Rating the Credibility of Online News Stories

J. A. D. P. K. A. JAYASINGHE

Rating the Credibility of Online News Stories

J. A. D. P. K. A. JAYASINGHE
Index No : 13000497

Supervisor : Dr. Ruvan Weerasinghe

December 2017

Submitted in partial fulfillment of the requirements of the BSc in Computer Science Final Year Project (SCS4124)

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: Ms. J.A.D.P.K.A. Jayasinghe

Signed: _____

Date: _____

This is to certify that this dissertation is based on the work of Ms. J.A.D.P.K.A Jayasinghe under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor Name: Dr. A. R. Weerasinghe

Signed: _____

Date: _____

Abstract

The internet has become one major platform where people get information regularly. News reading on the web has increased throughout these years. At the same time number of visits for a news website has increased. Hence today people try to manipulate information on the web in many ways. So here comes a new problem called fake news which can do considerable influence on events such as elections, natural disasters etc. With the involvement of the social media, this problem has become even bigger, because the information in social media is not monitored can be manipulated easily. And social media has the power of spreading information in less amount of time. Therefore identifying credible information from the web has become really important today.

In this thesis, we propose a novel approach for ranking the credibility of the website/source based on their behavior on the web. In order to determine the credibility of online news stories, first we need to determine the credibility of the news website. Hence this research is focusing on determining the credibility of the news website. So the approach followed in this research is based on the behavior of news stories on Twitter. The literature discusses three main user influencing factors in Twitter. They are In-degree, Mentions and URL Recommendation. So based on these factors three different models are developed that produce credibility rankings for news websites. And finally, a survey is conducted with experienced and reputed journalists in Sri Lanka to evaluate credibility ranking values produced by the models. According to the experiments carried out, it indicates, the factor URL Recommendation is the most influencing factor of news credibility in Twitter. So this research contributes a Credibility Network Model that produces credibility ranking values for Sri Lankan news website by considering the factor URL Recommendations in Twitter.

Acknowledgements

This thesis is the result of me being fortunate to have the unconditional assistance of several people who have been extremely supportive in various ways. First and foremost I would like offer my humble gratitude to Dr. A. R. Weerasinghe my supervisor for their tremendous encouragement, support and the guidance given throughout this research.

I would like to sincerely thank all the lecturers at the University Of Colombo School Of Computing for their valuable advices and comments given at various stages of this research. Without your support, I could not have completed this research with success.

I like to thank all my dear friends who were there around me, with best of their encouragement, suggestions and support throughout this research. Finally, I would like to express my heartfelt thanks towards my family for their support and encouragement through the many days and nights dedicated to the completion of this research

Contents

List of Figures

List of Tables

Definitions

Intrinsic Features: Properties that are within the text

Extrinsic Features: Properties that are not essentially within the text but linked with the text

Credibility-Network-Model: A directed network of news websites. And again this model represents the rating values of each website.

Rank/ Ranking and Rate/Rating: The words Rank and Ranking both referred to the meaning of Rate/Rating of a particular news website. Both words indicate the same meaning throughout the thesis.

Alexa Rank of a website: This is a ranking system set by alexa.com (a subsidiary of amazon.com) that basically audits and makes public the frequency of visits to various Web sites. The algorithm according to which Alexa traffic ranking is calculated, is simple. It is based on the amount of traffic recorded from users that have the Alexa toolbar installed over a period of three months.

Chapter 1

Introduction

Today Internet has become one major platform where lots of information travels around the world in considerably less amount of time. At the same time Internet has undertaken many human activities which are performed daily in life. News reading is one such activity which people do in daily basis. Hence over the past few years number of news websites and visitors to news websites have steadily increased. The huge supply of news online is a good indicator of users urge and desire to be informed. Moreover, the availability of multiple sources of news provides new opportunities to different social systems to convey their own opinions in different ways. Eventually, the Internet has become a major platform for millions of users to get involved with latest news stories. Because of this involvement the way of spreading news information has changed. Today we can see that how social media and micro blogging services have made positive and negative effects on the spread of information. When talking about the positive side, this has democratized and accelerated content creation and sharing. On the negative side, it has made people vulnerable to manipulation, as the information in social media is typically not monitored or moderated in any way. People tend to do such manipulations based on different reasons. Thus, it has become increasingly harder to distinguish real news from misinformation, rumors, unvaried, manipulative, and even fraudulent or fake content.

So here comes a new term fake news which is actually introduced by the news

media to illustrate stories on the internet posted by websites with questionable integrity. Or else it can be defined as false information published under the guise of being authentic or factual news. In recent past, there are many events such as elections which came under the influence of fake news stories. Therefore it has become really important to identify credible information content that provides an unbiased narrative of an event. These fake news websites may have several intentions, mainly they try to mislead their consumers through fake news content and motivate them to spread misleading information via social networks or systems. Identifying fake news stories is not a straightforward task. Just by looking at the news content even the professionals may not be able to verify it as fake or not, without a proper evaluation. In recent past, several countries have paid their attention to the problem of misleading information in news sources, social media and microblogging services as well. When addressing the problem of fake news, it is directly related to the credibility of news. News credibility is one of the most important factors in media perceptions. Therefore rating the credibility of online news stories plays an important role in every phase of the information world. The concept of the credibility of news lies on source credibility, content credibility and medium credibility. Source credibility is associated with the credibility of the originator of a particular news story, content credibility can be analyzed using the characteristics of the content and medium credibility involve with the credibility of the channel which the news story travels.

Online news stories can make a major influence on important events such as elections, making awareness of natural disasters, critical social incidents etc. And also online news stories have the dynamism of spreading rapidly than the news in traditional media. So it is important to evaluate the credibility of the information we get all the time. Then only we can address the problem of fake news. This can be achieved through evaluating credibility of news content or news source with suitable measurements and analyze it to identify how much it tends to be misleading or fake.

1.1 Background to the Research

For many years from now, the term credibility, or news credibility, has been an important area of research in persuasion theory. Distinguished source credibility becomes an increasingly important variable to examine within social media and other microblogging services, especially in terms of disaster, danger or risk information. However online information today is suffering lots of criticisms on its credibility. Overtimes, the existence of misleading, biased, falsified information have forced many to question the credibility of online information or online news narratives and the credibility of sources that broadcasts such information. Research in this area focusing on the intrinsic and extrinsic feature of linguistic analysis in order to distinguish the real from the misleading information. The researchers in the past have targeted different domains of news such as political or sports, different mediums such as online news or traditional and different social media platforms such as Twitter or Facebook when determining the credibility of news sources and news content.

In recent past, researchers have categorized four major categories of misleading news which also can be called as fraudulent or fake news. The first category is fake, false, or regularly misleading websites that are shared on social systems such as Facebook, Twitter etc. Some of these misleading websites may rely on outrage by using distorted headlines and decontextualized or suspicious information in order to generate likes, shares, and profits. There can be numerous reasons for developing this type of websites, for example, the reason may be financial benefits or political benefits.

