An Approach to Detect Unreliable News Articles on Online Media

K.A.N.Malkanthi
2023

An Approach to Detect Unreliable News Articles on Online Media

A dissertation submitted for the Degree of Master of Business Analytics

K.A.N.Malkanthi
University of Colombo School of Computing
2023

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: K.A.N.Malkanthi

Registration Number: 2019/BA/017

Index Number: 19880172

Signature: Date: 02/03/2023

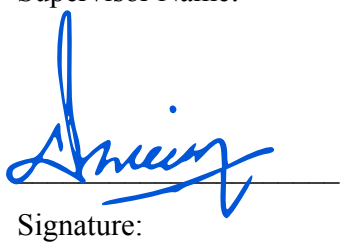
This is to certify that this thesis is based on the work of

~~Mr~~./Ms. K.A.N.Malkanthi

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name:



Signature: Date: 04/03/2023

Acknowledgements

First and foremost, I would like to offer my sincere gratitude to my project supervisor Dr. Ruvan Weerasinghe, Senior Lecturer at University of Colombo School of Computing. Who guided me by providing supervision and direction, which truly helped the progression and smoothness of the research.

Also, I would like to convey my utmost gratitude to all the other academic members of University of Colombo School of Computing – UCSC for the knowledge they passed on to me throughout this degree program.

Last but not least special thanks should be given to all of my batch mates for extending their supportive hands of friendship towards the successful drive of the research.

K.A.N.Malkanthi

2019/BA/017

University of Colombo School of Computing

Abstract

Internet-based technologies have quickly influenced all aspects of human life. It is commonly used in smartphones and tablets. People are quickly adjusting to social media platforms like Facebook and Twitter to exchange information easily and swiftly. People are searching and consuming news online rather than through traditional media such as newspapers, television, and radio. So, nowadays, internet is a major source of information. Social media and other platforms have enabled people to share information and ideas with others. In this case, the information's credibility and source are in doubt.

This study proposes novel way filter the credibility tweets which are posted in twitter using a hybrid model which a combination of checking the user level credibility and tweet content credibility. So, the focus was narrowed to the verifiability of the news text content against credible sources and the credibility of the user account that published news contents based on the Twitter posts. The two types of data—news from credible sources and ordinary user posts were encoded using FastText, word-embedding techniques. Then, the vector similarity was calculated using the cosine similarity technique against credible news items. Then, the user account features were selected, and points were assigned to them according to their contribution to the user account's authenticity. Thereafter, by considering both user account features and verified status with credible news, machine learning classifiers were trained. This study has used an SVM classifier and a Random Forest classifier, and better results were obtained with the Random Forest for the test data set.

Table of Contents

List of Figures

List of Tables

CHAPTER 1 -INTRODUCTION

1.1. Project Overview

Internet-based technologies have impacted every aspect of human life in a short period of time. It is widely used in smartphones, tablets, and other devices. In this environment, people are rapidly adapting to social media platforms like Facebook and Twitter, which are ways to share social media information easily and quickly (Google, 2021a; Murayama et al., 2021). Hence, people are consuming and seeking news from online sources rather than from traditional sources like newspapers, television, or radio. So, today, the internet is one of the main sources of information for users. Social media and other platforms have enabled people to publish several types of information and thoughts and discuss them with others who are using them (Ahlers, 2006). In this situation, credibility and the source of the information have raised uncertainties (Osatuyi, 2013).

The most common definition of false news is the deliberate spreading of false information, the deliberate spreading of false information, or the pushing of something specific as a false report of real events. Furthermore, this fake news can be considered a major threat to the economy, journalism, and influenceate spreading of false information, or the pushing of something specific as a false report of real events. Furthermore, this fake news can be considered a major threat to the economy, journalism, and influence. Democracy exists around the world. The most common example of untrustworthy news in the United States presidential election It discovered that 25% of the news outlets linked from tweets prior to the 2016 U.S. presidential election were either fake or extremely biased, and their causal analysis suggests that Trump supporters' activities influenced the activities of the top fake news spreader (Bovet and Makse, 2019). False information spreads very quickly, and it shows one thing that was removed from the site and another thing that is immediately replaced. Furthermore, people can download articles, share information, and redistribute it to others, and in the end, the false information goes so far as to be indistinguishable from the real news from its original site. Moreover, most people do not have the habit of checking the sources in correspondence. Even when the article was published, it was not read. Therefore, this can lead to the spread of untrustworthy news very quickly (Google, 2021a; Murayama et al., 2021).As a result, finding unbelievable news from platforms on the internet is especially important. Because unbelievable news and ideas spread like wildfire and mislead the people who consume them (DePaulo et al., 1997). Now several research concepts have been applied to this field of research, considering machine learning, text mining, AI (artificial intelligence), etc., and through these studies, several models have been developed. This study uses the social context information revealed by news advertising to identify network-based fake news. In general, it investigates two types of networks: homogeneous and heterogeneous networks.

Homogeneous networks are networks that contain a single type of nodes and edges, while heterogeneous networks have several types of nodes or edges.

1.2. Motivation

Considering the past decade, "fake news" has become a popular word and is the most important goal for the entire nation around the world to fight against. The most common definition of "fake news" is spreading false information, deliberately spreading false information, or pushing something specific as a dishonest report of real.

Considering the Sri Lankan context, the Sri Lankan population consumes news and other relevant information through various media such as television, newspapers, and radio. Furthermore, people are gravitating toward web-based media, which requires less time. In Sri Lanka, the spread of such false news is linked to the COVID-19 epidemic. In 2021, a large number of WhatsApp messages containing false vaccination information were widely distributed throughout the country. As a result, the vaccination program has been affected. Also, the government of Sri Lanka has twice banned social media to stop the spread of false information about the Easter Sunday attacks and the "Digana" incident (Google, 2021b, 2017)

So, research on counterfeit news is essential for the Sri Lankan context, just as it is for other countries all over the world. Furthermore, people cannot easily distinguish false news from the style of writing or content presentation (Zhou and Zafarani, 2019). The solution is to use hybrid model including user profile extracting and comparing the content of tweets and may help figure out how reliable the information is based on users and the content(Langin, 2018).

1.3. Objectives

The main objective of this study is to develop a model that can detect unreliable news articles that are spread on the internet using a hybrid approach that includes both user-level analysis and tweet content analysis. The project has followed the sub-objectives listed below to achieve the project's main goal.

Main Objective:

The main goal is to make a model for analysing tweet users and tweets with credible user tweets that can be used to find news tweets that aren't reliable.

Sub objectives:

- Identify the previous models and approaches and do a systematic review.
- Identify the user-level features from the literature.
- Determine credible sources and cross-check with the Sri Lankan Twitter community.
- Build the model for user-level analysis.
- Make a model to compare tweets from regular people with tweets from reliable sources and figure out how similar they are.
- Labelling the tweets using two models as "fake or real"

- Verify the model with SVM and Random Forest classification algorithms.

1.4. Background of the study

Fake news is a piece of information that contains unreliable details. Further, it spreads faster like a virus and cannot be identified as fake or real. This is a real issue in today's society, and it causes a loss of trust in society (de Beer and Matthee, 2021). Further detection of unreliable news articles is not an easy task, and it is a big challenge (Shu et al., 2017) . Because if these are spread, people will believe this news, and it will become a cybercrime due to affecting individuals, organizations, etc. In the 2016 US presidential election, people's opinions and decisions were affected by fake news (Bovet and Makse, 2019).To detect these unreliable news and opinions, various techniques and methods are used in past literature. Further, a lot of studies are done by most researchers. In the literature, these methods are categorized into five main categories. They are: 1) language approach; 2) topic based approach, (3) machine learning approach, (4) knowledge-based approach, and (5) hybrid approach. The language approach can be divided into sentiment analysis, or "bag of words" analysis. Furthermore, sentiment analysis is the method, which is based on each sentence's truthfulness in an article and uses scores for detecting fake news. Bag of Words Analysis is an analysis that considers each word in the article independently and uses scores for each word to detect fake news.(Kaushik, et al., 2015; Shu et al., 2017) . The N-gram model is one feature classification model in the language approach, and a lot of studies have been done in the past based on this to detect unreliable news articles (Castelo et al., 2019). The topic-based method is another approach that is used to detect fake news based on the topic of an article (Pérez-Rosas et al., 2017)[18]. The machine-learning approach is widely presented, and various algorithms and techniques are used to detect untrustworthy news articles. Models that are developed to identify unreliable news are based on these machine-learning methods. They are as follows: (1). decision tree, (2). Random Forest (3) SVM (support vector machine) (4). Naive Bayes (5) KNN (k-nearest neighbours) (Kaushik, et al., 2015; Pérez-Rosas et al., 2017).Considering the Sri Lankan context and social media, a smaller number of studies have been conducted to identify unreliable content. Network-based methods are methods that use information that is revealed in news advertising. Furthermore, network-based methods can be divided into two types: homogeneous and heterogeneous methods (Jin et al., 2016; Shu et al., 2018).

1.5. Scope of the study

Under the guise of identifying fake news, several models have been developed by researchers after many studies using different methods, such as the machine learning approach and text mining. This new study also adds new value to the area of fake news detection, and the model develops through a study based on network-based methods. This study is efficient and reduces the time required to identify fake news on social media. Therefore, fake news that is prevalent on social media can be easily identified by users.

