UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

COMPUTER SCIENCE HONOURS
FINAL PAPER
2017

Title: Developing a Machine-Learning Classifier for Assessing the Credibility of Information on Twitter

Author: Michelle Lu

Project Abbreviation: SASITWIT

Supervisor(s): Selvas Mwanza, Associate Professor Hussein Suleman

Category	Min	Max	Chosen
Requirement Analysis and Design	0	20	0
Theoretical Analysis	0	25	0
Experiment Design and Execution	0	20	15
System Development and Implementation	0	15	10
Results, Findings and Conclusion	10	20	20
Aim Formulation and Background Work	10	15	15
Quality of Paper Writing and Presentation	10		10
Quality of Deliverables	10		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	0	10	
Total marks		**80**	

Developing a Machine-Learning Classifier for Assessing the Credibility of Information on Twitter

Michelle Lu
University of Cape Town
lxxwei013@myuct.ac.za

ABSTRACT

This paper presents the development of a machine-learning classifier that can rate the credibility of tweets from Twitter, a microblogging platform. Due to the rise of fake news on social media platforms and the ease of information dissemination, specifically on Twitter, it has become important to distinguish rumour from fact. A variety of features were used for classification, which can be divided into three areas: graph-based, tweet-based and user-based features. In particular, the interest lied in whether the network structure of the diffusion of information has any significant impact on the credibility, therefore graphs were constructed to show the relationships of the retweets and replies of a tweet. Through statistical analysis of all the features, it was found that network features do play a significant role in determining the credibility. A variety of different machine-learning algorithms were evaluated through cross-validation and pairwise t-tests, with the results indicating that the J48 Decision Tree provided the best performance with regards to precision, recall, F-measure and could classify tweets with an accuracy of 70.02%.

CCS CONCEPTS

• **Computing methodologies → Machine learning algorithms**;

KEYWORDS

Twitter, Information credibility, Fake news, Machine-learning

1 INTRODUCTION

In modern day society, access to social media has become widely available, and has evolved into a medium for users to share information about events occurring around them. Social media platforms, such as Twitter, have moved beyond being just a social network, it now acts as source of news and a medium for activists. Twitter is a Web service that allows users to post short messages , known as tweets, which have become popular because information can be received much faster than conventional media [16]. Tweets are automatically uploaded and available to the public and this exposure means that anyone who has access to Internet can see this information even if they do not know the user [26]. The freedom for users to publicize anything without much control has given rise to spammers, those who spread rumours and fake news, which can have negative impacts. This leads to the question of whether one can really trust the credibility of what they see on social media.

1.1 Project Significance

The impact of fake news has been a concern, especially during the United States 2016 election, where it was suggested that the influence of fake news resulted in the election of Donald Trump as president [5]. This statement was due to a combination of factors, which included a report stating 62% of adults in the US receive their news through social media and the most popular, widely shared news stories were mostly fake, which tended to favour Donald Trump over Hilary Clinton [5].

Twitter has played important roles in reporting news, such as the 2008 Mumbai bomb blasts and the 2009 US Airways plane crash [26]. This leads to the argument that Twitter allows ordinary citizens to generate and consume news created by other ordinary citizens. An example of this is evident from the 2009 US Airline jet plane crash, where a citizen was able to report the news before the media crew could even arrive, which instantly transformed him into an on-scene journalist [26]. Unlike traditional media, where news will be filtered by journalists and editors, social media updates are not restricted in their content [14], and research has shown that people are poor at detecting disinformation in text content alone [25], thus research on how information credibility can be automatically assessed may mitigate the problem of misinformation.

1.2 Project Aim

This project is a subset of a larger one that involves developing an automated credibility system for tweets on Twitter. In this project, various machine-learning algorithms will be trained with a selection of features to classify the credibility of a tweet. The credibility of a tweet can be defined as the level of believability associated with the information contained in the tweet[40]. We would like to observe which features play the most important role in credibility classification, as well as which machine-learning algorithm will give the best performance. The research question we would ultimately like to answer is: Can distinctive features be extracted from fake and legitimate information on Twitter to train a machine-learning classifier that can assess tweets with high precision and recall?

1.3 Structure of Report

This paper is organized as follows. Section 2 is a literature review of previous studies related to this project. Section 3 describes the experiment design, such as which features were chosen to be analyzed,and which machine-learning classifiers were used. Section 4 describes the methodology for data collection and how the relevant data was extracted and classified. Section 5 will report the findings from the statistical analysis of the features and evaluation of machine-learning classifiers. Section 6 will discuss the ethical and legal issues and lastly, section 7 will present the conclusion and discussion of future work.

2 RELATED WORK

There are numerous studies on credibility assessment related to Twitter. In this section, we highlight approaches previous studies have used that are relevant to this project.