The second category is websites that may circulate misleading and/or potentially unreliable information. This category is different from the first one because these websites are seems to be stable and reliable. Most of the politically unreliable news websites fall into this category. The stableness of this type of websites makes it hard to determine whether the content of the website is false or not. Therefore the intrinsic analysis is not enough to identify the website and extrinsic features

are mostly targeted here.

The third category is Websites which sometimes use click-bait headlines and social media descriptions. The main purpose of click baits is to attract attention and encourage visitors to click on a link to a particular web page. This is another way of spreading fake news stories on the web. We will be focusing more on this under the literature review.

The fourth one is satire/comedy sites, which can offer important critical commentary on politics and society but have the potential to be shared as actual/literal news. Satire is an attractive subject in deception detection research. It is a type of deception that intentionally incorporates cues unveiling its own deceptiveness. Whereas other types of fabrications aim to instil a false sense of truth in the reader, a successful satirical hoax must eventually be exposed as a joke. We can see most people get offended by this kind of news stories. Research in this category follows intrinsic methodologies to distinguish satire news stories.

Even though there are four categories, some articles fall under more than one category. Assessing the quality of the content is crucial to understanding whether what you are viewing is true or not. For example, false or misleading medical news may be entirely fabricated (Category 1), may intentionally misinterpret facts or misrepresent data (Category 2), may be accurate or partially accurate but use an alarmist title to get your attention (Category 3) or maybe a critique of modern medical practice (Category 4).

So the whole four categories of misleading news information come under the research area of credibility of online news stories. The whole research area of credibility of news online or the fake news detection falls under two categories such as linguistic feature based credibility analysis and network feature based credibility analysis. All four categories of misleading news which are described above can be analyzed solely based one of the two methods or combining those two methods.

Apart from them, there is logic programming based approaches to rate the credibility of news stories which can be found in very recent researches.

Evidently, literature review performed (which will be discussed later) seems to suggest that credibility is key in news. That means how important to identify credible news stories. Today, instead of just being a passive recipient of messages, readers or audiences are able to select their news channels. When the public (news recipient) considers a medium to be more credible than other media, they are also more likely to rely on that medium for information search and sharing than other media. Therefore the importance of analyzing the credibility of news providers or the news sources has also been discussed in this research area.

1.2 Research Problem and Research Questions

1.2.1 Research Problem

Today online news stories play an important role in the information world. Hence the problem of measuring the credibility of information emerged. Because of the ultimate technological freedom, some people/websites tend to make misleading information for different purposes. Today these misleading information on several important events which can be called as fake news being used to influence elections and many other events. Clearly, the problem is how to measure the credibility of online news stories. Determining the credibility of online news stories is a vast and time-consuming problem. So the problem can deviate into phases from the origination to the distribution of a news story. Therefore the first phase is to determine the credibility of news originator or the news provider. It will lead the way to determine the credibility of the news story. So in order to face the problem of fake news determining the credibility of the news source is important. Typically the misleading or fake news stories travel through different news sources. Those news websites/sources can be stable or completely fake. Most misleading news sources tend to share their information through social media hence their consumers will share that information without having a proper idea of its credibility. Hence

those misleading information can influence the world in different ways. Therefore it is important to have an understanding of most credible news sources online.

1.2.2 Research Questions

1) How to evaluate the credibility of online news sites/sources?

This question addresses the problem of having misleading news websites which continuously provide misleading news stories. Hence their consumers will be get excited and persuaded to share those news stories on social media sites. Social media is a platform which has an excessive power to spread something rapidly. That news story may have travelled so far when eventually get to know that is a fake or misleading news story. Hence there should be a proper method to evaluate the credibility of a news website, which misleading news stories are published, by its features. This research question determined to find a proper method to evaluate the credibility of a news website/source.

2) How to evaluate the credibility of a news story?

This question will produce a better analysis of the solution we provide for above question. A particular news story may have been shared on several websites. Even though the news story can be misleading or true. So by this question, we are addressing how to evaluate the credibility of a particular news story which has been published on several websites. That is by considering the credibility rating values of each website that the news has been published on. The website which a news story is been published may be completely fake or a stable one. Therefore if we can measure the credibility of a news website, we can decide the credibility of the news content of that particular website. So this problem addresses determining the credibility of a new story based on the news publisher. The solution for this question is solely based the solution for the previous question

1.3 Justification for the research

So far the research in this area falls into a combination of natural language processing and machine learning based studies theoretically. There are several classification based approaches to analyze the credibility of news stories in different categories such as satire, click-baits etc. But most of them represent one single domain. These classification based approaches mostly fall into the linguistic category (this will be discussed more in the literature review) where intrinsic features are analyzed predominantly. But the problem is by only analyzing intrinsic features it is hard to predict credibility measurements. At the same time analyzing the credibility of news is time-consuming and dynamically changes time to time. Therefore it is important to always analyze current data. Hence network-based approaches give more capacity to get a broader view of credibility. That is by considering the behavior of social networks and other knowledge-based hierarchical models. Even though there are network-based approaches most of them are focusing on determining the credibility of the news event. However, analyzing the credibility of the news source or the news publisher is an important area which is not addressed considerably well. So this research is focusing on analyzing the credibility of a news source and build a model that rates news websites based on its credibility. Based on that credibility of the news story is determined. But again it is hard to do such network-based analysis. Because in order to do that a massive process of data collection and time is needed. Hence in this research Sri Lankan news environment is being considered. And there is no prior knowledge or analysis exists related to Sri Lankan context.

1.4 Methodology

Generally, this research is done with the intention of finding a promising approach to credibility evaluation of online news stories. Following approaches are the methods that are going to be used in the research.

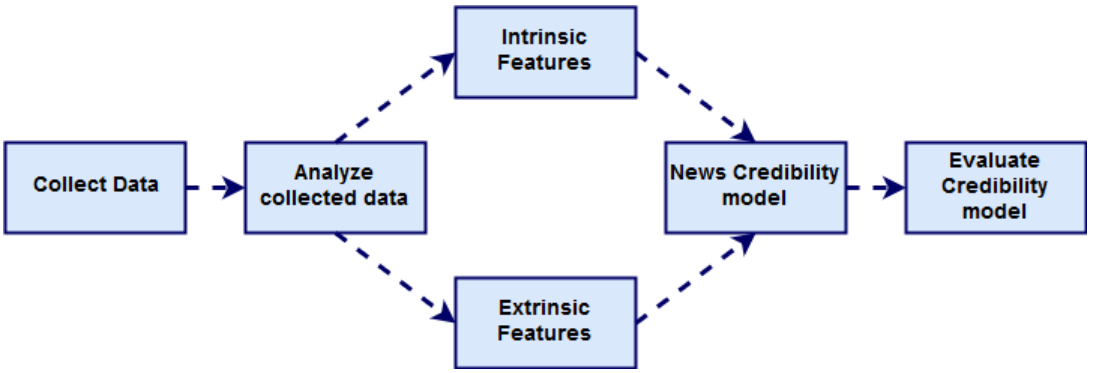

FIGURE 1.1: Research Methodology

1) Generate a suitable dataset using existing data.