1.6. Structure of the dissertation

- Introduction
- Literature review
- Methodology
- Results and Evaluation
- Conclusion and future works

CHAPTER 2: LITERATURE REVIEW

This section reviews previous literature on the study of the distribution of unreliable news, network analysis, text mining, and other methods.

2.1. What is Fake News

Fake news is not a new term, however, it gained popularity in the United States during the 2016 presidential election (Bovet and Makse, 2019). Prior to 2016, a significant number of researchers, as well as a huge number of intellectuals, had looked at the issue. Satirical fake news and American political discourse by (Reilly, 2012)[21] and when fake news becomes real: Combined exposure too many news sources and political attitudes of inefficacy, alienation, and cynicism by (Reilly, 2012) are two well-known prior literatures in the field (Balmas, 2014). The most popular definition of "fake news" is the deliberate transmission of incorrect information or the pushing of something specific, such as a false account of real occurrences. Furthermore, fake news poses a serious danger to the economy, journalism, and political power. It's possible. Around the world, there is democracy. The most well-known example of skewed news comes from the 2016 presidential election in the United States. It discovered that 25% of news outlets linked from tweets before the 2016 U.S. presidential election were either fake or heavily biased, and its investigation into how this occurred indicates that Trump supporters influenced the top fake news spreader's activities (Bovet and Makse, 2019) . In Sri Lanka, people get news and other important information from a variety of sources, like TV, newspapers, and radio. Furthermore, individuals are gravitating toward web-based media, which requires less time. The COVID-19 pandemic in Sri Lanka has been connected to the spread of false information. In 2021, there were a large number of WhatsApp messages circulating throughout Sri Lanka about the COVID-19 vaccination, and all of the information was false. As a result of the false information, the vaccination effort suffered a setback. Also, the government of Sri Lanka has twice banned social media to stop the spread of false information about the Easter Sunday attacks and the "Digana" incident (Google, 2017; Google, 2021b)

2.2. What is the impact of Fake News

Internet-based technology has influenced every aspect of human life in a short period of time. It can be found in a variety of smartphones, tablets, and other electronic devices. In this atmosphere, people are rapidly adopting social media platforms such as Facebook and Twitter, which allow them to simply and quickly communicate social media material (Google, 2021a; Murayama et al., 2021). Furthermore, in 2017 (Albright, 2017), Facebook was the most popular social media network. As a result, individuals prefer social media to traditional media such as newspapers, television, and radio for news. In this environment, the internet has grown and it important as a source of information for consumers. People can now use social media and other platforms to share a wide range of information and perspectives, as well as discuss them with other users (Ahlers, 2006). There are concerns about the information's veracity and source in this circumstance. Disinformation has the capacity to cause problems for millions of individuals in minutes

(Figueira and Oliveira, 2017). As a result, it may cause harm to the public, including election processes, disagreements, and public animosity.

2.3. Distribution of fake news and Detecting fake news

Due to the nature of social media, it's easy to advertise fake news, because a user can share fake news with friends, and then transmit it to their friends. Comments on fake news can sometimes increase its "credibility," leading to faster sharing and the spread of more fake news (Albright, 2017), and two types of bots known as social bots and clickbait also help spread fake news by targeting the main users in social networks (Chen et al., 2015; Shao et al., 2018) .

Detecting fake news from social networking platforms on the internet is extremely vital. Because implausible news and ideas spread like wildfire and misinform the public (DePaulo et al., 1997). Machine learning, text mining, Artificial Intelligence (AI), and other concepts have now been applied to this field of research, and various models have been built as a result of these investigations. The social context information supplied by news advertising is used in this study to identify network-based fake news. It studies two types of networks in general: homogeneous and heterogeneous networks. Here, these methods are categorized into three segments, as shown below.

(a) Content-Based Approach

The content-based approach method focuses on a human or software computer using linguistics to detect bogus news. Inside this method, it considers all the sentences and all the words in a sentence, the style of their writing, and grammar and syntax (Yang et al., 2022). This content-based approach mainly uses lexicon-based methods and machine learning approaches.

The lexicon-based approach mainly focuses on the sentences and words in a paragraph, and it has three main methods: the bag of words approach, semantic analysis, and deep syntax approach, while the machine-learning approach is based on different types of machine-learning algorithms like Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) that have been used to identify fake news. This is achieved through different types of training datasets to refine the algorithms. Further, these datasets have enabled the development of new machine learning approaches and techniques (Pérez-Rosas et al., 2017) . For most reasons, the most popular data set, the Bengali data set, used both lexicon and machine learning approaches to detect fake news (Hossain et al., 2020) .

(b) Context or Source-Based Approach

Social computing is a subfield of information technology that studies human behaviour and social connections in relation to computer networks. This is a new research area, and it is also known as social computing (Tavakolifard and Almeroth, 2012) , which is the intersection of computer science and social science. Social network analysis is a method for retrieving information about user groups and profiles, as well as social networks with inbound and outbound ties to other social profiles. This makes it easier to tell the

difference between phony and real news (De Silva, 2021). Graph representation, content mining, and semantic analysis were used to get social network profiles (Andry Alamsyah et al., 2013). The degree of differences in their profiles were compared and evaluated to find user profile attributes after identifying user groups that are more likely to propagate fake or real news. The user account aspects were investigated from various perspectives, including implicit and explicit (De Silva, 2021).

(C) Hybrid Approach

Due to the limitation, the desired result cannot be achieved while utilizing one approach. The most recent study in this situation employed a hybrid model that combines human and machine learning techniques to assist in the identification of false information on social media (Okoro et al., 2018). In a study (Deokate, 2019) designed to spot fake news on the internet, especially on social media, the study also made use of datasets from Twitter and Buzzfeed, a machine learning classifier, and the Support vector technique. Some research is also put into groups based on how polarized it is and how true the information is (Tavakolifard and Almeroth, 2012) .

CHAPTER 3 -Methodology

With the help of the previous literature review chapter, a planned research approach may be used to examine and solve a research problem. This chapter illustrates the many research methodologies used by computer scientists. This study's research method, strategy, and approach are acknowledged and discussed.

3.1. Research Problem and Research Approach

3.1.1. Research Problem

The goal of this study is to develop a hybrid approach of credibility analysis that can be used to measure the credibility of Twitter users. This approach combines network structure analysis with a credibility study of the social graph's nodes based on the characteristics humans believe are significant in determining user trustworthiness. According to studies, Twitter is one of the most active sites for crowdsourcing news during times of crisis. Twitter's social structure is set up in such a way that no two users need to follow one another in order to receive updates from one another. This is beneficial for a news story's reach since it allows ostensibly reliable information to reach a big audience. Yet, it is difficult to determine whether this information is accurate and to distinguish between news that has been verified as true, fake news, and gossips.

According to previous literature, it has created a variety of models and methodologies for determining the veracity of and filtering out bogus news that is spreading through social media platforms. However, these solutions are more time-consuming, complex, and difficult to implement. The method of solution offered is to assess the user network based on social ties and credibility based on publicly available user metadata. By building a trustworthy network of Twitter users in a community, fake news and rumors will be less likely to spread.

3.1.2. Research Approach- Mixed Method

Five distinct research methods can be used in a study: experiment, survey, archive analysis, case study, and history. Quantitative techniques encompass surveying and experimental research. Qualitative techniques include case studies, research, ethnography, action research, and grounded theory.

The goal of this thesis is to identify fake news in social networks and credible sources using a mixed method.

On one hand, this research mostly employs quantitative methodologies due to the subject matter of analysing users' perceived credibility inside a microblogging network and achieving greater accuracy. To do this, the data collected in this research will primarily be numerical in nature, with a greater emphasis on data that reflects a user's connections with and between other users in same network, as well as statistical information based on user activity that will be used to assess trustworthiness. On the other hand, because this research is dependent on characteristics of perceived user credibility, it is critical to examine the experiment results qualitatively as well. Due to their higher cognitive

capacities, humans naturally excel at recognizing and comprehending credibility and trust, as well as ranking rather than putting a score on something. So, it makes sense to look at the results of studies done in this research with people as participants.

3.2. Data Collection and Pre-processing

Since this study is about finding fake news on social networks and is based on a major incident in Sri Lanka, the Easter Sunday attack in April 2019 is used as a case study.

3.2.1. Why selecting this incident

This is a clear incident because many Sri Lankans used Twitter during the Easter attack. Using the Easter Attacks in Sri Lanka in April 2019 as a case study, this study aims to develop a mechanism for determining Twitter users' credibility in a breaking news event. Due to this incident, no publicly available data set has been collected, and a new dataset has been created.

Data collection based on Twitter and Twitter news items, as well as user information, was accomplished through the use of Twitter's open developer API. The dataset included tweets from the start of the year 2019. In this instance, two distinct sorts of data are utilized:

- Ordinary users tweet
- Tweets collected from credible sources

Considering the credible news sources, first credible source list is identified and through the survey, these sources are ranked and selected top eight credible sources and data collected based on this data sources for same incident. The tweets posted under below hashtags like "EasterSundayAttacks," "EasterSundayAttacksLK," "EasterAttacks," "EasterAttacksLK," "EasterAttacksSL,", "Sri Lanka," and "LKA" were collected.