2.1 Network Graphs

Network graphs can be used to graphically show the interconnections between entities. Chat et al. [12] used directed edges of a graph to demonstrate the movement of information between users. A similar study was done by Wang [39] who used a directed social graph that indicated the follower and friend relationships. It is directed because the 'follow' relationship is not always mutual. Wang [39] used a Bayesian classifier to differentiate suspicious and normal activity, and achieved an accuracy of 89%. The classifier showed that the majority of spammers have few followers [39], which corresponds to Twitter's spam policy stating that, if an account has few followers in comparison to the number of users they follow, this could mean that they are a spam account. Castillo et al. [11] concludes that users who have a low number of friends and little tweet history, tend to be related to low credibility information.

According to Qazvinian et al. [33], network-based features, such as the number of retweets a user has or their tweet history, can be a good measure of rumour detection. The Web service 'Truthy', developed by Ratkiewicz et al. [34] analyzes the circulation of information in Twitter through the use of directed graphs, with the edges representing retweets between users. Castillo et al. [11] took a similar approach with the use of a propagation tree to observe retweeting behaviour. They found that the depth of the propagation tree is an important feature, as more retweets are linked with more credible news [11]. There, however, needs to be caution when modelling retweets. It is important to have a clear separation between the original user who posted a tweet and the user who retweeted it, because sometimes people change the retweeted post, which can result in a different meaning [33]. By using a J48 Decision tree, it was found that the ratio of the number of followers to friends and the number of retweets were very good features for identifying fake news [33].

However, research about the propagation of fake URLs during Hurricane Sandy by Gupta et al.[17], showed that during crisis situations, people tended to retweet whatever they find on the topic, irrespective of whether they follow the poster or not. This was shown when they observed the follower relationships via network graphs and found that this contributed to only 11% of the spread of fake URLs [17]. This could possibly mean that the number of followers a user has is not accurate enough on its own in distinguishing fake news.

2.2 Tweet-based Features

The most common feature used to identify phishing is URL links in tweets. Nowadays, phishers use URL shorteners to hide the identity of the domain, which makes identifying suspicious URLS much more difficult [13]. Many studies[17, 28], show that URLs are still an effective feature for credibility detection, as spam and phishing tweets tend to include a higher number of URL links.

Some other distinguishing features in detecting spammers result from their behaviour patterns. Wang [39] found that spammers tend to post numerous duplicates. He used the method of Levenshtein distance to calculate whether tweets were duplicates or not, however, the results also showed that this alone cannot be used to detect spammers, because some legitimate users also demonstrate this behaviour [39]. In conjunction to duplication, spammers try to gain more visibility by using hashtags to take advantage of trending topics, and use @username mentions to get more retweets [13]. They will usually post many tweets in set intervals of time [3]. Overall, it appears that the best indicators of suspicious behaviour from spammers include a high number of mentions and hashtags for tweets posted in set intervals of time.

2.3 Sentiment Analysis

The content of a tweet can be classified into positive and negative sentiments via sentiment analysis. Castillo et al.[11] used a J48 Decision Tree and found that tweets with negative sentiments tended to relate to credible news. Kang et al.[21] built on top of Castillo et al. 's[11] work to develop their content-based model. They tried to identify term patterns and other properties of tweets by using a probabilistic-language approach, which would try to find if these patterns could lead to more retweets, or higher user credibility ratings. However, this method only achieved a 62% accuracy for this content-based model [21]. Another method by Ikegami et al.[20] used Latent Dirichlet Allocation to locate tweets on specific topics, and classified whether the opinion on the topic was positive or negative by using a semantic orientation dictionary. The credibility of the information was then based on the ratio of positive to negative opinions. The accuracy of both the topic and opinion classification was only 47.6% [20].

2.4 Machine-Learning Classification

There are various schemes proposed for spam detection or credibility assessment on Twitter that use machine learning methods as a means of classification. These methods include Naïve Bayesian classifiers [3, 17, 38, 39], Decision Tree classifiers [3, 11, 17, 38, 39] and Random Forest classifiers [3]. Wang's [39] Bayesian classifier was used to differentiate suspicious and normal activity, and achieved an accuracy of 89%. Gupta et al. [17] used a J48 Decision Tree and achieved a 97% accuracy in determining credibility of tweets containing URLs to images associated with Hurricane Sandy. Aggarwal et al. [3] achieved 92.52% accuracy in detecting phishing tweets.

2.5 Existing Credibility Systems

There are systems that have been developed to assess the credibility of information. There is PhishAri, which is a real-time detector of phishing in Twitter [3] and TweetCred, which is a Web browser extension that gives credibility ratings to tweets on Twitter [16]. These two systems use different algorithms to classify information. PhishAri adopted machine learning classification techniques that include Naive Bayes, Decision Trees and Random Forest. The Random Forest Classifier algorithm proved to be the most accurate with a precision of 92.52% in detecting phishing tweets [3]. They found that the most revealing features include how long the account has been active, the number of trending hashtags and retweets, follower-to-friends ratio and @username mentions [3]. TweetCred, on the other hand, used a combination of 45 features extracted from

Twitter and a semi-supervised ranking algorithm for credibility assessment [16]. Their results from their user evaluation, however, showed that only 43% of the users agreed with the rating. Majority of the users gave higher ranking scores than the ones given by TweetCred, which could be due to the ranking algorithm not taking into consideration the relationships between users on Twitter [16].