In In this phase, a suitable dataset for the research will be generated. There is no available dataset for Sri Lankan websites with their features. So those data will be collected. Part of the data includes social media behavior of a particular website. Selected social media platform for this research is Twitter. In order collect those twitter data, Twitter APIs will be used.

2) Analysis of collected data

Analysis will be conducted in the following for the collected dataset,

- Analysis of behavior of features related to websites
- Analyze similarities and dissimilarities of feature
- Review credibility evaluating techniques available in the literature
- Explore existing extrinsic feature based techniques.
- Explore existing intrinsic feature based techniques.
- Examine and review most appropriate credibility evaluating techniques applied by previous researchers.

3) Design a method to determine the credibility of online news stories and news websites and implement a prototype

- Design most appropriate research to evaluate the credibility of news websites by considering the analysis of the data.
- Design the credibility news website network based on twitter data
- Design the evaluator of news stories based on credibility news website network

4) Evaluate the Implemented model with the association of field expert

- In this phase, the implemented credibility evaluation model will be evaluated for its accuracy with the participation of some field expert who has some prominent knowledge in news credibility evaluating
- Also evaluated with human adaptability of this method to identify most credible news over the online environment.

1.5 Outline of the Dissertation

Today the internet has become a major platform for millions of users to get involved with latest news stories. Even the way of information spreading has changed because of the existence of social media and microblogging services. At the same these time social media platforms and news websites have the power of spreading rapidly than the traditional media. On the positive side, this has democratized and accelerated content creation and sharing. On the negative side, it has made people vulnerable to manipulation, as the information in social media is typically not monitored or moderated in any way. Thus, it has become increasingly harder to distinguish real news from misinformation, rumors, unvaried, manipulative, and even fake content. Therefore there should be a proper mechanism to analyze the credibility of the news information providers or the publisher. So analyzing the credibility of the news source, news content and the news medium is important in order to achieve online news credibility.

There are identifiable four categories exists for misleading news stories based on the news source and the news content. A particular misleading news event can be a combination of above mentioned four categories. Research in this area can be mainly categorized into two as linguistic-based approaches and network-based approaches. Linguistic-based approaches are based on analyzing intrinsic features of news content and network-based approaches are based on analyzing extrinsic features of news websites and news content. Apart from them, there are methods that combine these two approaches. At the same time, there are few logic programming approaches as well.

This research follows a methodology which consists of different phases. There are basically four main phases. First phase is the data collecting phase. In order to have a fine dataset, data related to Srilankan news websites will be collected throughout this phase. A part of the data collection holds social network behavior of SriLankan news stories. The second phase is the analysis of collected data. In this phase, the most prominent feature will be extracted in order to identify credibility measurements. As the third phase, a model is generated in order to predict credibility ratings of each news website. The final phase is the evaluation of generated model. Evaluation of the model is conducted based on the feedbacks of experienced and prominent journalists in Sri Lanka.

1.6 Delimitations of Scope

The dataset is generated based on Sri Lankan news websites and the behavior of Sri Lankan news on social media. Basically the social media network that will be considering is Twitter. In order to collect Twitter data, Twitter APIs will be used. So the credibility-network-model of news websites is built based on above mentioned data.
The considered language will be English.

1.7 Summary

So far the in this chapter, the importance of analyzing the credibility of news stories and news sources is being discussed. Though that the potential harm of fake or misleading new stories is identified. By considering the area of this research, It can be seen that the network approaches can yield broader picture on online news credibility. So based on that this research is focusing on building a credibility-network-model of news websites in order to get credibility ranking values of each news website.

As described in the methodology, the first phase will generate a full-featured dataset on Sri Lankan news websites with including social media data. Twitter is the considered social media platform since according to the literature 85% of the Tweets is about news events. So in order to collect Twitter data, Twitter APIs will be used. So this will fulfills the unavailability of a dataset related to Sri Lankan news websites which can be efficiently used to build more credible environment for Sri Lankan people.

The main objective of this research is to build the credibility network model of Sri Lankan news websites. Basically, this network can identify most credible news websites because websites with more credible features are given higher ranks. So by looking at this credibility network human can decide which websites are more credible and which websites tend to be misleading. Like explained above this method is fallen into the category of network approaches. So the final aim is to build a credibility network model which can be analyzed by human intelligence.

Chapter 2

Literature Review

The Internet evidently, has become one of the influential news sources. The literature reports that over 3 billion people around the world now use the internet via various devices [25]. In addition, news websites have become more not only profitable but also effective, and media organizations tend to invest in online journalism. Consequently, the number of the news website is rapidly increasing today. At the same time number of visits for those websites steadily increasing. The Newspaper Association of American (NAA, 2006) reports that 112 million people visited online news sites during the first quarter of 2006. Nearly one-quarter (24%) of Americans say the Internet is their main source of news, while 44% obtain news from online sources at least once a week (Pew Research Center, 2005). Hence the question has been raised on the online news credibility hereafter. Today most of the countries have paid their attention towards this problem. For example, the USA has put a lot more attention into fake news during their presidential election 2016 [30]. They discuss the bad consequences towards the election made by fake news stories. Furthermore, discuss the involvement of social media with fake news and the potential harm that can make to important events. So currently this problem of fake news has become a threat. And according to their sources, there are currently more than 200 news websites which have been identified as fake news websites. All these fake news websites were engaged in providing fraudulent information to their consumers. Some of these websites were duplicating the names of prominent news websites. Hence identifying source credibility emerged as an

important problem to all.

Apart from new websites itself, in the present, there are several commercial multi-source news providers on the Web, such as Google News (http://news.google.com/), Yahoo! News (http://news.yahoo.com/.), etc. They have their own mechanism in order to select or filter news from the web. Although none of them has unveiled the technical details underlying the way news stories or news events are selected, conglomerated and ranked. Although it is evident that factors such as freshness, sources and popularity measures are taken into consideration. They use special procedure to measure popularity of a particular source in the web. It can be changed in time to time because popularity is dependent upon time. The information provided in news reports may not always be fully verifiable and therefore another important factor that can help select news is trust or credibility. Since that commercial multi-source news providers may not be fully correct all the time because factors like popularity may not always represent credible news. Hence there should be a proper mechanism for determining the credibility of a news story. But it is harder to measure the credibility than measuring the freshness or the popularity. Research on the multi-source news has generally overlooked the dynamics of news credibility. Mostly credibility of news has been studied through quantitative approaches (e.g. [25]). There is also documented evidence [27] of Google News proposals to build a database of news source credibility based on information such as average story length, number of staff a news source employs, the volume of internet traffic to its website and the number of countries accessing the site.

