be found in Table 1. In total between the four cases, we collected 20,061 misinformation related tweets (after removing unrelated tweets from the data, explained in Section 2.3). The search criteria used for all cases were a combination of case insensitive keywords and hashtags that related to the specific case (e.g., immigration and #harvey; hawaii and #missile). The search criteria were chosen via an extensive Twitter Advanced Search to find the major keywords and hashtags that identified tweets related to the false rumors and false alerts. The criteria were searched in English.

Dataset	Collection Began	Collection Ended	Total Tweets
Manchester Bombing Holiday Inn Rumor	5/25/17	6/21/17	3,882
Hurricane Harvey Immigration Rumor	8/28/17	9/24/17	2,034
Hawaii Missile False Alert	1/14/18	2/10/18	6,691
East Coast Tsunami False Alert	2/7/18	3/6/18	7,454

Table 1: Collection dates and total tweets collected for all four cases

2.3 Coding Scheme

Utilizing latent content analysis, the text of each tweet was coded to categorize the information within it. Two students participated in the coding process for all of the tweets. The coders were required to become familiar with all four of the misinformation cases in this study before coding began. The tweets were coded into the following four mutually exclusive categories: true, false, neutral, and non-related. A rubric for the coding scheme can be found in Table 2. After both coders completed the datasets independently, they then worked together to cross-validate all of the tweets in which they disagreed on the category. There ended up being 4,216 of these instances (79% accuracy between coders;

Category	Definition
True	Contains valid content regarding the misinformation, delivering correct information to Twitter's network
False	Contains false content regarding the misinformation, delivering incorrect information to Twitter's network
Neutral	Comments, questions, or shares an opinion on the misinformation, and does not offer valid or false information
Non-related	Not related to the misinformation in any way; these tweets are removed from the dataset before training ML

Table 2: Coding rubric used to define the different categories

15,845 out of 20,061 coded the same), and the coders worked together to come to agreement on which category was more prominent in the content of these tweets. After cross-validation was completed, all non-related tweets were discarded, leaving us with the final numbers reported in Section 2.2. The final results from coding can be found in Table 3, and the unbalanced nature of the data becomes clear within the independent datasets, and also across the combined datasets. We use the combined data to train and test the classifiers, in order to have a large dataset which consists of different events and textual features.

Dataset	True	False	Neutral
Manchester Bombing Holiday Inn Rumor	299	3,498	85
Hurricane Harvey Immigration Rumor	1,474	99	461
Hawaii Missile False Alert	4,269	471	1,951
East Coast Tsunami False Alert	5,477	402	1,575
Total	**11,519**	**4,470**	**4,072**

Table 3: Results from data coding

2.4 Supervised Learning

In this work, we implement and compare the following seven classifiers: k-NN, SVM, decision tree, random forest, AdaBoost, MNB, and XGBoost. Due to the unbalanced nature of the data, we also implement and compare the following three sampling algorithms: SMOTE, ADASYN, and ROS. For all seven classifiers, we apply the three sampling algorithms, giving us 21 different combinations to compare. First, we pre-process the data by converting all characters to lower case, converting all URLs to strings, converting the "@" symbol to "at_user", and converting multiple spaces to a single space. Next, we run the sampling algorithms before training the classifiers, and then conduct feature extraction on the text of the tweets. For feature extraction, we use term frequency-inverse document frequency (TF-IDF). TF-IDF is the product of term frequency (how frequently a term occurs in a document) and inverse document frequency (how rare or unique a word is), and is a very popular statistical information retrieval method. The next step is to train the classifiers using 70% of the balanced data, and subsequently test the classifiers using 30% of the unbalanced data. Lastly, we review and compare the performance of all of the balancing and classification algorithms by retrieving the overall accuracy, precision, recall, and F-1 scores for all 21 combinations.

3. Results

The results from all classifier-sampling algorithm combinations are presented in Table 4. It is clear that all of the classifiers performed very well, with the lowest overall accuracy (OA) coming from k-NN combined with ADASYN, at 69.8%. The highest OA was achieved via SVM combined with ROS, reaching 89.52%. This combination had the best overall performance in three of the four metrics, with the OA being the highest, precision (86.66%) being the second highest, and recall and F-1 being the highest (87.95% and 87.12%, respectively). **For k-NN** the best results in classification came when using the ROS balancing algorithm, where we achieved a 79.93% OA, 75.15% precision, 78.82% recall, and 76.54% F-1 score. The worst results using the k-NN classifier came when balancing the data with ADASYN. **For SVM** the best results in classification came when using the ROS balancing algorithm, as reported above. The worst results came when the data was balanced with ADASYN. **For decision tree** the best results in classification came when balancing the data with SMOTE, where we achieved an 86.76% OA, 82.91% precision, 83.13% recall, and 83.02% F-1 score. The worst results came when balancing the data with ADASYN. **For random forest** the best results in classification came when balancing the data with ROS. We achieved an OA of 88.81%, precision of 85.86%, recall of 87.03%, and F-1 score of 86.3%. The worst results for random forest were obtained when balancing

the data with ADASYN. **For AdaBoost** the best results in classification came when balancing the data with SMOTE, where we achieved an 84.22% OA, 82.02% precision, 83.44% recall, an 81.78% F-1 score. The worst results came when balancing the data with ADASYN. **For MNB** the best results in classification came when using the SMOTE balancing algorithm, where we achieved an 86.03% OA, 83.54% precision, 84.69% recall, and 83.76% F-1 score. The worst results came when using the ADASYN balancing algorithm. Lastly, **for XGBoost** the best result came when balancing the data with SMOTE, where we achieved an OA of 85.61%, precision of 83.8%, recall of 85.45%, and F-1 score of 83.44%. The worst results for XGBoost came when balancing the data with ADASYN.