3 EXPERIMENT DESIGN

This section discusses the features selected for classification and how analysis was performed to determine which are the most significant. It will also discuss which machine-learning classifiers were used and how their performance was measured.

3.1 Feature Extraction

The features of interest were based off those mentioned in Section 2, with additional interest in the network propagation features, through retweets and replies. A retweet is when a user republishes a post that another user has uploaded and a reply is a means through which users can interact with each other in the comment section of every tweet [1]. A network can be represented as a graph with Twitter users as nodes, and the edges representing the type of relationship between the users, such as a retweet or reply. All the selected features can be divided into three categories: Tweet-based features, User-based features and Graph-based features. A summary of these features is provided in the Tables 1, 2 and 3.

Table 1: Summary of the Tweet-Based Features

Feature	Description
Sentiment Score	The sentiment of the text portion of the tweet.
Word Count	Number of words in a tweet
Character Count	Number of characters in a tweet
Retweet Count	Number of reposts of the tweet [1]
Reply Count	Number of responses to the tweet [1]
Like Count	Number of likes
User Mentions Count	Number of other usernames contained in the tweet [1] e.g. @user1
Hashtag Count	Number of indexed keywords or topics [1]. e.g. #SA
Contains URL	The tweet contains an URL
Is Retweet	The tweet is a retweet

Table 2: Summary of User-Based Features

Feature	Description
Contains Profile Image	User account has a profile image
Is Verified	User account is authenticated[1]
Followers Count	Number of people who subscribe to a user [1]
Friends Count	Number of people a user has subscribed to[1]
Friend-Follower Ratio	Followers count divided by the friends count
Contains Description URL	User Profile Description contains URL

Table 3: Summary of Graph-Based Features

Feature	Description
Node Count	Number of connected entities in the graph
Edge Count	Number of connections in the graph
Community Count	Number of vertices that are completely connected to each other forming a group [24]
Clustering Coefficient	The extent to which the neighbours of a given node connect to each other [19].
Connected Components Count	Number of vertices that are reachable from one another[24].
Density	Number of edges divided by the maximum number of possible edges [19].
Diameter	The maximum distance between any two nodes[19].
Path Length	The longest shortest distance between two nodes [19].
Average Eccentricity	The maximum distance from a node to all other nodes [19]. (Averaged)
Average Closeness Centrality	The time it takes for information from one node to reach all other nodes [4]. (Averaged)
Average Betweeness Centrality	Frequency of a node found on a shortest path between two nodes [4]. (Averaged)
Average Eigencentrality	Measurement of the influence of a node in a network [4]. (Averaged)
Average Degree	Average number of connections each node has [4].
Average Indegree	Average number of incoming connections for each node [4].
Average Outdegree	Average number of outgoing connections for each node [4].
Average Pageranks	Measurement of the importance of each node in a network [24]. (Averaged)
Average Authority	Score of the value of the information at each node [24]. (Averaged)
Average Hub	Quality of the outgoing links from a node [24].(Averaged)

3.2 Feature Analysis

To determine which features were the most significant, statistical analysis was performed on the data through RStudio [1]. Since the response variable, in this case the credibility label, has more than two classes, multinomial logistic regression was be used. The two-tailed z test was then performed to calculate the corresponding p-values of each feature, and from these values, it was determined which features were significant to the model. The relative risk was also calculated for each feature in the model, in order to compare the likelihood of an observed outcome, based on the feature.

3.3 Machine-Learning Algorithms

The Waikato Environment for Knowledge Analysis (Weka) Java library [18] provides an assortment of machine-learning algorithms that can be used for classification. The working environment it provides allows those who aren't machine-learning experts to have access to these tools. It is freely available and can run on any computing platform. In order to evaluate which of the machine-learning algorithms will have the best performance, important algorithms for classification were selected based off the Weka documentation[2] as well as previous studies. More details about each algorithm are provided below.

3.3.1 Naive Bayes. This algorithm relies on a probabilistic model based on Bayes theorem. It calculates probabilities by recording the frequency and combinations of values in a dataset, and works on the assumption that all attributes are independent. It is called 'naive' since this assumption rarely holds for real world applications, however it still performs well in classification[29].

3.3.2 J48 Decision Tree. The J48 classifier is a C4.5 decision tree, which is the approach that is most useful in classification problems [29]. The algorithm uses a divide and conquer approach, whereby at each node of the tree, it chooses the attribute that most successfully splits the data into subsets. The characterizing trait is the difference in entropy, where entropy is a measure of the disorder of data[30].

3.3.3 Support Vector Machines. A Support Vector Machine (SVM) algorithm plots each data point in a n-dimensional space, where n represents the number of attributes. Classification is then performed by finding the hyper-plane that differentiates the two classes. A hyper-plane is a n-1 dimensional subset of the n-dimensional space [23]. It is used to classify binary outcomes, however Weka accommodates multi-class classification.

3.3.4 k-Nearest Neighbour. K-nearest-neighbor classification is mostly used when there is little or no prior information about the underlying distribution of the data. Classification is done by assigning an object to the class that the majority of its k nearest neighbours belong to, with k being a small positive integer [31].