When comes to the theoretical background of this research area, there are two major categories that can be identified. They are Linguistic Approaches and Network-based Approaches. Linguistic Approaches in which the content of misleading information is extracted and analyzed to associate language patterns with deception or fraud and Network-Based Approaches in which network information, such as message metadata or structured knowledge network queries can be

harnessed to provide aggregate deception measures[1]. Both forms typically incorporate machine learning techniques for training classifiers to suit the analysis. It is solely incumbent upon researchers to understand these different areas. A considerable amount of literature has been published on both linguistic approaches and network-based approaches. In recent years, researchers have investigated further on combining these two main approaches together to form different models to evaluate the credibility of news stories. Apart from that there are ranking methods for news stories build upon logic programming methods. But those logic programming models are recently emerged and there are very few. So the following sections are elaborating these approaches separately in order to get a better understanding of how these approaches are actually administered.

2.1 Linguistic Approaches

In recent years, researchers have investigated a variety of approaches based on linguistic features. These linguistic feature-based approaches are applied on different news domains. Basically, they have used style based techniques which relies on computational linguistics and natural language processing. Furthermore, deception detection methods are applied to identify statements at the sentence-level that constitute prevarications and lies[6]. Rubin[10] contributed the first actual attempt at fake news detection by separating satire news as a representative of humorous fakes from real news. In his research investigation, a dataset of 180 news articles was analysed. This is basically style-based approach falls into the linguistic category.

However Linguistic feature based research area is a huge research area where many distinguishable categories can be identified when comes credibility analysis of news or fake news detection. So focusing more on linguistic approaches mainly following categories can be identified. They are Deep Syntax, Rhetorical Structure and Discourse Analysis, Classifiers, Semantic Analysis and Data Representation. When

considering each of this category, we can see that they have used linguistic features in an appropriate way to determine the credibility or detect false information.

As mentioned above the first linguistic approach is the Deep syntax approach. According to deep Syntax, Analysis of word use is often not enough in predicting deception or identifying false information. So in this approach, deeper language structures (syntax) have been analyzed to predict instances of deception. Deep syntax analysis is implemented through Probability Context Free Grammars (PCFG). Sentences are transformed to a set of rewrite rules (which can be called as a parse tree) to describe syntax structure, for example noun and verb phrases, which are in turn rewritten by their syntactic constituent parts [11]. The final set of rewrites produces a parse tree with a certain probability assigned. This method is used to discover rule categories (lexicalized, unlexicalized, parent nodes, etc.) for deception detection with 85-91% accuracy (depending on the rule category used) [11].

Rhetorical Structure and Discourse Analysis is another linguistic approach that identifies instances of rhetoric relations between linguistic elements. Systematic differences between deceptive and truthful or credible messages in terms of their coherence and structure has been combined with a Vector Space Model (VSM) that assesses each messages position in multi-dimensional RST space with respect to its distance to truth and deceptive centers [12]. At this level of linguistic analysis, the prominent use of specific rhetorical relations can be suggestive of deception. Regarding fake news detection, Chen [13] point out the need for an automatic crap detector for news, but there is no report of his actual experiments regarding this case. But according to the literature Rubin [12] apply, for the first time, deception detection approaches to fake news detection using rhetorical structure theory as a measure of story coherence.

In recent past, there have been many classification-based, semantic-based research methodologies on credibility analysis of the news. But they certainly focusing on

one particular domain like politics, sports or a category of news such as satire, click-baits etc. So focusing on related work on these categories, following models can be identified as outcomes of previous research work. Ivan Koychev, Preslav Nakov [7] proposed a language-independent approach for automatically distinguishing credible from fake news, based on a rich feature set. In particular, they use linguistic (n-gram), credibility-related features (capitalization, punctuation, pronoun use, sentiment polarity), and semantic (embeddings and DB-Pedia data) features. This research is purely linguistic. But again there could be limitations because achieving language independence is not an easy task and the model may depend on the alphabet of the language.

Yimin Chen,Niall J.[8] proposed a potential method for the automatic detection of click-bait as a form of deception. In their work, they propose methods for recognizing both textual and non-textual click-baiting cues. Click-bait is another way of spreading misleading news stories. We can recognize some part of click-baits are intentionally made for spread rumors. Click-bait refers to content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page. Click-bait has been implicated in the rapid spread of rumor and misinformation online.

So they are solely linguistic feature-based approaches which can be identified from the current work of this research area. Apart from deep syntax approaches, other approaches mostly depend on the news domain, which is being considered in the research, for example political, sports etc. So the generated model cannot be generalized into different domains. When considering applying deep syntax and Rhetorical Structure and Discourse Analysis into fake news detection; they are hard and time consuming than other approaches. Even though these solely linguistic approaches are good at identifying false information that may not be enough in order to identify fake news stories. Because today fake news providers or websites are ambitious on producing fake news stories so it is hard to distinguish fake from real only based on linguistic features. And existence of social networking platform

make the situation even more complex. So Network Based approaches are more appropriate for analyzing source credibility. As described earlier, analyzing the credibility of the source is important to measure the credibility of news stories which are published on that particular website.

2.2 Network Based Approaches

As mentioned above the second category of news credibility analysis is Network-based approaches. In this category, extrinsic features are analyzed in order to build credibility measurements. Basically, network approaches are innovative and varied because they examine network properties and behaviors. So this network behavior or the structure mostly can be seen on social media platforms such as twitter, facebook and other micro-blogging services. As real-time content on current events is increasingly proliferated through micro-blogging applications such as Twitter. According to research in this category, there are several investigations that analyze network properties of micro-blogging services in order to determine how news stories are spread. So based on those studies, models are built for analysing news credibility online.

By considering the mechanisms of social networks, new angles on the problem of fake news propagation come into reach. Context-based fake news detection again belongs to this category of research. Acemoglu, Asuman [18] model how false information is spread in social networks. This is one important attempt at credibility analysis based on social networks. This is not directly on news but their studies is a good example of the spreading behavior of false information in social networks.

A considerable amount of literature has been published on network-based approaches recently. When focusing more on those investigations, several models and methods that analyze the credibility of online information can be found in the literature. Tambuscio [15] study the spread of misinformation in social media; however, they also study the efficacy of countermeasures such as debunking sites.

In particular, they find that by exceeding a certain threshold in spreading the refutation is sufficient to remove the misinformation from the network, and that this threshold does not depend on the spreading rate but on credulity and forgetfulness.

When comes to the social media platform, Twitter is being targeted in several times in order to identify credible information over misleading information. Social media platforms like Twitter facilitates real-time propagation of information to a large group of users. This makes it an ideal environment for the dissemination of breaking-news directly from the news source and/or geographical location of events[5]. Kwak, Ponnurangam [20] discovered that 85% of discussion topics on Twitter are related to news. Hence it derives twitter is the most suitable platform to do a news credibility analysis. So It can be seen that Twitter is the most suitable social media platform to analyze the behavior of news stories.

Zareen Sharf, Anwar[22] produce an application called Twitter news credibility meter which is based on a network analysis of Twitter. That system proposed for the credibility assessment of a potential news tweet.It is heavily inspired by early knowledge models. The developed system validates credibility based on Number of re-tweets received in a specified interval of time, The geographical location of the users tweeting and the event are identical, The tweets contain reference links or URLs for more information regarding the event.