3.2.2. How to access the twitter and data collection

- To access Twitter data, it must first obtain an API from Twitter and explain why, what, and how it will be used.
- To find the most relevant hashtags for the incident.
- Identify hashtags, and through the Twitter Search API using Tweepy, a Python package, Twitter data is collected.

3.3. Approach for building the Proposed Model

The ultimate objective of this research is to detect whether the tweets are fake or not. To do this, we need to identify users' credibility and tweets' content credibility. So, the prototype solution include with three phases. The first phase is building a model that could identify and analyse the user profiles of users who posted tweets. This will be accomplished by extracting the major and minor features of user profiles and analysing all user profile metadata. The user's credibility score will be calculated as a result of this.

The second phase is building a model that calculates the cosine similarity between ordinary tweets and credible source tweets, here, all tweets are vectorised based on the semantic meaning of the tweets using FastText.

The third phase is the combination of models one and two. This third model figures out the total score and says whether or not a normal tweet is real or fake.

Finally, the accuracy of the above model was verified using SVM and Random Forest classifiers.

3.4. Overview of the system architecture

Using primary data, this study seeks to uncover the characteristics that contribute to node (user) trustworthiness in microblogging networks. Experiments are designed to gather quantifiable data. As mentioned previously, the reliability of individuals on microblogging sites varies. Thus, the acquired data will need to be extensively examined for causal relationships between factors or variables. Inductive research will be used in this study to understand the data and come up with theories about how trustworthy users are on microblogging sites like Twitter.

High level approach in relation to the model shown in Figure 3.1 below

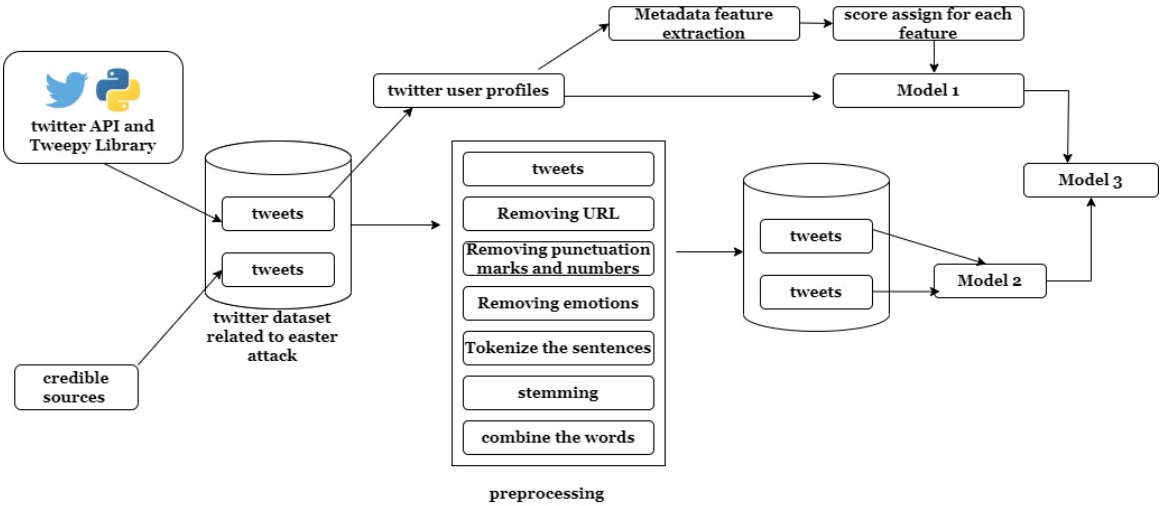


Figure 3. 1: High Level Design of Proposed Model

Before building the model, data collection is the main part of the solution design, which is described under Section 3.2. The primary dataset collected consists of ordinary user tweets and credible source tweets.

As per the existing literature discussed in Chapter 2, based on metadata collected from Twitter users, there are user credibility analysis techniques that have been investigated. Metadata is information about a single user on the network that is made available to the public and is provided by the Twitter platform. The current study in this area has produced some intriguing findings. (YeKang Yang et al., 2015). Based on the results of their work, this study tries to come up with a credibility scoring algorithm to figure out how trustworthy a user is.

Also, the proposed process will use the user metadata features that have been found in other research and give each feature a weight based on how important it is. For example, if the features are major, they will be weighed 1.0; if they are minor, they will be weighed 0.5.

Furthermore, model 2 has extracted vectors for every tweet in the dataset using FastText, which was introduced by Facebook. Both the ordinary tweets dataset and the credible tweets dataset are vectorised in this case. Finally, it will calculate the cosine similarity for every ordinary tweet in the credible source tweet dataset.

The last model, which is a mix of models 1 and 2, will decide if normal tweets are real or fake.

3.4.1. Data Collection

Data collection is the main part of the solution design and is described under Section 3.2.

3.4.2. Data Preparation

The model is built using two datasets: the first contains tweets that were posted between April 19th and May 20th, 2019. Further, this data set is related to the tweets about the Easter attack in Sri Lanka. In addition to that, this dataset was extracted with the expectation of recording verified news, rumours, and all the chaos that ensued in the immediate aftermath. Further, these tweets posted under diffrent hashtags and they are #EasterSunday, #EasterSundayAttacks, #EasterSundayAttacksLK, #EasterAttacksSL, #srilanka, and #lka.

A second set of data was taken from reliable sources for the same time period, and the survey was used to check that these sources were reliable.

It gathered information about both unique users, like ordinary Twitter users, and credible users.

Below Table 3.1 shown the list of the attributes of the dataset, which includes tweets from both ordinary and trustworthy sources.

Table 3. 1: Attributes list of extracted tweet datasets

Tweet Id	Name
Tweet URL	Screen Name
Tweet Posted Time (UTC)	User Bio
Tweet Content	Verified or Non-Verified
Tweet Type	Profile URL
Client	Protected or Non-protected
Retweets Received	User Followers
Favourites Received	User Following
Tweet Location	Favourites Count
Tweet Language	Statuses Count
User Id	User Account Creation Date

Below Table 3.2 is a list of attributes that are related to the user's metadata, made from existing literature. These features persuade every Twitter user to validate user-level credibility (YeKang Yang et al., 2015).

Table 3. 2: Features list influence to the user level credibility

screen_name
User_Id
Name
followers_count
friends_count
created_at
statuses_count
listed_count
location
verified
User_Bio

3.4.3. Data pre processing

Prior to employing any method for similarity discovery, the gathered tweets' text fields needed to be analysed. As shown in Figure 3.2, mentions, extraneous punctuation and numerals, URLs, emoticons, and retweets were eliminated.

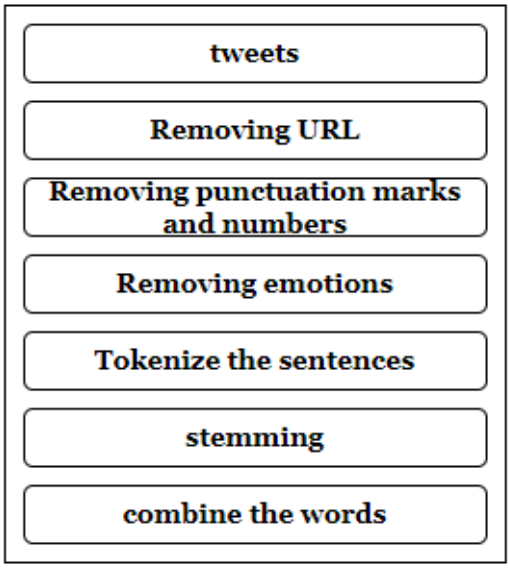


Figure 3. 2:Pre-Processing Steps

Keywords are a set of significant words in a document that give a high-level description of the content for investigating readers and are useful tools for many purposes. Furthermore, keyword extraction is necessary for many purposes, such as text categorization. Two types of approaches are used in feature extraction. The first is keyword extraction from text using a prior dictionary, and the second is keyword extraction from text without a prior dictionary. For that, a set of techniques has to be followed.

Remove Punctuation: All punctuation marks that are in the reviews have to be removed. It has also included special characters.

Remove Numbers: All numbers that do not give any meaning to the text have to be removed.

Filter noisy data: All noisy words are filtered into their correct form, such as happppppy into happy.

Remove stop words: Stop words are words that can be used to exclude something from a description. Articles (a, an, the), prepositions (in, for, from, with, within, etc.), and conjunctions (and, or, but, etc.) are included in this list. These words don't contain information about the items.

Stemming: All words that mean one word have been reduced into their stem form. Within an example, experience, experiences—all these words can be reduced to their stem form as "experience." This technique is very useful to determine the similarity and distance according to how consumers present the same feature.

Removing emotions: All emotions in the text are removed.

Remove URL: All URLs in the tweet text are removed.

Combine all words: After doing all the pre-processing, it may need to combine and make words out of bags for each and every tweet.