3.3.5 Holte's OneR. Holte's OneR is a classifier that has the goal of creating one rule that predicts the resulting class. Given a set of attributes, it chooses the most informative one to base the rule on [27].

3.3.6 Logistic Regression. Logistic regression is a technique borrowed from statistics, that is typically used to classify binary outcomes [3]. It uses a logistic function to estimate the probabilities of the categorical outcome, based on the predictor variables [22]. Similarly to SVM, Weka accommodates for multi-class classification.

3.3.7 Random Forest. Random Forest is an ensemble machine-learning algorithm. This means that it is a combination of algorithms working together for classification[4]. It creates a number of tree classifiers from subsets of the training data, and the accuracy of the classifier improves as more trees are added [5].

3.4 Testing of Machine-Learning Algorithms

The measures for evaluation of the classifier will be precision, recall and the F-measure which are commonly used for machine-learning [32]. Precision is the ratio of how many of the predicted values are actually correct [43].Recall is the ratio of how many of the actual truth labels were predicted [43]. F-measure is the harmonic mean between precision and recall[36]. Prediction of the classifier will be done via k-fold cross validation. This technique divides the dataset into k random samples, from which k-1 samples are used to train the classifier, and the remaining sample is used to test it. This process repeats k times until each and every sample has been used to test and train the classifier. Once that has been completed, the results are then averaged. For each classification, precision, recall and F-measure are automatically computed by Weka, and will be recorded. Once cross-validation has been performed, pairwise t-tests will then be used to determine which algorithm results in the best performance.

4 METHODOLOGY

This section will detail how data was collected, and how the data was processed to extract the relevant features.

4.1 Data Collection

A labelled dataset was kindly provided by Carlos Castillo from his previous work on *Information Credibility on Twitter* [11] . The data consists of over 10 000 tweets that were obtained from trending topics over a two month period. These tweets were classified whether they were newsworthy or conversational, from which, the newsworthy tweets were given a credibility rating of "CREDIBLE" or "NOT CREDIBLE". Since the conversational tweets were not able to be provided with a rating, we have labelled them as "NEUTRAL". From this data, around 8000 tweets were extracted for this project. The dataset only included the tweet IDs and the label, therefore the Twitter API was used to extract additional information needed. This information included the raw JSON file of the tweet, retweets, likes and replies, and the follow or friend relationship between the retweeters and the original user, This set of relationships will be used to determine the retweet chain, i.e. who retweeted from who. The data collected was then stored in an MySQL database to be used for training.

[1] https://www.rstudio.com/
[2] https://www.cs.tufts.edu/~ablumer/weka/README.html
[3] https://machinelearningmastery.com/logistic-regression-for-machine-learning/
[4] https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1
[5] http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

4.2 Data Pre-Processing

Figure 1[6] illustrates a concatenated example of a tweet JSON file. This file is stored as a JSON in the database, but Java recognises it as a string, so a JSON Java package[7] was used to create a JSON object from the string, and the package provided a method to get certain objects within the file. Through this method, the tweet-based and user-based features were easily extracted from the JSON files.


```
{
 "created_at":"Thu Apr 06 15:24:15 +0000 2017",
 "id": 850006245121695744,
 "id_str": "850006245121695744",
 "text": "1/ Today we're sharing our vision for the future of the Twitter API platform!\nhttps://t.co/XweGngmxlP",
 "user": {},
 "entities": {}
}
```


Figure 1: An example of a Tweet JSON file

The StanfordNLP Java Library [8] was used to perform sentiment analysis on the tweet, where it would calculate the main sentiment on the longest sentence of the tweet with a score from 0-4.

- 0 : Very Negative
- 1 : Negative
- 2 : Neutral
- 3 : Positive
- 4 : Very Positive

In regards to the network features, a retweet and a reply graph for each tweet needed to be constructed, however, to construct a graph, there had to be an indication of a relationship between the users. This was straight-forward enough for the reply graph because Twitter provides the information of which user replied to which. It was a different case for retweets, because Twitter only indicates a relationship from the user who retweeted to the the original user. So if user A posted a tweet, user B retweets it from A, and then user C retweets it from B, Twitter will not indicate this relationship; it will only state that C retweeted it from A. In order to extract this retweet chain, we iterate through the list of retweeters and find out their relationships, in other words, which retweeters are following the original poster and which are following each other. A few assumptions are made: Let user A be the original poster and user B and C retweeted from A.

- If B is following A, it is assumed that B retweeted from A.
- If C is following B, but not A, then it is assumed C retweeted from B.
- If C is following both A and B, it is assumed C retweeted from the original user, A.
- If C is following neither A nor B, then C just retweeted from the original user, A.

The relationships are stored in a string where the two unique usernames are separated with a whitespace. These strings are stored in an array and sent to a method called buildGraph(), where the graph is constructed. The pseudocode to dynamically build the graphs is found at algorithm 1. Once the graph has been constructed, the metrics can be easily computed. The data is then converted into a CSV file which is a format that can be read by Weka, as well Rstudio for statistical analysis.