Zhiwei Jin, Juan Cao [21] propose to exploit the conflicting viewpoints in microblogs to detect relations among news tweets and construct a credibility network of tweets with these relations. First, conflicting viewpoints are mined through a topic model method. Then they construct a credibility network by linking tweets with detected relations to evaluate them as a whole. Their model is a good example of network approach on Twitter. Here the accuracy of the minded viewpoints directly affects the accuracy of the resulted network.

The Knowledge-based network is one prominent approach in Network-based approaches. The use of knowledge networks may represent a significant step towards scalable computational fact-checking methods. Knowledge-based fake news detection (also called fact checking), is tackled with methods borrowed from information retrieval, semantic web, and linked open data (LOD) research[14]. For example, Etzioni, Banko [14] propose to use their well-known tool Text Runner proposed by Yates, Banko[16] to extract and index accurate or factual knowledge from the web, and to use the same technology to extract factual statements from a given text in question, matching them against the indexed facts to distinguish inconsistencies. Magdy, Wanas [15] develop a statistical model to check factual statements extracted from a given document in question, analyzing how frequently they are supported by documents retrieved from the web. Both approaches presume that web resources (or the frequency by which a fact is mentioned) can be used as an indication of its truth.

In recent past, Martin Potthast, Johannes Kiesel [2] have conducted their research on a writing style analysis of hyperpartisan news in connection to fake news. They proposed and demonstrated a new way of assessing style similarity between text categories via Unmasking, revealing that the style of left-wing and right-wing news have a lot more in common than any of the two have with the mainstream. Basically, their intention was to determine how left-wing or right-wing news can be fallen into the category of fake news. They have used existing knowledge in order to determine whether the news events are true or false. Furthermore, they show style-based fake news detection does not live up to scratch. Their approach can be recognized as an approach which comes under knowledge-based fake news detection. This research is targeting the behaviour of fake news on Facebook.

There are several network effect variables that are exploited to derive truth probabilities [17], so the outlook for exploiting structured data repositories for fact-checking remains encouraging. From the short list of existing published work in this area, results using sample facts from four different subject areas range from

61% to 95%. Success was measured based on whether the machine was able to assign higher true values to true statements than to false ones [17]. A problem with this method, however, rests in the fact that statements must reside in a pre-existing knowledge base. Therefore the accuracy of pre-existing facts of knowledge base influences on the fact-checking model.

2.3 Mixed Methods

2.3.1 Combination of Both Network and Linguistic Approaches

In recently researchers have focused on combining both linguistic and network features into their research. Basically, they have used network methods to link data elements and also by using linguistic methods they have analyzed the structure of the language elements. Some of those studies have shown more prominent result on identifying fake news stories. Carlos Castillo, Marcelo Mendoza [5] propose automatic methods for assessing the credibility of a given set of tweets. Specifically, they analyze micro blog postings related to trending topics, and classify them as credible or not credible, based on features extracted from them. They use features from users posting and re-posting (re-tweeting) behavior, from the text of the posts, and from citations to external sources. They have basically focused on the newsworthy topics on Twitter. Their investigations are based comparisons of various classification algorithms and interesting features for the task. Features are extracted from four aspects: the message, user, topic, and propagation features. Therefore this research is based on both linguistic and network features and provides better insight into the problem of identifying credible information on Twitter.

2.4 Approaches on Hyperlink Behavior of Websites

Apart from network-based studies on social networking platforms, there exist investigations based network approaches which analyze hyperlinks behavior of websites in order to evaluate the credibility of a set of websites. Porismita Borah [9] attempts to understand first, how hyperlinks can influence individuals perceptions of news credibility and information-seeking behavior. Second, the paper extends previous research by examining the interaction of hyperlinks with the content of the story. In doing Furthermore this research examines the influence of hyperlinks on news frames.

2.5 Logic Programming Approaches

This is a different research approach for the problem of determining news credibility. And this attempt is based on a Logic programing approach for rating the credibility of news stories. Gabriela, Carlos, Ricardo[23] proposed a qualitative and personalized trust-based news service which allows news viewers to access and compare the trustworthiness of news sources and their reports. Viewers trust statements on sources and reports can be based on the viewers subjective beliefs or, when absent, trust assumptions can be obtained indirectly from other viewers beliefs. However, in order to derive trust from other viewers, a trust relationship between viewers must exist. In this proposal, trust is modelled using DeLP, a defeasible argumentation framework based on logic programming [24]. This allows us to integrate argumentative reasoning into a news service, which will provide a reasoned basis for the news presented to the user.

In that way, all those approaches from the past can be categorized briefly. All of them together present different aspects we can consider in order to measure credibility of news.

2.6 Twitter as a News media

Twitter, apart from a social media platform, has been studied broadly from a media perspective as a news dissemination mechanism. And that is for both regular or periodic news and for emergency situations such as natural disasters, and other high impact situations [5][20][29][36][37]. For example, Thomson [29] models the credibility of various tweet sources throughout the Fukushima Daiichi nuclear disaster in Japan. According to their studies proximity to the crisis seemed to moderate an increased tendency to share information from highly credible sources. There are several social media platforms. So why select Twitter as a news media? Actually this question is answered in many times in the past by several investigations. Kwak, Ponnurangam [20] discovered that 85% of discussion topics on Twitter are related to news. Twitter acts not only as a social network, but as a news source [29]. Notifying oneself regarding breaking news is a common motivation for examining public tweets [31]. For example when seeking updates about local emergencies [32] mostly public tweets are examined or searched. So Twitter is now used to disseminate substantive content such as breaking news. Unfortunately, there are not only of breaking news but also undesirable memes such as spam [33] and rumors [34].

Therefore it increases the importance of evaluating the credibility of tweets. So these studies show that Twitter is not less than a news media for millions of Twitter users.

2.6.1 User Influence In Twitter

Direct links in social media can represent any relationship to common interest or passion and interest towards a breaking news or even celebrity gossip. Such directed links determine the flow of information and hence indicate a users influence on others. Therefore when news stories are propagating on Twitter, users are the ones who influence those news stories to be an interest of other users. According to the literature, there are three measures of influence: indegree, retweets, and mentions. Meeyoung, Hamed[28] studied on how different types of influentials interact with their audience. according to their investigation, influential users can

hold significant influence over a variety of topics. Furthermore, they say that mainstream news organizations consistently spawned a high level of retweets over diverse topics. And the name value of the mention helped them get responses from others.

The proposed method for this research consider these influential factors (indegree, retweets and mentions) of Twitter to determine the credibility of news stories. Because when propagating news stories in twitter role of the user gives good insight into determining the credibility.