3.4.4. Model 1 (User Metadata based)

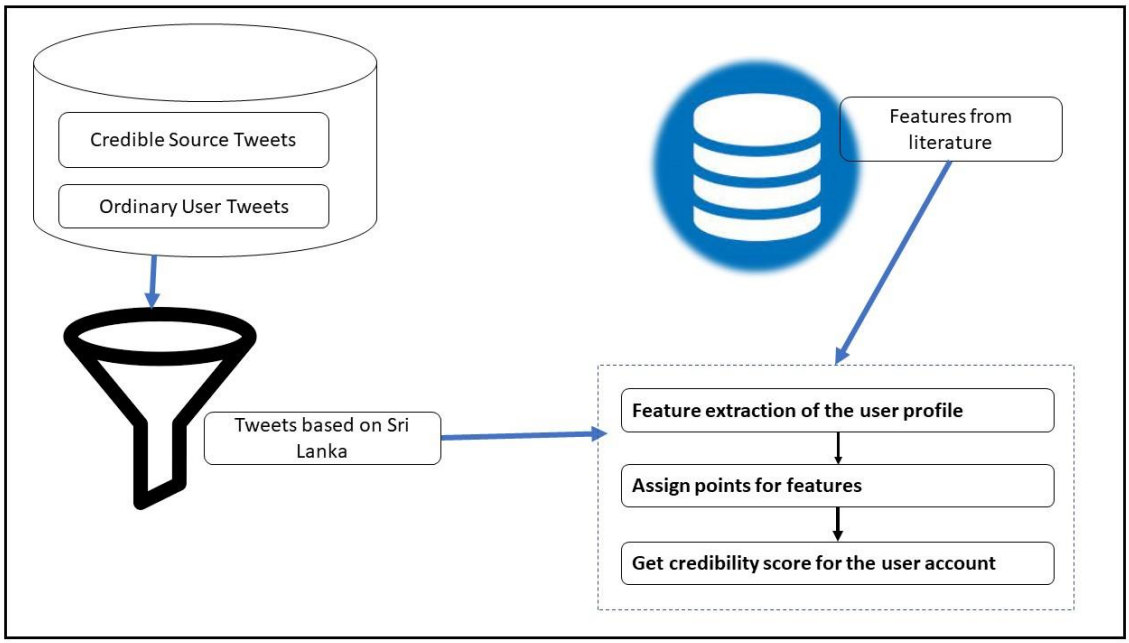


Figure 3. 3: Overview of Model 1

This model phase consists of four parts: data collection and pre-processing, extraction of features from the user profiles, scoring the features, and credibility scoring.

Data collection and pre-processing

The primary goal of this research is to develop a model for classifying tweets as fake or real; additionally, this model was developed to gain credibility among users in the Sri Lankan community, and it was used in an Easter attack in 2019. There are no publicly available datasets for this purpose, so we collected a new dataset from Twitter for a one-month period.

Extraction features of the user profiles

So, all of the data collected was related to the hashtags mentioned under the section data preparation, and all of the data was filtered based on Sri Lanka as the location. For this model, user profiles are analysed. As a result, more filters are applied to the filtered data to extract unique ordinary users and there, publicly available metadata has collected.

Table 3.3 contains the collection of metadata properties.

Table 3. 3: Collected List of metadata

No	Feature
1	Screen name of the user
2	Followers count of the user
3	Friends count of the user
4	Account created date
5	Number of status count
6	Number of listed count
7	Tweet user location
8	User account verified from twitter
9	User account has description/bio
10	User id
11	User Name

Scoring the features and credibility scoring

This step in the model employs the scoring for each user that was obtained in the previous phase. This scoring is based on the credibility features identified in prior literature as important**.** The work that has already been done in this field has identified several user credibility features that perform best as well as some other characteristics that perform averagely. In this study, a combination of these features is employed to provide a more accurate user trustworthiness rating.

This is consistent with the research goal of achieving a more precise credibility rating for all 340 network members, including those who have already been given a credible rating by the system in existence.

There are two types of features that can be found in the literature: major features and minor features. So, here are the major attributes weighted as 1.0 and the minor attributes weighted as 0.5, and Table 3.4 shows the weighted features identified from the existing literature (YeKang Yang et al., 2015)

Table 3. 4: List of credibility features with scoring

no	Feature	weight
1	No of followers/followers	1.0
2	No of Friends	1.0
3	No of status/tweets	1.0
4	Age of the account	1.0
5	Verified status	1.0
6	Location of the user	0.5
7	User description/bio	0.5
8	Number of lists	0.5

Before using the above credible features to figure out how much each user account is worth, the study has to give each user profile a score based on the above features. The scoring matrices shown in Table 3.5 are used for this purpose. Further, this scoring matrix shows how credibility scores are distributed for each feature a particular user will be evaluated against.

Table 3. 5: Scoring Matrix

Feature	**score**					
	0	1	2	3	4	5
No of tweets/status	<10	10-100	101-1000	1001-5000	5001-10000	>10000
Age of the account	<1 month	1-6 months	6 months to 24 months	24 months - 60 months	5 - 10 years	>10 years
Follower count/friends count	<0.5	0.5-1.0	1.0-2.0	2.0-5.0	5.0-10.10	>10.0

Verified status	No					Yes
Has user description/bio	No					Yes
Location of the users	No or other locations					Sri Lanka
No of lists	<2	2-10	10-30	30-60	60-90	>90

In the above scoring matrix, if a user has no location details or is not located in Sri Lanka, they receive a score of 0, and only users who are located in Sri Lanka receive a score of 5.Along this line, it identified credible Twitter users in the Sri Lankan community.

Based on the above scoring matrix, user-level features are scored between 0 and 5, and based on the weight for the feature as per Table 3.5, feature-level scores are weighted, and finally the weighted value is normalized as a 0–1 value.

Then, using the credibility equation, the following steps are taken to figure out a user's credibility score:

Major attribute weight (maj_{wt}) =1.0

Minor attribute weight (min_{wt}) =0.5

Major attributes trust score (maj_{sc}) =0.0

Minor attributes trust score (min_{sc}) =0.0

Credibility score of major attribute (maj_{credsc})

Credibility score of minor attribute (min_{credsc})

Final weighted score ($final_{wtsc}$)

Credibility score of major attribute (maj_{credsc}) calculated using below formula

$$\mathrm{maj_{sc}} = \mathrm{maj_{sc}} + \frac{\sum(\mathrm{maj_{wt}} * \mathrm{maj_{credsc}})}{number\ of\ maj\ features}$$

Credibility score of minor attribute (min_{credsc}) calculated using below formula

$$\mathrm{min_{sc}} = \mathrm{min_{sc}} + \frac{\sum(\mathrm{min_{wt}} * \mathrm{min_{credsc}})}{number\ of\ min\ features}$$

Total credibility score is the combination of both minor attributes score and major attribute score, so final score calculated using below formula

$$\text{final wighted score}(\mathrm{final_{wtsc}}) = \mathrm{maj_{sc}} + \mathrm{min_{sc}}$$

Normalized the score as 0.0 to 1.0 value and that value calculate using below formula

Here two constant values used based on literature [19].

Normalized minimum trust score (minX) =0

Normalized maximum trust score (maxX) =7.5

$$\text{Normalized Trust score} = \frac{(\mathrm{final_{wtsc}} - \mathrm{minX})}{(\mathrm{maxX} - \mathrm{minX})}$$

3.4.5. Model 2 –Text similarity model

Model two is the model used to score the tweets using the semantic meaning of the content. So, this model consists of three major modules: preparation of two datasets, including ordinary tweets and credible source tweets, vector representation, finding the cosine similarity.

Also, regular tweets are contrasted with tweets from reliable sources, allowing Mode 2 to capitalize on people's superior capacity for trustworthiness judgments.

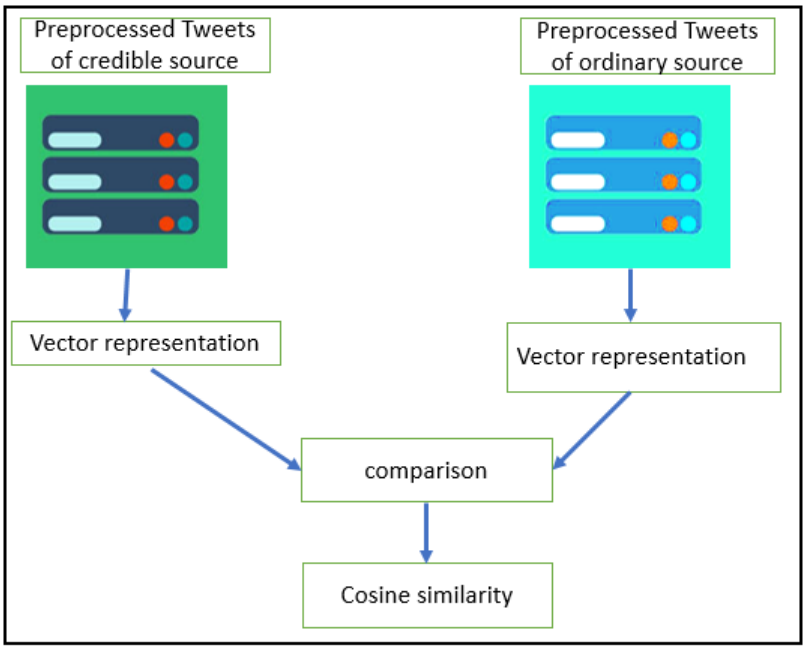


Figure 3. 4 : Overview of Model 2

Data collection and Preparations

Tweets posted by ordinary and credible users were collected for a month beginning April 19, 2019; the details of this data content are described under the section on "data preparation."

For this module, only tweet content with user details was extracted, and all tweet content was pre-processed from both datasets. Two datasets are stored separately.