Algorithm 1 Graph Builder

```
1:  procedure BUILDGRAPH(ArrayList<String> relations)
2:      nodes = new HashMap()
3:      edges = new HashMap()
4:      for String s : relations do
5:          Split s into source and destination
6:          if source is not contained in nodes then
7:              create a new source node
8:              add the source node into nodes
9:          if destination is not contained in nodes then
10:             create a new destination node
11:             add the destination node into nodes
12:         create a new edge between the source node and destination node
13:         add the edge to edges
```

4.3 Classification

Once the training data was processed into a CSV file, the Weka Java Library was used to read in the dataset and cross-validation was performed using the machine-learning algorithms provided in section 3.3, and the results can be found in section 5. The best performing algorithm was then saved to be loaded later. To classify unlabelled tweets, the server sends all the relevant information related to a tweet, the features are then extracted and converted to a CSV file where the saved classifier will then be loaded to classify the credibility.

5 ANALYSIS OF RESULTS

This section details the analysis of the features used for classification and a comparison of the performance of the machine-learning classifiers.

5.1 Performance of Feature Analysis

The extracted training dataset was analyzed in RStudio. While observing the data, it appeared that the following features : is retweet, contains profile URL, contains description URL, all appeared to have the 'true' value for every tweet. This implies that every tweet in the dataset was posted by the original user, and was not a retweet. It also implies that all the users had a profile image and an URL in their profile description, therefore these features were removed due to computational reasons. Since the response variable is categorical and has 3 classes, the use of multinomial logistic regression was appropriate. To use this regression, one of the response classes had to be chosen as a reference class, and in this case the 'NEUTRAL' class was chosen. In order to observe which features were significant, the two-tailed z test was performed and the corresponding p-values for each feature was computed. The final model was then established in an iterative fashion, whereby an insignificant feature, one with a p-value of greater than 0.05, was removed one by one, and the two-tailed z test was repeatedly computed until all features had significant p-values. The final results indicated that the features:

[6] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object
[7] https://github.com/stleary/JSON-java
[8] https://nlp.stanford.edu/software/

reply graph clustering coefficient, reply graph average betweenness centrality, followers count, friends count, friends-to-followers ratio, retweet count, is verified, and user mentions count were not significant, therefore were removed from the model.

To verify that the model has improved, the Akaike information criterion(AIC) is used. The AIC criterion is used for selecting the best model, out of a selection of models, with the best one having the lowest AIC [9]. The final model achieved an AIC of 1335.61, which is lower compared to the model of all the features with an AIC of 13356.03, therefore there was an improvement.

There were many features that had a p-value of 0, indicating that they were the most significant features, and these can be summarized in Table 4. The observed results suggest that they were all related to the retweet and reply graph metrics, therefore the graph features do play an important role in determining the credibility label. The data also suggests that the only user-based feature that remained in the model, was if the user account was verified or not, which means that these features do not seem to play a significant role in classifying credibility.

To get a closer look at the probabilities that the features have on the outcome, we calculate the relative risk by taking the exponential of the multinomial logistic coefficients. The relative risk is the ratio of probabilities, comparing the odds of one group with another, and in this case, comparing the results of CREDIBLE and NOT CREDIBLE against NEUTRAL. Taking a look at Table 5, we highlight the following results:

- Relative Risk(RR) can be values anywhere from 0 to infinity. When RR is 0, this means that none of the tweets with this feature were labelled with that particular outcome, which implies that the feature does not contribute to that outcome[35]. These values can be observed for retweet graph metrics such as the edge count, clustering coefficient, connected components count, path length, average betweenness centrality, average pageranks, average hub for the 'NOT CREDIBLE' outcome, suggesting that these features do not contribute to the tweet being fake in comparison to a neutral outcome. The data also suggests that tweets with the retweet clustering coefficient and retweet path length are very important for resulting in a 'NEUTRAL' label, as the RR is 0 for both 'NOT CREDIBLE' and 'CREDIBLE'.
- When RR is less than 1 but greater than 0, it indicates the outcome is more likely to be in the referent group 'NEUTRAL' [35]. From the results, it can be observed that majority of the reply graph metrics, and some retweet graph metrics for the 'NOT CREDIBLE' outcome are very small values, suggesting that tweets with these features are more likely to be labelled as 'NEUTRAL' than 'NOT CREDIBLE'. It is also observed that tweets with tweet-based features such as the sentiment score, character count, word count, reply count, and hashtag count are more likely to be labelled 'NEUTRAL' in comparison to both 'CREDIBLE and 'NOT CREDIBLE'.
- When RR is greater than 1, it indicates that the particular outcome was observed with a greater chance for tweets with that specific feature than for tweets without that feature. This indicates that the feature promotes the occurrence of the outcome [35]. From the results, it can be observed that majority of the reply graph metrics and some retweet graph metrics for the 'CREDIBLE' outcome seem to be very large values, indicating that tweets with these features have a greater likelihood of being labelled 'CREDIBLE' in comparison to 'NEUTRAL'. The data also suggests that if a tweet does contain a URL link, the RR of 13.149 indicates that it is also more likely to be labelled 'CREDIBLE' than 'NEUTRAL'.
- When RR is infinite, it suggests that there are tweets with that feature which exhibit the observed outcome, but tweets without that feature, do not [35]. There are several infinite values, particularly for retweet graph metrics of the 'NOT CREDIBLE' outcome. Any one unit increase in features such as the retweet node count, community count, density, diameter, average eccentricity, average closeness centrality and average eigencentrality means that it is infinitely more likely to be labelled 'NOT CREDIBLE' in comparison to 'NEUTRAL'. This suggests that these metrics are very important distinguishing factors between a tweet being labelled 'NEUTRAL in comparison to 'NOT CREDIBLE'.