According to the studies of Jilin, Rowan[29], URL of the content recommendation on Twitter is a better way of getting users attention. In a modular approach, they have explored three separate dimensions in designing such a recommender: content sources, topic interest models for users, and social voting. So it is an example of examining the recommendation mechanism of Twitter. According to their studies URL recommendation is a useful way of streaming information. Therefore in this research, we are focusing on URL recommendations on news stories by different users. So in our studies, we propose and examine how this recommendation mechanism can be applied to determine the credibility of news websites.

2.7 PageRank Algorithm

Since this research is focusing on rating/ranking a news website based on its credibility, there should be a proper algorithm to achieve that. In order to determine the importance of web pages, Brin and Page [39, 40] proposed a ranking algorithm, called PageRank, which computes a ranking for every Web page based on the graph of the Web. However, PageRank algorithm is the powerful weapon behind the success of Google Internet search engine market. However, the authors have focused more on providing quality search outcomes efficiently [39]. This method brings the most appropriate search results to the top since it takes the assumption that web hyperlinks as the important votes and ranks search results based on links

interlinking them [38]. Thus, PageRank has created an innovative way of ranking the web pages. It is very simple, content-free, and scalable.

2.8 Summary

This chapter discusses different approaches present in this research area. Basically area of this research can be divided into two as Linguistic and Network. Furthermore, this chapter discusses existing approaches and models for evaluating credibility based on Linguistic and Network approaches. Apart from that, methods that combine both the linguistic and network approaches are discussed. And again discusses how Twitter can be considered as news media and user influencing factors on Twitter. Then finally discusses the importance of page ranking algorithm. So based on the literature, our proposed approach suggests a network-based approach considering user influence of Twitter to build a fine model for rating the credibility of online news stories. In order to develop a proper and appropriate rating mechanism, the page ranking algorithm is used. Later sections elaborate how these approaches are used to design and implement the Credibility Network Model.

Chapter 3

Design

Proposed Architecture

This chapter outlines the overall design of the proposed solution. As the investigation done under the previous chapter, It is been clarified that Network Based Approach is more appropriate for this research. Therefore a Network-based approach is being used to generate the Credibility Network Model for news websites in Sri Lanka. The intention here is to provide a solid design for a model which can predict credibility of Sri Lankan news websites by considering online news propagating environment.

3.1 Design Approach

The architecture of this research is designed based on a Network-based approach. Based on the literature it is determined that Network-based approach is most appropriate to conduct the research. Therefore the network behavior of Srilankan news websites is analyzed in this research. According to the literature it says that 85% of the Twitter is about news. So Twitter is the most suitable social media platform to analyze news events. Therefore Twitter has been selected as the social media platform to analyze news events based on the literature. When considering the user influence in Twitter, there are three influential factors as In-degree, Mentions and URL Recommendations. So in this research, all the three factors are analyzed with the perspective of online news credibility. Therefore three

models based on In-Degree, Mentions and URL Recommendations are designed in this chapter. A page ranking algorithm is used in order to determine the credibility ranking values for news websites.

3.2 System Architecture

This section is devoted to present the architectural and functional design of our Credibility Network Model. The first phase is the Data collection. Then from the collected data, the connecting environment of news websites with other users is observed. As discussed in the literature there are influential factors such as In-degree, mentions [28] that influence the linking environment in Twitter. Therefore those factors were considered when building connections among news websites and other users. So the linking environment is extracted based on collected data. And as explained in the literature, URL recommendations[29] are again influencing the connecting environment of websites. Hence these factors are considered separately in three different models that represent,

- In-degree,
- Mentions and
- URL recommendations.

All the three models generate credibility rankings based on the same rating algorithm. So finally the most accurate model is selected based on the evaluation results. Evaluation phase is conducted with some prominent journalists in Sri Lanka.

The proposed architecture of our experiment is depicted in Figure 3.1. We are following three approaches based on main three factors: In-degree, Mentions, URL Recommendations.

Therefore the model 01 represents the approach based on "In-degree", that is in this model, connecting (linking) environment of news websites are determined

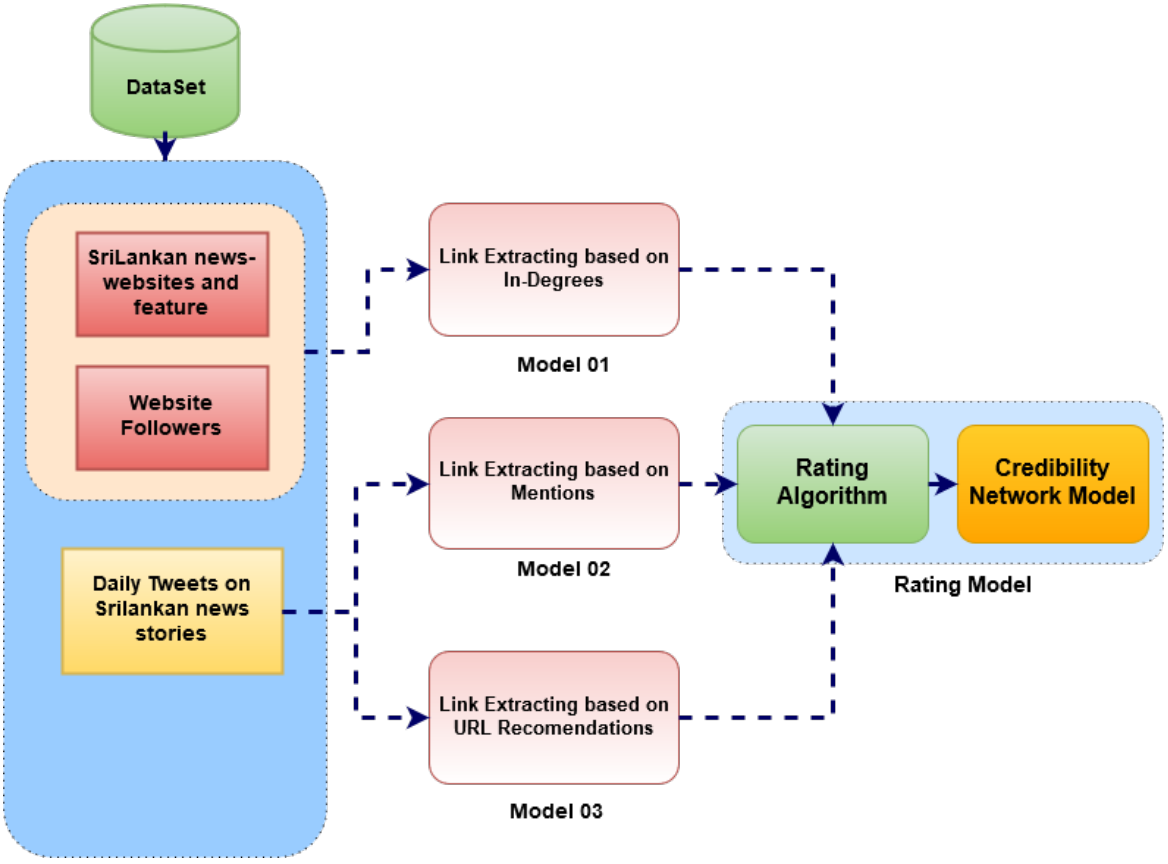

FIGURE 3.1: Design of credibility rating model

based on the in-degrees. The model 02 represents the approach based on "Mentions", that is, the connecting environment of the news websites are determined based on mentions.