Vectorization of the tweet content

To compare the two tweets' content, it needs to be converted into numeric format. For that, different types of techniques can be used, including the most popular ones such as BOW, TF-IDF, and word embedding techniques. Considering the BOW and TF-IDF techniques, they only captured the frequency of the words present in the tweets. As a result, it will not provide the most accurate comparison. So using word embedding techniques, tweet content can be vectorised based on the semantic meaning of the content. Here, we used FastText models, which were made by Facebook and trained with data from Wikipedia and CommonCrawl.

FastText has two vector representations: "getSentenceVector" and "getWordVector." So, in this module, the content of every tweet is turned into a vector based on this, and each content is made up of both vectors.

Compare and find out the Cosine Similarity

In this phase, vector comparisons were done. Here, the vector similarity of two sentences (an ordinary user tweet and a credible tweet) was calculated using the cosine similarity technique.

3.4.6. Model 3

Model three is the final phase of the fake news detection modal, and it is a summation of models one and two. So, this model three helps get the final result, which was partially obtained from models one and two.

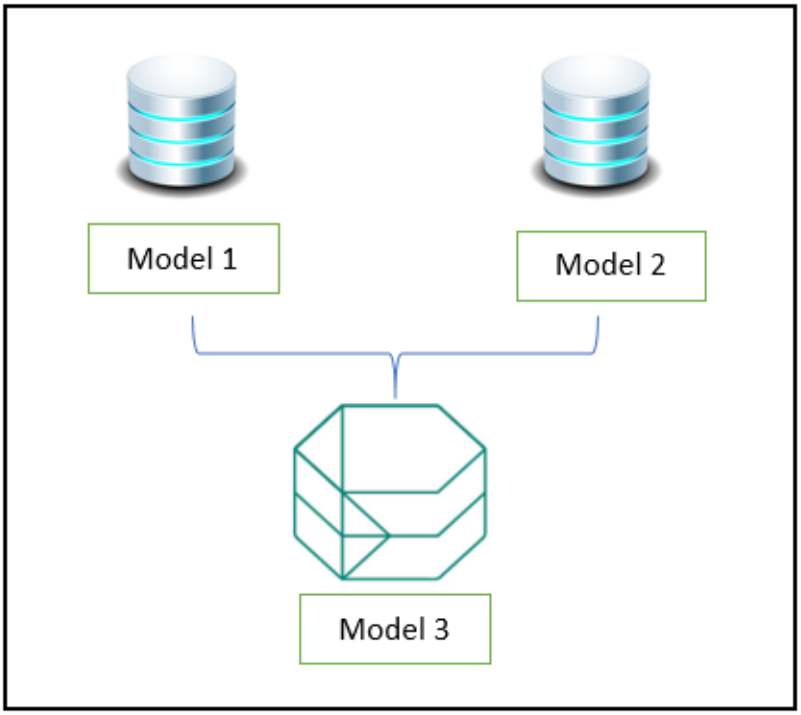


Figure 3. 5: Overview of Model 3

In model three, final credible scores are calculated based on the below calculation, and both model one and two data scores are collected for this. Compare the final score to the threshold value to determine whether the tweets are fake or real. The threshold value is calculated by comparing the data set to previously published literature. Finally, all tweets are labelled as real or fake based on this.

T : Threshold value

Normalized Trust Score: Model one score

Cosine similarity score: Model two score

Final score;

$$\text{Final score} = (\text{T} * \text{Normalized Trust score}) + ((1 - \text{T}) * \text{cosine similariy score})$$

Labelling as fake or real using as below;

$$\text{If Final score} > = 1 \text{ tweet content label} = '\text{REAL}' \text{ else tweet content label} = '\text{FAKE}'$$

3.5. Implementation of the Proposed System

Through this study, it proposed a model for labelling the tweets as fake or real through three phases of the model. So, this section discusses a proper implementation method to build the proposed credible scoring model as discussed in the previous section.

The architecture of the system comprises multiple important components. Starting with data collection, data cleaning, and analysis, and ending with building the three models, this section gives a detailed look at how to choose the tools used to get information, how to clean data, and how to build the models on the Twitter user network.

3.5.1. Tools and Technologies

Table 3.6 lists the software and hardware requirements for the research project.

Table 3. 6: List of Tools and Technologies for Project

Technology	Tools
Python	Google colab
API	Twitter developer API

3.5.2. Data Collection and Pre-processing

The aim of this study is to identify the credibility of Sri Lankan Twitter users and detect whether the tweets they post are fake or real. For this, we gathered tweets from Twitter about the deadly Easter attacks in Sri Lanka in 2019. For this reason, the authors focused on filtering tweets that were about this event. This study identifies most of the Sri Lankan tweets posted under hashtags including #lk, #lka, #SL, #SriLanka, and many more. Further, to extract the incident, it considers hashtags, including key words, as "Easter Attack."

Authors select the hashtags that are most relevant to the incident and use them to collect data.

Location of the data set: Sri Lanka

Time period: 19th Apr 2019 -20th May 2019

Ordinary data set

As per the above section, using selected hashtags, we collected ordinary tweets and put them into CSV format. After analysis, the dataset and unique users were identified. In addition, user metadata for each and every unique user is extracted.

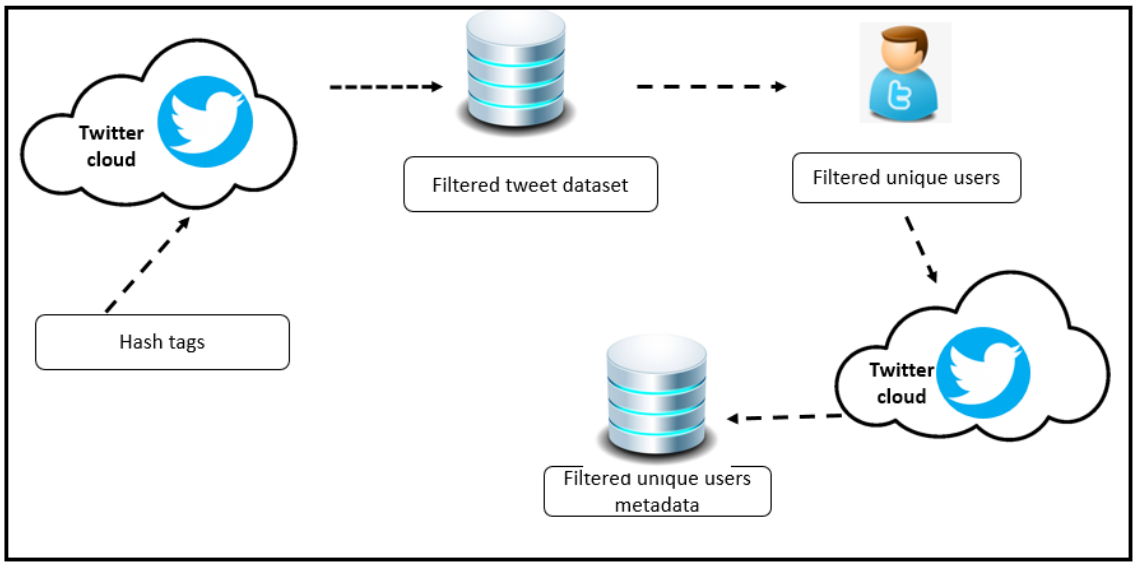


Figure 3. 6: Overview of collecting Ordinary Data set

Credible user's tweet dataset

The authors of this study have chosen 10 reliable sources from the Sri Lankan community based on the existing literature. To verify these sources' credibility, this study conducted a survey and ranked the 10 users as per the responses from the Twitter community in Sri Lanka. Based on data observation, the top eight credible sources are chosen for the next level of study.

Both metadata and tweets posted by these eight credible sources are collected

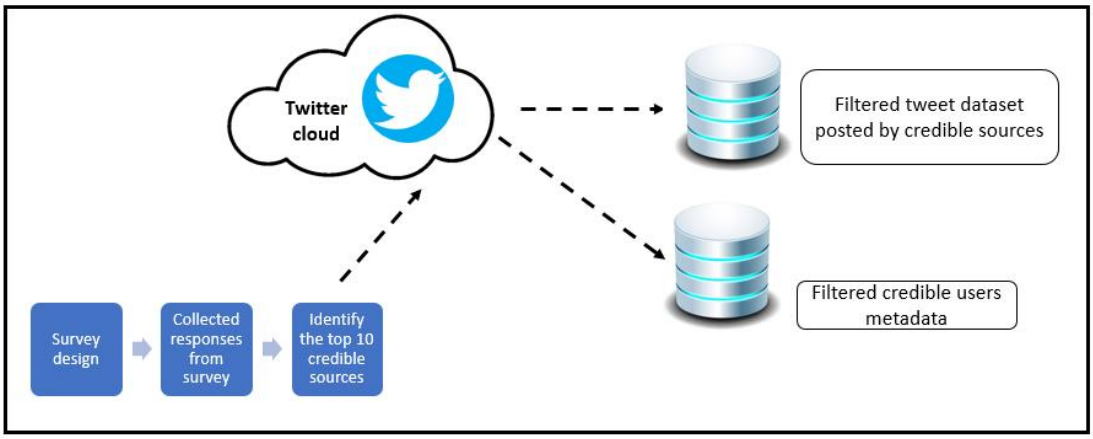


Figure 3. 7: Overview of collecting credible source tweets

3.5.3. Implement the Model 1

As described in Section 3.4.4, Model 1 was implemented to extract the user-level credibility score. For that, pre-collected user metadata is imported into the Python workbook, which applies rules and weights as discussed in the same section.