Table 4: Summary of the most significant features with p-value of 0

Retweet Graph Features	
Node count	Edges Count
Community Count	Clustering Coefficient
Connected Components Count	Density
Diameter	Path Length
Average Eccentricity	Average Closeness Centrality
Average Betweenness Centrality	Average Indegree
Average Outdegree	Average Degree
Average Pageranks	Average Pageranks
Average Authority	Average Hub
Reply Graph Features	
Node Count	Edges Count
Community Count	Connected Components
Density	Diameter
Path Length	Average Eccentricity
Average Closeness Centrality	Average Eigencentrality
Average Indegree	Average Outdegree
Average Degree	Average Pageranks
Average Authority	Average Hub

Table 5: Relative Risk of the final features used for the multinomial logistic regression model with NEUTRAL as the reference category

Feature	Credible	Not Credible
Retweet Node Count	$3.961e^{21}$	Inf
Retweet Edge Count	$2.309e^{20}$	$8.418e^{-93}$
Retweet Community Count	$1.0006e^{-13}$	Inf
Retweet Clustering Coefficient	0	0
Retweet Connected Components Count	1778651564	0
Retweet Density	Inf	Inf
Retweet Diameter	$1.810e^{-24}$	Inf
Retweet Path Length	0	0
Retweet Average Eccentricity	$5.689e^{21}$	Inf
Retweet Average Closeness Centrality	$9.287e^{-24}$	Inf
Retweet Average Betweenness Centrality	Inf	0
Retweet Average Eigencentrality	3.521	Inf
Retweet Average Outdegree	$2.132e^{87}$	$6.607e^{-37}$
Retweet Average Indegree	$2.132e^{87}$	$6.607e^{-37}$
Retweet Average Degree	$4.546e^{174}$	$3.683e^{-73}$
Retweet Average Pageranks	$1.336e^{-207}$	0
Retweet Average Hub	$1.340e^{-207}$	0
Reply Node Count	$1.077e^{44}$	$9.209e^{-01}$
Reply Edge Count	$1.454e^{-85}$	$1.010e^{03}$
Reply Community Count	$7.409e^{128}$	$9.121e^{-04}$
Reply Connected Components Count	$7.409e^{128}$	$9.121e^{04}$
Reply Density	$3.597e^{93}$	$4.838e^{-01}$
Reply Diameter	$7.409e^{128}$	$9.121e^{-4}$
Reply Path Length	$7.409e^{128}$	$9.121e^{-04}$
Reply Average Eccentricity	$3.596e^{93}$	$4.838e^{-01}$
Reply Average Closeness Centrality	$3.596e^{93}$	$4.838e^{-01}$
Reply Average Eigencentrality	$2.060e^{35}$	$1.885e^{-03}$
Reply Average Outdegree	$2.060e^{35}$	$1.885e^{-03}$
Reply Average Indegree	$2.060e^{35}$	$1.885e^{-03}$
Reply Average Degree	$4.245e^{70}$	$3.554e^{-06}$
Reply Average Pageranks	$3.596e^{93}$	$4.838e^{-01}$
Reply Average Authority	$3.596e^{93}$	$4.838e^{-01}$
Reply Average Hub	$3.596e^{93}$	$4.838e^{-01}$
Sentiment Score	0.462	0.721
Word Count	0.997	0.861
Character Count	0.998	1.028
Reply Count	$1.438e^{-}173$	$7.372e^{-02}$
Like Count	1.027	1.055
Hashtag Count	0.904	0.446
Contains URL : true	13.149	1.384

5.2 Performance of Machine-Learning Algorithms

The machine-learning algorithms were run on the resulting dataset that was acheived from feature analysis, and the results from the cross-validation can be found in Table 6. Since precision, recall, and F-measure were computed for each class, 'NEUTRAL', 'CREDIBLE', and 'NOT CREDIBLE', Weka calculates a weighted average for each. It appears that the J48 Decision Tree results in the highest accuracy of 70.02%, as well as the highest scores for averaged precision (0.705), recall (0.700), F-Measure (0.684) . To determine whether these results are significantly different statistically, pairwise comparisons are performed using the t-test with the J48 Decision Tree as the base class, these results are recorded in Table 5. The negative sign indicates that the algorithm has statistically worse performance than the J48 Decision Tree, which further confirmed that the J48 Decision tree is significantly better than the other algorithms.