Given these results, for classifying misinformed Twitter data, we rank the ML algorithms from best to worst as follows: (1) SVM, (2) random forest, (3) decision tree, (4) MNB, (5) XGBoost, (6) AdaBoost, and (7) k-NN. These rankings are based on the highest OA achieved for every classifier. The sampling algorithms are ranked as follows: (1) SMOTE, (2) ROS, and (3) ADASYN. These rankings are based on the highest OAs achieved across all of the classifiers (SMOTE had the highest OA for four out of seven classifiers).

Classifier	Method	OA	Precision	Recall	F1
k-NN	SMOTE	72.11	72.02	77.45	71.38
	ADASYN	69.8	68.55	74.16	68.12
	ROS	79.93	75.15	78.82	76.54
SVM	SMOTE	89.47	86.71	87.48	86.97
	ADASYN	89.38	86.27	87.45	86.79
	ROS	89.52	86.66	87.95	87.12
Decision Tree	SMOTE	86.76	82.91	83.13	83.02
	ADASYN	85.82	81.76	82.45	82.07
	ROS	86.59	82.6	83.07	82.82
Random Forest	SMOTE	88.56	85.85	86.57	86.06
	ADASYN	88.47	85.44	86.58	85.91
	ROS	88.81	85.86	87.03	86.3
AdaBoost	SMOTE	84.22	82.02	83.44	81.78
	ADASYN	82.54	78.21	80.95	79.16
	ROS	83.55	81.59	83.01	81.21
MNB	SMOTE	86.03	83.54	84.69	83.76
	ADASYN	85.51	81.84	84.76	83.07
	ROS	85.78	83.33	84.2	83.46
XGBoost	SMOTE	85.61	83.8	85.45	83.44
	ADASYN	84.15	80.23	82.79	81.15
	ROS	85.2	83.63	85.43	83.16

Table 4: Classification results showing: overall accuracy (OA), precision, recall, and F-1 score (F1)

4. Conclusions and Future Research Directions

In this work, we were able to successfully classify misinformed Twitter data with supervised learning approaches, achieving considerable performance metrics. Training the ML algorithms with manually coded data, we automatically classified the data into tweets that were true, false, or neutral, with the highest overall accuracy reaching 89.52% via the SVM classifier. Utilizing seven different ML algorithms and three different sampling algorithms, we offer scholars, researchers, and emergency managers with insights and methods to help detect and track misinformation on social media. For example, government and emergency response organizations can apply these findings to track rumors in real time that are spreading on social media, to help them make informed decisions regarding intervention (such as posting misinformation correction/debunking tweets). By knowing how many users are tweeting true or false information regarding an event, timely actions can be taken to contain the spread of incorrect information. Addition-

ally, researchers worldwide can use these techniques and algorithm combinations to automatically classify their social media data into different categories after training the algorithms, saving hours of labor that would be spent manually coding the data to further analyze. Given the extraordinary losses associated with natural and man-made disasters, it is critical that workers in the research sector, industry sector, and government sector can spend their time efficiently in order to help mitigate future loss.

In future research, this work can be extended by applying neural networks in the same domain in order to increase the performance and accuracy levels of the classifiers. Similarly, the techniques can be applied to classify the Twitter users into different behavioral groups (e.g., users spreading false rumors, debunking false rumors, questioning the rumors, commenting, etc.) in order to study features of rumor diffusion and network structure.

Acknowledgements

This research was supported by the National Science Foundation (NSF) under award numbers 1760586 and 1762807. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also thank the referees for providing constructive comments.

References

[1] Phuvipadawat, S., and Murata, T. (2010). Breaking news detection and tracking in Twitter. In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 120-123). IEEE.

[2] Fraustino, J. D., Liu, B., and Jin, Y. (2012). Social media use during disasters: A review of the knowledge base and gaps (final report, start). Technical Report. Human Factors/Behavioral Sciences Division, Science and Technology Directorate, US Department of Homeland Security.

[3] Procter, R., Vis, F., and Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. International journal of social research methodology, 16(3), 197-214.

[4] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR), 51(2), 32.

[5] Vosoughi, S., Roy, D., Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

[6] Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., and Takayasu, H. (2015). Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study. PLoS one, 10(4), e0121443.

[7] Wang, B., and Zhuang, J. (2018). Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters. Natural Hazards, 93(3), 1145-1162.

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

[9] Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT?. In Proceedings of the first workshop on social media analytics (pp. 71-79). ACM.

[10] Oh, O., Agrawal, M., Rao, and H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. Mis Quarterly, 37(2).

[11] Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., and Lukasik, M. (2016). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In Proceedings of the 26th International Conference on Computational Linguistics. Association for Computational Linguistics, 2438-2448.

[12] Zeng, L., Starbird, K., and Spiro, E. S. (2016). #Unconfirmed: Classifying rumor stance in crisis-related social media messages. In Tenth International AAAI Conference on Web and Social Media.

[13] Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y. (2013, December). Prominent features of rumor propagation in online social media. In 2013 IEEE 13th International Conference on Data Mining (pp. 1103-1108). IEEE.

[14] Federal Emergency Management Agency. Harvey rumor control. online, Aug 29 2017. https://www.fema.gov/disaster/4332/updates/rumor-control.