3.5.4. Implement the Model 2

As described about in Section 3.4.5, two tweet datasets that have already been processed are brought into the Python Colab workbook. These two datasets are the credible source tweet dataset and the ordinary tweet dataset. Through the data pre-processing steps described in Section 3.4.3, both datasets are pre-processed for further steps. In addition to that, the FastText model was also imported with other Python libraries, and vectorization was done for the datasets. Finally, these two sets are evaluated, and cosine similarity is generated.

3.5.5. Implement the Model 3

As discussed in Section 3.4.6, all two models' scores are imported into the Python Colab workbook, and a final credibility score is generated. In addition to that, ordinary tweet items are labelled as "fake" or "real" based on the threshold value.

3.6. Classification and validation

Then a data normalization process was performed on the prepared dataset, which came from Model 3. This data set was split into two groups. 70% data for training and 30% data for testing. Here, for the study, it used two machine learning classifiers to verify and validate the results from the above-implemented model. For that, we use support vector machines (SVM) and random forests (RF). First, we trained the classifiers and then tested the classifiers. Finally, the model's accuracy was determined by comparing the results to the model's results.

CHAPTER 4 –Results and Evaluation

4.1 Introduction

This chapter's goal is to summarize the results of the experiments that were designed and carried out based on the research method described in Chapter 3. This proposed study discussed three models that are intertwined: the first model scores users based on their metadata. The second model used Fasttext, and Fasttext introduces two possible ways to get vector representations of sentences: "getWordVector" and "getSentenceVector." In this section, all tweets about the Easter attack from Sri Lanka are compared to those from credible sources about the same incident. Further, the final model is designed to take the overall score and label the tweets as fake or real.

4.2 Data Collection

The data collection for models 1 and 2 is done. In model 1, all tweets about the Easter Attacks are collected, filtered based on where they came from (Sri Lanka), and all user information is taken. The second model captures the credible sources and their tweets for the Easter attack in Sri Lanka for the same time period, and these tweets are included with the source tags, which can be verified to ensure that the tweets are credible.

4.2.1 Data Collection Results

For the first and second models, a total of 10000 ordinary tweets were collected, and 8096 unique users were extracted based on their location to create 400 user profiles. Some samples are shown in below Table 4.1.

Table 4. 1: Tweet content collected from twitter

Tweet user	Tweet content
@munza14	Explosions at St. Anthony's Church, Kochchikade, several other churches in Negombo & Batticaloa and at Shangri-La Hotel, Colombo as well as Cinnamon Grand.#LKA #SriLanka OH GOD! WHAT JUST HAPPENED!
@ChandaniKirinde	#eastersundayattackslk #srilankaattacks #srilankan #colombo #attacks #sri #srilanka #lka #india #isis #easter #news #muslims #police #country #sunday #government #today https://t.co/xI4A8BjXVI

Further, all tweet content shown in a word cloud as Figure 4.1 in below.

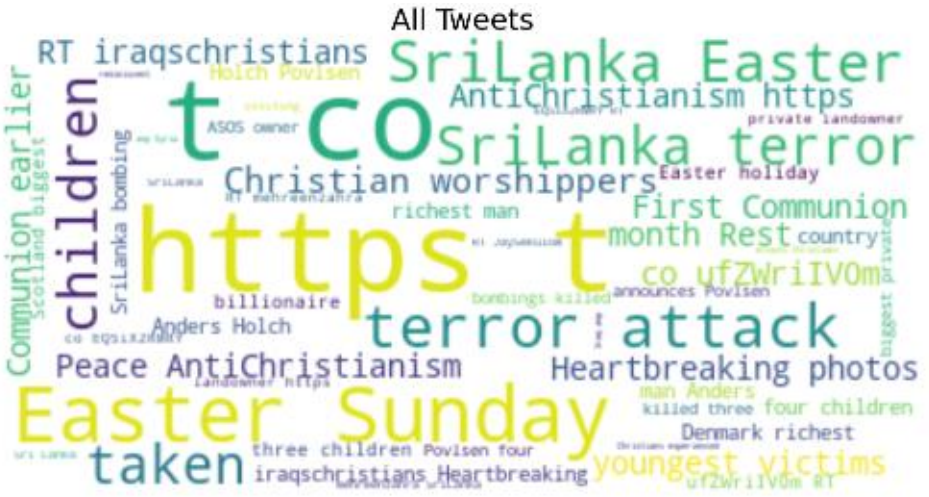


Figure 4. 1: Word Cloud of all ordinary tweets

Taking into account the credible sources, we have selected the users listed below as credible sources, and their metadata is as follows:

screen_name	User_Id	Name	followers_count	friends_count	created_at	statuses_count	listed_count	location	verified	User_Bio
BBCWorld	742143	BBC News (Wor	38978742	20	2007-02-01 7:44	354333	133158	London, UK	TRUE	News, features ar
AzzamAmeen	140784400	Azzam Ameen	401196	877	2010-05-06 10:5	30661	539	Colombo	TRUE	Journalist - Colom
SriLankaTweet	41786801	Sri Lanka Tweet	240768	36704	2009-05-22 10:1	45181	434	Sri Lanka	TRUE	Ayubovan 🙏 LK!
adaderana	176337215	Ada Derana	528141	8	2010-08-09 7:35	76588	497	Sri Lanka	FALSE	A premier breakin
NewsfirstSL	339564751	Newsfirst.lk Sri L	494757	6	2011-07-21 9:17	127232	566	Colombo, Sri La	FALSE	Sri Lanka's Large https://t.co/b5felZ https://t.co/mvJor https://t.co/tiqxNR https://t.co/HfbA2
DailyNews_lk	1879618076	Daily News	15069	89	2013-09-18 14:0	30137	158	Lake House, Sri	FALSE	Sri Lanka's premi
MarianneDavid2	2449625713	Marianne David	32604	640	2014-04-17 12:0	22848	196	Colombo, Sri La	FALSE	Journalist. Deputy Editor - TI #everylittlecounts https://t.co/jkuMF
HarshadeSilvaM	1968865952	Harsha de Silva	349263	30	2013-10-18 13:1	5657	261	Colombo, Sri La	TRUE	Father. Husband.
RiviraNews	1222141574840	Rivira.lk	479	24	2020-01-28 12:5	1512	7		FALSE	Sri Lanka News V #BreakingNews #
Dailymirror_SL	66329707	DailyMirror	585576	36	2009-08-17 10:5	88223	763	Colombo, Sri La	TRUE	Premier Breaking

Figure 4. 2: Credible sources Metadata details

4.2.2. Survey Results

As previously discussed, a survey was conducted to identify the most credible news sources on Twitter. For that, consider the 10 credible sources, which are listed and ranked. This survey was done in the Sri Lankan Twitter community and received 100 responses. The survey results are shown below in Figure 4.3.

As per the survey result, it will consider the first eight sources for the next step in the research.

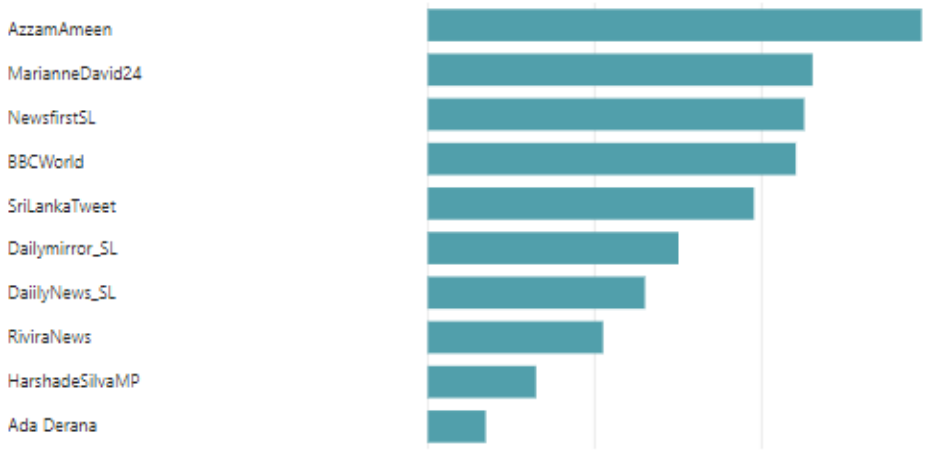


Figure 4. 3: Survey Results

The list of highly credible Twitter profiles consists mostly of well-known journalists in Sri Lanka and news media outlets, both local and international. After analysing the survey results, it has identified Aszam Ameen, a journalist from Newswire.lk who has also worked as Sri Lanka's news correspondent for BBC News, as the most credible Twitter profile for obtaining credible news in the local arena, based on an overwhelming number of votes of confidence. Further down the list, it found less trust in Ada Derana and Harshade Silva MP. The other eight sources are considered highly credible, and here we used them to extract tweets for the incident described in the previous section.

4.3. Model 1 (User Metadata based) – Results and Analysis

Model 1 is designed to calculate and analyse the user profiles based on the user metadata features and weight. Here is the analysis based on the features of the user profiles that were identified based on the literature, and there are two types of features based on level of influence. There are major attributes and minor attributes, as discussed in the previous section. Furthermore, major attributes are weighted at 1.0 and minor attributes are weighted at 0.5.