And model 03 represents the approach based on "URL Recommendations", that is, the connecting environment of news websites are determined based on URL Recommendations.

So following sections describe the design of each model. Though the method of extracting links is different in each model, the rating algorithm used, to develop the credibility ranking values, is same for all three models. So the proposed architecture of each model comprises with three main modules; Data Collecting Module (Dataset), Link Extracting module and Rating Module. Most accurate model for rating the credibility of online news stories is chosen based on the evaluation results.

3.2.1 Model 01 (In-Degree)

This model consists of three main parts. They are data collecting module, Link extracting module and Rating model. The Figure 3.2 depicts the design of Model 01.

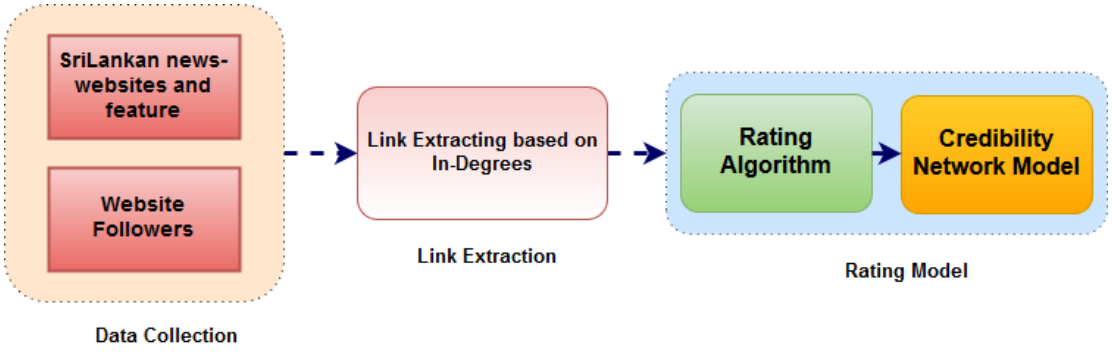

FIGURE 3.2: Design of Model 01

3.2.1.1 Data Collection

Data Collection phase consists of 2 main parts. First, a dataset consist of Sri Lankan news websites and features of those websites is generated. So that dataset contains 85 Sri Lankan news websites and features. As the second part, Twitter data related to each Sri Lankan news website is collected.

Data Collection Part 01

Collection of 85 Sri Lankan news websites and features. The considered features are as following.

- Alexa Rank of the website
- Google plus connections per a website
- Number of tweets a particular website has posted in its twitter account
- Number of twitter accounts a news website follows
- Number of followers per a news website

Data Collection Part 02

Collection of details of 500 followers of each news website separately.

3.2.1.1.1 Data Collecting Methods

1. Sri Lankan News Websites: Reliable sources of Sri Lanka
2. Twitter Data: Twitter Search API, Twitter Tread API, Twitter Steaming API with python

3.2.1.2 Link Extracting Module (Based on In-degrees)

As described earlier, in this model, the link extracting approach is based on In-degrees. So basically the total number of In-degree of a particular news website is the total number of followers of that website. Therefore having a large number of followers means the in-degree of that website is high. If we consider the followers as consumers of news websites and two websites have the same set of consumers, which means those two websites have a common set of properties. So in this approach websites having a common set of followers(consumers) are linked together. The links are extracted based on the implementation of the algorithm 01.
So this matrix M is created considering all the edges between USERs and MENTIONs. That matrix is the input for the Rating Model.

Algorithm 1 Link Extraction Algorithm

```
1: W ← All websites
2: T ← All threshold
3: for A in W do
4:     for B in W do
5:         Common_Followers_set AB = Followers(A) ∩ Followers(B).
6:         if sizeof[Common_Followers_set AB] > T then
7:             Link Aand B.
8:
9:             Where,
10:            A and B are news website "Common Followers A B" is a Set
11:            containing all the common user that follows both
12:            the websites A and B.
13:
14:            If the numbers of the common followers is greater
15:            than the selected Threshold(T), link the website A and B.
16:
17:            Here T is selected based on a median values which represented as
18:                Size[Common P Q] < ... < size[CommonXY]
19:                 < .... < size[CommonRS] P,Q,R,S are websites and
20:            Median = size [Common X Y] = T
21:
22:            Generate W*W Matrix M
23:
24:            for A in W do
25:                for B in W do
26:                    if A and B are linked then
27:                        mark the position (A,B) of M as 1
28:                    else
29:                        mark the position (A,B) as 0
30:            return M
```

Another point about this network is this will contain only the Sri Lankan news websites. As described earlier, Rating model is same for all three models. The rating algorithm is described in section 3.2.4

3.2.2 Model 02 (Mentions)

This model consists of three main parts. They are data collecting module, Link extracting module and Rating model. The Figure 3.3 depicts the design of Model 02.

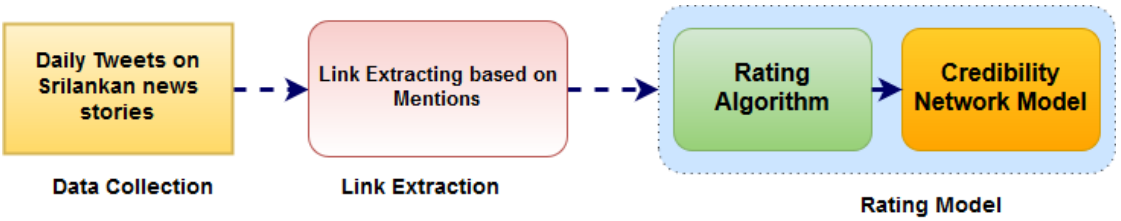

FIGURE 3.3: Design of Model 02

#SriLanka #srilanka #Srilanka #lk #lka were filtered daily. Tweets with those tags are considered as news related tweets

3.2.2.1 Link Extracting Module (Based on Mentions)

In this approach link among Twitter users and news websites are extracted based on mentions. For example if a user A has posted a tweet having @B in the tweet body, then A has a directed link towards B. Algorithm 02 is designed to extract those links from collected data.

Algorithm 2 Mentions based link extraction algorithm

```
for each Tweet do
    Extract username of the USER who post the tweet
    Extract all the "MENTIONS" in the tweet body
    Link Twitter USER with "MENTIONS"

W ← All USERs + all "MENTIONS"

Generate W*W Matrix M
for A in W do
    for B in W do
        if A and B are linked then
            mark the position (A,B) of M as 1
        else
            mark the position (A,B) as 0
return M
```

So this matrix M is created considering all the edges between USERs and MENTIONs. That matrix is the input for the Rating Model.