Table 6: Summary of the Performance of the Machine-Learning Algorithms on the Significant Features

Algorithm	Accuracy (%)	Averaged Precision	Averaged Recall	Averaged F-Measure	Pairwise t-test
J48 Decision Tree	70.0241	0.705	0.700	.684	-
Naive Bayes	63.8118	0.614	0.638	0.610	-
Holte's OneR	65.5006	0.566	0.655	0.601	-
SVM	65.5006	0.566	0.655	0.601	-
k-Nearest Neighbours	64.415	0.653	0.644	0.646	-
Logistic	66.8275	0.665	0.668	0.645	-
Random Forest	66.4053	0.663	0.664	0.663	-

6 ETHICAL, PROFESSIONAL AND LEGAL ISSUES

Permission was given by Carlos Castillo to use the dataset from *Information Credibility on Twitter* for training the classifier. The Weka, Gephi and StandfordNLP Java Libraries all operate under the open source licenses.

7 CONCLUSIONS

In this section, conclusions from the development and outcome of the project are discussed.

7.1 Graph-based features Play an Important Role in Credibility Classification

Graph-based features proved to be more significant than user-based or tweet-based features in the credibility classification of the tweets. The data also suggested that user-based features do not play an important role, as the only feature that was kept after statistical

analysis, was if the user account was verified or not. This is interesting as the follower-to-friends ratio do not play a significant role, in contrast with previous studies [3, 39], where they found the follower-to-friends ratio to be effective in detecting spam accounts. It was also observed through the calculation of relative risk ratios, that generally, reply graph metrics are more likely to result in a credible outcome compared to a neutral one, suggesting that tweets with more replies are more likely to be true. On the other hand, these features make it more unlikely for a tweet to be labelled fake in comparison to neutral, suggesting that users do not tend to reply to tweets that are fake. Retweet graph metrics including the node count, community count, density, average eccentricity, average closeness centrality, average eigencentrality are important from distingushing fake news from neutral, as the relative risk calculated was infinite. This suggests that fake tweets have more retweets than neutral ones, and the users who retweeted seem to be closely connected to each other, possibly implying that users who post fake news, have friends or connections to help them retweet to gain more attention. Overall, distinctive features can be extracted from tweets on Twitter that can classify them with a credibility label.

7.2 The J48 Decision Tree is an Effective Algorithm for Tweet Credibility Classification

After performing cross-validation and pairwise t-tests, the J48 Decision Tree was the best performing algorithm with the highest average precision, recall and F-measure, and could classify the tweets with the highest accuracy of 70.02%. These results can be compared to the research done in *Information Credibility on Twitter* as Castillo et. al were able to achieve a precision and recall between the range of 70% - 80% using a J48 Decision Tree [11], therefore our results using their dataset, fall into that range as well, however they had managed to gain an accuracy of 86.01%[11]. This means that our classifier does have room for more improvement, and possibly the features that were chosen for this project were not the most effective in comparison. Overall our classifier has satisfactory precision and recall, however accuracy can be improved.

8 FUTURE WORK

In this section, potential areas for future work are discussed. Due to the time constraints in regards to project completion as well as scope for each teammate, my teammate could not deploy the crowdsourcing application in time to collect training data for the classifiers, This would have been more ideal in gathering recent tweets as the data used for this project was collected around 2011, so it may be outdated. Also, from requirements gathering earlier on in the project, the classification categories were split into 5 classes. However, we did not have a labelled dataset with these specific labels, therefore could not implement it as originally planned. There is also the potential to continuously improve the classifier as we get feedback from the users through Kristin Kinmont's web browser extension. Should these changes come into development, it could improve the accuracy of the classifier and benefit the research community.

9 ACKNOWLEDGEMENTS

I would like to thank my teammates, Kristin Kinmont and Shaheen Karodia for their support, as well as Carlos Castillo for providing a dataset to work with. Finally, I want to give my sincere thanks to my project supervisor Selvas Mwanza for his constant support,advice and willingness to help, and co-supervisor Hussein Suleman, whose guidance helped me navigate through this project.