The first ordinary dataset that had already been processed was filtered by location as Sri Lanka, and unique users were found. Some of the unique users are listed below in Figure 4.4.

User_Id	Name	Screen_Name
1951067186	Chandani Kirinde LK	ChandaniKirinde
9909452	Vipulananda LK	vipulananda
957681564284006401	Asiri Fernando	AsiriFernandoLK
1100241609721958400	Lankanenews	lankanenews
1869917396	Wasitha Pinnawala	WasithaKeminda
4611985095	MarcMartín	MarcMartinn
122884856	Lankanorg News	LankanOrg
40320649	neetwit	Neetwit

Figure 4. 4: Ordinary Users in Sri Lankan Location

Using this unique user list, their metadata is extracted, and a portion of the metadata is shown in Figure 4.5.

screen_name	User_Id	Name	followers_count	friends_count	created_at	statuses_count	listed_count	location	verified	
ChandaniKirinde	1951067186	Chandani Kirinde LK	11129	1392	2013-10-10 06:53:14	50061	152	Colombo,Sri Lanka	False	
vipulananda	9909452	Vipulananda LK	3046	5000	2007-11-03 07:37:38	124277	353	Colombo, Sri Lanka	False	6
AsiriFernandoLK	957681564284006401	Asiri Fernando	1207	2119	2018-01-28 18:27:32	17718	23	Sri Lanka	False	
lankanenews	1100241609721958400	Lankanenews	597	1784	2019-02-26 03:50:16	3123	0	Sri Lanka	False	
WasithaKeminda	1869917396	Wasitha Pinnawala	337	546	2013-09-16 02:07:12	80	0	Pinnawala	False	
MarcMartinn	4611985095	MarcMartín	57	155	2015-12-26 15:30:38	1820	0	Sri Lanka	False	
LankanOrg	122884856	Lankanorg News	140	574	2010-03-14 06:54:48	6430	0	Sri Lanka	False	#S
Neetwit	40320649	neetwit	2934	2867	2009-05-15 20:06:00	191839	15	Colombo	False	
Francanstyle	301807277	Francanstyle	81	347	2011-05-20 03:02:43	1611	0	Sri Lanka	False	
vikalpavoices	114960150	Vikalpa	16105	151	2010-02-17 04:34:35	35414	143	Colombo, Srilanka	True	

Figure 4. 5: Metadata for Ordinary Users

Finally, the ordinary unique user's metadata is extracted and applied to Model 1, and the results of Model 1 are shown here step by step.

First step

Using the rules described in Chapter 3, scores are calculated and stored in a Pandas data frame according to the user level.It has shown in Figure 4.6

User_Id	Name	verifiedStatus	getAgeFromCreatedDate	calculateAgeScore	followerRatioScore	tweetCountScore	User_Bio	user_location	listCountScore
1951067186	Chandani Kirinde LK	0	68	4	4	5	5	5	5
9909452	Vipulananda LK	0	140	5	1	5	5	5	5
957681564284006401	Asiri Fernando	0	15	2	1	5	5	5	0
1100241609721958400	Lankanenews	0	2	1	0	3	5	5	0
4611985095	MarcMartín	0	41	3	0	3	0	5	0

Figure 4. 6: User Level scores for features

Second Step

Here Major attributes are the age of the account, the number of followers and friends, the verified status, and the number of tweets. In addition to that, the following are considered optional attributes: location, number of lists, and user profile description. Based on the formula in the section 3.4.4. Weighted mandatory trust score and a weighted optional trust score were calculated. The final weighted score is the sum of both the mandatory and optional trust scores.

User_Id	Name	weighted_mandatory_trust_score	weighted_optional_trust_score	final_weighted_score
1951067186	Chandani Kirinde LK	3.25	2.500000	5.750000
9909452	Vipulananda LK	2.75	2.500000	5.250000
957681564284006401	Asiri Fernando	2.00	1.666667	3.666667
1100241609721958400	Lankanenews	1.00	1.666667	2.666667
4611985095	MarcMartín	1.50	0.833333	2.333333
122884856	Lankanorg News	2.00	1.666667	3.666667
40320649	neetwit	3.00	2.000000	5.000000
301807277	Francanstyle	1.75	1.666667	3.416667
114960150	Vikalpa	4.75	2.500000	7.250000

Figure 4. 7: Model 1 Final weighted score

Third step

Further, the final weighted score is normalized and put into a normalized value, which is 0 to 1, based on the formula described in Chapter 3 under 3.4.4. Section.

User_Id	Name	normalized_final_weighted_score
1951067186	Chandani Kirinde LK	0.766667
9909452	Vipulananda LK	0.700000
957681564284006401	Asiri Fernando	0.488889
1100241609721958400	Lankanenews	0.355556
4611985095	MarcMartín	0.311111
122884856	Lankanorg News	0.488889
40320649	neetwit	0.666667
301807277	Francanstyle	0.455556
114960150	Vikalpa	0.966667

Figure 4. 8: Normalized Trust Score

4.4 Model 2 – Results and Analysis

In model 2, each tweet's content was used and vectorised as word and sentence vectors using fastText, which was introduced by Facebook. On the other hand, for credible source tweets, the same procedure was followed, and two vectors were created for each and every tweet.

User_Id	Tweet_Content	Tweet_sentence_vec	Tweet_word_vec
1100241609721958400	eastersundayattacksik srilankaattacks srilankan colombo attack sri srilanka lka isi easter news ...	[-0.04657008, -0.005448333, -0.020174017, 0.027212488, -0.044868317, -0.016738586, 0.05076768, 0...	[-0.0034655333, -0.0006858973, -0.0005793686, 0.0075466475, 0.0033139696, -0.016494058, -0.00753...

Figure 4. 9: Model 2 two vectors

Additionally, using cosine similarity, each ordinary tweet was compared to credible source tweets, and the similarity for each tweet was calculated.

User_Id	Tweet_Content	Tweet_sentence_vec	Tweet_word_vec	Tweet_sentence_com	cosine_sim_senetence	Tweet_word_vec_com	cosine_sim_word
22595284	srilanka hotel occupancy drop following easter sunday bombing	[-0.04319242, 0.044449653, -0.017805653, 0.053325735, -0.004718261, 0.025657924, 0.037084598, 0....	[-0.009995355, 0.0011599448, -0.0055440776, 0.01640812, -0.0026002633, -0.0134268375, -0.0027348...	[-0.04842251539230347, -0.009690969251096249, -0.021630277857184 41, 0.01949372328817 8444, -0.048...	0.651410	[-0.04842251539230347, -0.009690969251096249, -0.021630277857184 41, 0.01949372328817 8444, -0.048...	0.321643
557916099	rt bbhuttozardari devastating news coming sri lanka always purveyor terror chosen target innocen...	[-0.018647559, 0.0030466549, 0.004072194, 0.073729455, -0.041547816, -0.0100892, 0.031269863, 0....	[0.00010299351, -0.006690341, -0.001178866, 0.00909946, -0.011208518, -0.0089183925, -0.00436873...	[-0.04842251539230347, -0.009690969251096249, -0.021630277857184 41, 0.01949372328817 8444, -0.048...	0.700531	[-0.04842251539230347, -0.009690969251096249, -0.021630277857184 41, 0.01949372328817 8444, -0.048...	0.379003

Figure 4. 10: Cosine similarities for both vectors

4.5 Model 3 – Results and Analysis

Model three is a combination of Models One and Two, and here is the sum of Models One and Two using a formula described in Chapter 3 under Section 3.4.6. Finally, it labels ordinary tweets as "real" or "fake" based on the final trust score and the threshold value.

User_Id	Tweet_Id	trust_score_sentence_level	trust_score_word_level
1.951067e+09	1130420279627395072	0.520259	0.462645
9.909452e+06	1130150275766001664	0.551221	0.466382
9.576816e+17	1130098187124051968	0.382172	0.327333
1.100242e+18	1130097107333836800	0.440195	0.276576
4.611985e+09	1130043387724353536	0.379851	0.241553

Figure 4. 11: Model 3 results

User_Id	Tweet_Id	trust_score_sentence_level	trust_score_word_level	Label_word_level	Label_senetence_level
1.951067e+09	1130420279627395072	0.520259	0.462645	Fake	Real
9.909452e+06	1130150275766001664	0.551221	0.466382	Fake	Real
9.576816e+17	1130098187124051968	0.382172	0.327333	Fake	Fake
1.100242e+18	1130097107333836800	0.440195	0.276576	Fake	Fake
4.611985e+09	1130043387724353536	0.379851	0.241553	Fake	Fake

Figure 4. 12: Labelled tweets data

4.6. Classification and Evaluation

The purpose of this study is to identify untrustworthy news pieces that are published on microblogging platforms such as Twitter. Thus, this study emphasized the need of following reliable individuals on Twitter in order to acquire credible news amid breaking news events, owing to the unprecedented volume of misinformation being spread due to the microblogging platform's absence of a fact-checking process. One technique to determine the trustworthiness of a tweet is to look at the individual who posted it.

Further, this will extract the ordinary tweets and compare the credible source tweets for the same incident. A list of credible sources was built from literature and a survey conducted in the Sri Lankan Twitter community to verify the list and identify the most highly credible sources. So, in both cases, assume that credible users are posting news that can be considered credible. This study focused on tweets about Sri Lanka during the 2019 Easter attack incident. All data for this study was lost due to users' failure to use standard hashtags. Furthermore, as always, the data tweets captured from credible sources include source links related to government news.