REFERENCES

[1] Help center. https://support.twitter.com/. Accessed: 2017-09-30.
[2] Adali, S., Escriva, R., Goldberg, M. K., Hayvanovych, M., Magdon-Ismail, M., Szymanski, B. K., Wallace, W. A., and Williams, G. Measuring behavioral trust in social networks. In *2010 IEEE International Conference on Intelligence and Security Informatics* (May 2010), pp. 150–152.
[3] Aggarwal, A., Rajadesingan, A., and Kumaraguru, P. Phishari: Automatic realtime phishing detection on twitter. In *2012 eCrime Researchers Summit* (Oct 2012), pp. 1–12.
[4] Aldhous, P. Network analysis with gephi. http://paldhous.github.io/NICAR/2015/gephi.html. Accessed: 2017-08-14.
[5] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. Tech. rep., National Bureau of Economic Research, 2017.
[6] Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. M., and Alamri, A. Reputation-based credibility analysis of twitter social network users. *Concurrency and Computation: Practice and Experience 29*, 7 (2017), e3873–n/a.
[7] Alrubaian, M., Al-Qurishi, M., Hassan, M., and Alamri, A. A credibility analysis system for assessing information on twitter. *IEEE Transactions on Dependable and Secure Computing PP*, 99 (2017), 1–1.
[8] Appelman, A., and Sundar, S. S. Measuring message credibility. *Journalism & Mass Communication Quarterly 93*, 1 (2016), 59–79.
[9] Burnham, K. P., and Anderson, D. R. Multimodel inference. *Sociological Methods & Research 33*, 2 (2004), 261–304.
[10] Canini, K. R., Suh, B., and Pirolli, P. L. Finding credible information sources in social networks based on content and social structure. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (Oct 2011), pp. 1–8.
[11] Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (New York, NY, USA, 2011), WWW '11, ACM, pp. 675–684.
[12] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. Measuring user influence in twitter: The million follower fallacy. *Icwsm 10*, 10-17 (2010), 30.
[13] Chhabra, S., Aggarwal, A., Benevenuto, F., and Kumaraguru, P. Phi.sh/$ocial: The phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference* (New York, NY, USA, 2011), CEAS '11, ACM, pp. 92–101.
[14] Conroy, N. J., Rubin, V. L., and Chen, Y. Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology 52*, 1 (2015), 1–4.
[15] DuBois, T., Golbeck, J., and Srinivasan, A. Predicting trust and distrust in social networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (Oct 2011), pp. 418–424.
[16] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. Tweetcred: A real-time web-based system for assessing credibility of content on twitter. In *Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain* (2014).
[17] Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22Nd International Conference on World Wide Web* (New York, NY, USA, 2013), WWW '13 Companion, ACM, pp. 729–736.
[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explorations 11*, 1 (2009), 10–18.
[19] Heymann, S. Gephi. https://github.com/gephi/gephi/wiki/. Accessed: 2017-08-14.
[20] Ikegami, Y., Kawai, K., Namihira, Y., and Tsuruta, S. Topic and opinion classification based information credibility analysis on twitter. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (Oct 2013), pp. 4676–4681.
[21] Kang, B., O'Donovan, J., and Höllerer, T. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (New York, NY, USA, 2012), IUI '12, ACM, pp. 179–188.
[22] Kurt, I., Ture, M., and Kurum, A. T. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications 34*, 1 (2008), 366 – 374.
[23] Mammone, A., Turchi, M., and Cristianini, N. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics 1*, 3 (2009), 283–289.

[24] McSweeney, P. J. Gephi network statistics. *Google Summer of Code*, 1–8.
[25] Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (New York, NY, USA, 2012), CSCW '12, ACM, pp. 441–450.
[26] Murthy, D. Twitter: Microphone for the masses? *Media, culture & society 33*, 5 (2011), 779–789.
[27] Nevill-Manning, C. G., Holmes, G., and Witten, I. H. The development of holte's 1r classifier. In *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems* (Nov 1995), pp. 239–242.
[28] O'Donovan, J., Kang, B., Meyer, G., HÃűllerer, T., and Adalii, S. Credibility in context: An analysis of feature distributions in twitter. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (Sept 2012), pp. 293–301.
[29] Patil, T. R., and Sherekar, S. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal of Computer Science and Applications 6*, 2 (2013), 256–261.
[30] Patil, T. R., and Sherekar, S. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal of Computer Science and Applications 6*, 2 (2013), 256–261.
[31] Peterson, L. E. K-nearest neighbor. *Scholarpedia 4*, 2 (2009), 1883.
[32] Powers, D. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies 2*, 1 (2007), 37–63.
[33] Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2011), EMNLP '11, Association for Computational Linguistics, pp. 1589–1599.
[34] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web* (New York, NY, USA, 2011), WWW '11, ACM, pp. 249–252.
[35] Sabo, D. Relative risk and the odds ratio. https://commons.bcit.ca/math/faculty/david_sabo/apples/math2441/section8/oddsratio/oddsratio.htm. Accessed: 2017-09-28.
[36] Sasaki, Y., et al. The truth of the f-measure. *Teach Tutor mater 1*, 5 (2007).
[37] Tapia, A. H., Bajpai, K., Jansen, B. J., Yen, J., and Giles, L. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In *Proceedings of the 8th International ISCRAM Conference* (2011), pp. 1–10.
[38] Wang, A. H. Detecting spam bots in online social networking sites: A machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy* (2010), Springer, pp. 335–342.
[39] Wang, A. H. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on* (2010), IEEE, pp. 1–10.
[40] WIDYANTORO, D., and WIBISONO, Y. Modeling credibility assessment and explanation for tweets based on sentiment analysis. *Journal of Theoretical and Applied Information Technology 70*, 3 (2014).
[41] Williams, N., Zander, S., and Armitage, G. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *SIGCOMM Comput. Commun. Rev. 36*, 5 (Oct. 2006), 5–16.
[42] Yang, J., Counts, S., Morris, M. R., and Hoff, A. Microblog credibility perceptions: Comparing the usa and china. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (New York, NY, USA, 2013), CSCW '13, ACM, pp. 575–586.
[43] Zhang, M. L., and Zhou, Z. H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering 26*, 8 (Aug 2014), 1819–1837.