4.6.1. Classification results

As discussed in Chapter 3, the results dataset from Model 3 is broken down into two sets: the train and test datasets. Further, FastText introduced two possible methods to vectorize the data: "Getsentencevector" and "Getwordvector". The first results dataset was classified and verified in order to determine the best suitable method. As a result, it was discovered that "Getsentencevecor" provides better performance, including accuracy.

For this small dataset has used.

Table 4. 2 : Comparison for two vectors

	get_word_vectors	get_sentence_vector
Accuracy	0.5036	0.73
Precision	0.536	0.71
Recall	0.589	0.79

This results obtained from below classifiers

(a) SVM classifier

Here we used two methods to identify whether the dataset was linear or nonlinear. As a result, training data were used for the classification process, which was carried out using SVM and two techniques: linear and RBF kernels.

Table 4. 3: Classification results of SVM

Label	F1 Score		Precision		Recall		Accuracy	
	SVM_linear	RBF	SVM_linear	RBF	SVM_linear	RBF	SVM_linear	RBF
Real	0.75	0.8	0.76	0.82	0.80	0.82	0.80	0.81
Fake	0.84	0.94	0.78	0.80	0.81	0.82	0.82	0.85

The F1-score of SVM using a linear kernel was 75% -84%, and accuracy was 80%-82%. The F1 score of SVM with a RBF kernel was 80%-94%, and accuracy was 81% -85%. SVM using the RBF kernel shows the highest F1-score and accuracy out of the two SVM models. Above table shows the test results of SVM with linear kernel and SVM with RBF kernel classifier. According to that, it shows data are more non-linear.

(b) Random Forest

Random Forest (RF) gave an accuracy of 85% and an F1-score of 85% as the outcome with the same dataset as in Table 10. When the results of SVM, RBF kernel, and RF are compared, it is clear that RF has the highest accuracy and F1 score.

Table 4. 4: Classifier results of Random Forest

Label	F1 Score	Precision	Recall	Accuracy
Fake	0.82	0.80	0.81	0.85
Real	0.92	0.85	0.83	0.83

CHAPTER 5 –Conclusion and Future Work

5.1. Conclusion

In this study, we suggested a hybrid method—a combination of user-account-associated characteristics and text verifiability against reliable sources—to identify fake news on social media platforms. Using a rule-based approach, the hybrid feature methodology was put into practice. Fake news is made-up information or a deception that is created to mislead a crowd. With the rise in social media news searches nowadays, the dissemination of disinformation has changed from random events to planned and methodical actions. Bogus news detection is a big topic, so we had to create a dataset with fake news in it. We applied five rules to determine a user account's credibility score. Features that were appropriate and affected a user account's believability were noted in the literature. The similarity score between each user's tweet and each reliable tweet was calculated in the text verification module. We then developed an overall trustworthiness score for user tweets using a different formula. Here, the contributions of the user account credibility score and the text verifiability score were both taken into account.

5.2. Limitations

This investigation concentrated on tweets concerning Sri Lanka during a big 2019 incident. Because it only considered tweets that included a particular set of hashtags, the tweets are from a Sri Lankan context. However, Twitter users might not always use hashtags, particularly during noteworthy events. As a result, it's probable that some of the important tweets that were posted in the first few hours after the catastrophe were missed in the data gathering.

5.3. Future Works

This model is based on the use of both content and context methods, resulting in a "hybrid" approach. As a result, the build model must be extended to achieve greater accuracy, and it can further improve the analysis by utilizing the network of users and datasets that can be extracted using it.

References

Ahlers, D., 2006. News Consumption and the New Electronic Media. Harv. Int. J. Press. 11, 29–52. https://doi.org/10.1177/1081180X05284317

Albright, J., 2017. Welcome to the Era of Fake News. Media Commun. 5, 87–89. https://doi.org/10.17645/mac.v5i2.977

Andry Alamsyah, Budi Rahardjo, Kuspriyanto, 2013. Social Network Analysis Taxonomy Based on Graph Representation. https://doi.org/10.13140/2.1.3221.2160

Balmas, M., 2014. When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism. Commun. Res. 41, 430–454. https://doi.org/10.1177/0093650212453600

Bovet, A., Makse, H.A., 2019. Influence of fake news in Twitter during the 2016 US presidential election. Nat. Commun. 10, 7. https://doi.org/10.1038/s41467-018-07761-2

Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., Freire, J., 2019. A Topic-Agnostic Approach for Identifying Fake News Pages, in: Companion Proceedings of The 2019 World Wide Web Conference. pp. 975–980. https://doi.org/10.1145/3308560.3316739

Chen, Y., Conroy, N.J., Rubin, V.L., 2015. Misleading Online Content: Recognizing Clickbait as "False News," in: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. Presented at the ICMI '15: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, ACM, Seattle Washington USA, pp. 15–19. https://doi.org/10.1145/2823465.2823467

de Beer, D., Matthee, M., 2021. Approaches to Identify Fake News: A Systematic Literature Review, in: Antipova, T. (Ed.), Integrated Science in Digital Age 2020, Lecture Notes in Networks and Systems. Springer International Publishing, Cham, pp. 13–22. https://doi.org/10.1007/978-3-030-49264-9_2

De Silva, T.H.M., 2021. A Network Analysis Based Credibility Ranking Model to Combat Misinformation on Twitter (Thesis).

Deokate, S.B., 2019. Fake News Detection using Support Vector Machine learning Algorithm. Int. J. Res. Appl. Sci. Eng. Technol. 7, 438–444. https://doi.org/10.22214/ijraset.2019.7067

DePaulo, B.M., Charlton, K., Cooper, H., Lindsay, J.J., Muhlenbruck, L., 1997. The Accuracy-Confidence Correlation in the Detection of Deception. Personal. Soc. Psychol. Rev. 1, 346–357. https://doi.org/10.1207/s15327957pspr0104_5

Figueira, Á., Oliveira, L., 2017. The current state of fake news: challenges and opportunities. Procedia Comput. Sci. 121, 817–825. https://doi.org/10.1016/j.procs.2017.11.106

Google, 2021a. How to Spot Real and Fake News - Critically Appraising Information [WWW Document]. URL https://www.mindtools.com/a0g6bjj/how-to-spot-real-and-fake-news (accessed 11.19.22).

Google, 2021b. Sri Lanka's Solution To #FakeNews: Appropriate Or Not? - Roar Media [WWW Document]. URL https://roar.media/english/life/current-affairs/what-is-sri-lanka-doing-to-combat-fake-news (accessed 11.19.22).

Google, 2017. Disinformation in Sri Lanka: An overview. Groundviews. URL https://groundviews.org/2017/07/04/disinformation-in-sri-lanka-an-overview/ (accessed 11.19.22).

Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S., 2020. BanFakeNews: A Dataset for Detecting Fake News in Bangla.

Kaushik, K., Kaushik, A., Naithani, S., 2015. A Study on Sentiment Analysis: Methods and Tools. Int. J. Sci. Res. IJSR 4, 287–292. https://doi.org/10.21275/v4i12.NOV151832

Langin, K., 2018. Fake news spreads faster than true news on Twitter—thanks to people, not bots [WWW Document]. URL https://www.science.org/content/article/fake-news-spreads-faster-true-news-twitter-thanks-people-not-bots (accessed 11.19.22).

Murayama, T., Wakamiya, S., Aramaki, E., Kobayashi, R., 2021. Modeling the spread of fake news on Twitter. PLOS ONE 16, e0250419. https://doi.org/10.1371/journal.pone.0250419

Okoro, E.M., Abara, B.A., Umagba, A.O., Ajonye, A.A., Isa, Z.S., 2018. A hybrid approach to fake news detection on social media. Niger. J. Technol. 37, 454. https://doi.org/10.4314/njt.v37i2.22

Osatuyi, B., 2013. Information sharing on social media sites. Comput. Hum. Behav. 29, 2622–2631. https://doi.org/10.1016/j.chb.2013.07.001

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R., 2017. Automatic Detection of Fake News.

Reilly, I., 2012. Satirical Fake News and/as American Political Discourse. J. Am. Cult. 35, 258–275. https://doi.org/10.1111/j.1542-734X.2012.00812.x

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.-C., Flammini, A., Menczer, F., 2018. The spread of low-credibility content by social bots. Nat. Commun. 9, 4787. https://doi.org/10.1038/s41467-018-06930-7

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explor. Newsl. 19, 22–36. https://doi.org/10.1145/3137597.3137600

Tavakolifard, M., Almeroth, K., 2012. Social computing: an intersection of recommender systems, trust/reputation systems, and social networks. IEEE Netw. 26, 53–58. https://doi.org/10.1109/MNET.2012.6246753

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S., 2022. TI-CNN: Convolutional Neural Networks for Fake News Detection.

YeKang Yang, Kai Niu, ZhiQiang He, 2015. Exploiting the topology property of social network for rumor detection, in: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). Presented at the 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, Songkhla, Thailand, pp. 41–46. https://doi.org/10.1109/JCSSE.2015.7219767

Zhou, X., Zafarani, R., 2019. Network-based Fake News Detection: A Pattern-driven Approach.