Another point about this network is this will contain not only the news websites but also the users who post tweets about Sri Lankan news. As described earlier,

Rating model is same for all three models. The rating algorithm is described in section 3.2.4

3.2.3 Model 03 (URL-Recommendations)

This model consists of three main parts. They are data collecting module, Link extracting module and Rating model. The Figure 3.4 depicts the design of Model 03.

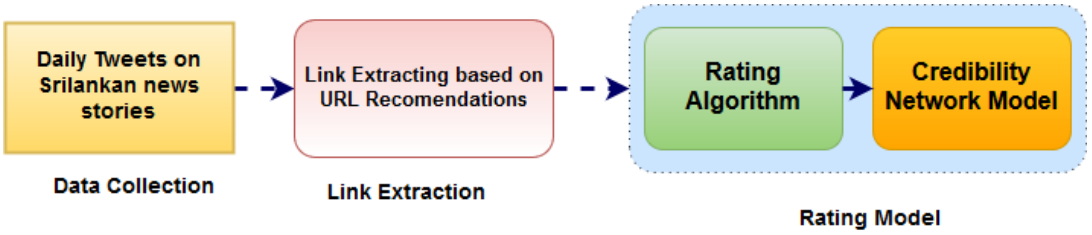

FIGURE 3.4: Design of Model 03

3.2.3.1 Data Collection

Data collection is same as the model 02. But here data is filtered in order to extract URLs from Tweets. Like in the previous model (model 02), Tweets containing tags of #SriLanka #srilanka #Srilanka #lk #lka were filtered daily. Tweets with those tags are considered as news related tweets.

3.2.3.2 Link Extracting Module (Based on URL Recommendation)

In this approach links between users and news websites are extracted based on users who are posting URLs related to news stories. For example, if the user A posting a URL such as www.news.lk, user A is linked with the domain name news of that posted URL. So the following algorithms explain the whole process of extracting connections between users and URLs.

Algorithm 03 is about extracting connections between USERs and URLs per a day. For each day a set of USER,URL edges are created according to this algorithm. Algorithm 04 is for a particular period of time. After generating (USER, URL) set for each day in a particular period (work done by algorithm 03), the common

Algorithm 3 Extracting URL with USER

```
Filter Tweets with URLs

for each Tweet with URLs do
    Extract username of the USER who post the tweet
    Unwrap the URL and get the Domain name of the URL
    Link the domain name of URL with the Twitter USER

Create (USER, URL) Set for each Day
```

set of (USER, URL) is extracted based on the algorithm 04. The purpose of doing this is to identify the frequent set of users who update tweets frequently on news.

Algorithm 4 Dataset creation algorithm with USER, URLs

```
Take the sets of (USER,URLs) in a certain time period
for each day do
    Set a Threshold for "Minimum number" of Tweets per Day by a USER

if Number_Of_Tweets_by_USER(A) > Threshold then
    Extract USER(A) with URLs_PostedBy(A)
for each Week do
    Extract Common USERs and linked URLs

Generate W*W Matrix M

for A in W do
    for B in W do
        if A and B are linked then
            mark the position (A,B) of M as 1
        else
            mark the position (A,B) as 0
return M
```

So this matrix M is created considering all the edges between USERs and domains name of URLs. That matrix is the input for the Rating Model. The rating model is same for all the three models and it is described in section 3.2.4

3.2.4 Rating Model (Credibility Network Model)

This model represent the implementation of Rating Algorithm of the news websites. So it produces credibility rating values for each news websites. And further

by considering the connectivity among news websites, the Credibility Network Model is built.

3.2.4.1 Rating Algorithm

The Rating mechanism of news websites is build according to the following algorithm. This algorithm is the pageRank algorithm.

Page Rank Algorithm For Rating News Website

$$\mathbf{PR}\,(\mathbf{A}) = (\mathbf{1}-\mathbf{d}) + \frac{\mathbf{d}(\mathbf{PR}(\mathbf{T1}))}{\mathbf{C}\,(\mathbf{T1})} + ... + \frac{\mathbf{d}(\mathbf{PR}(\mathbf{Tn}))}{\mathbf{C}\,(\mathbf{Tn})}$$

where

PR(A) is the PageRank of page A,
PR(Ti) is the PageRank of pages Ti which link to page A,
C(Ti) is the number of outbound links on page Ti and
d is a damping factor which can be set between 0 and 1.

PageRank of page A is recursively defined by the PageRanks of those pages which link to page A. The PageRank of pages Ti which link to page A does not influence the PageRank of page A uniformly. Within the PageRank algorithm, the PageRank of a page T is always weighted by the number of outbound links C(T) on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T. The weighted PageRank of pages Ti is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's PageRank.

Finally, the sum of the weighted PageRanks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1. Thereby, the extend of PageRank benefit for a page by another page linking to it is reduced.

So this mechanism is used in order to rate websites. So all the websites that are connected to a particular website is considered in order to calculate the rating value of the second website.

3.3 Limitations of the methodology

In the Twitter data collecting phase Tweets are collected having #SriLanka #srilanka #Srilanka #lk #lka in the tweet body. So those tweets will be considered as tweets related to Sri Lankan news stories. So other news related tweets which do not have above tags will not be collected.

3.4 Summary

This chapter elaborates the complete design in detail. In the system architecture, three models are designed based on factors: In-degree, Mentions and URL Recommendations. The algorithms used in each model is described under different sections. As the Ranking algorithm, page Ranking algorithm is chosen and it is described in this chapter. And finally, this chapter discusses the limitations of the research methodology.

Chapter 4

Implementation

This chapter focuses to introduce comprehensive arrangement to carefully analyze the dataset and implement the architecture proposed in the previous chapter. The important features which are extracted and data filtering methods on Twitter are discussed in this section. As described in the previous architecture, three models are developed based on the factors: In-degree, Mentions and URL recommendations. And several problems came across when developing these models. So different techniques are used to overcome those problems when implementing the models.

4.1 Dataset- Data filtering techniques

This research is focusing on data related to Sri Lankan news stories. First, a dataset is created with 85 Sri Lankan news websites and their features. Those data were collected by referring reliable sources and Twitter APIs. Secondly followers details of each website is collected using Twitter APIs. These collected data were used in the analysis of the Model 01(In-degree). As described in the section 3.2.1, Model 01 is developed based on the In-degree of each news website.

As the second part of the data collection, Tweets related to Sri Lankan news stories are collected. Therefore only the tweets related to Sri Lankan news stories were filtered. Mostly Sri Lankan news stories appear with Sri Lanka tag. But the tag

can be seen in different ways such as #lk, #lka, #SriLanka, #SL, #srilanka etc. So mostly tweets with those tags are related to news. Data filtering techniques are used to filter twitter data in order to extract tweets with Sri Lankan news stories. Therefore data is filtered by the time data is loaded using twitter search API with python. That is based on having "Sri Lanka" tag in the tweet body. Figure 4.1 is example tweets posted by ground views and FT Sri Lanka news, having tags #SriLanka, #lka and a URL inside the tweet body. Figure 4.2 is an example tweet of having mentions inside the tweet body.

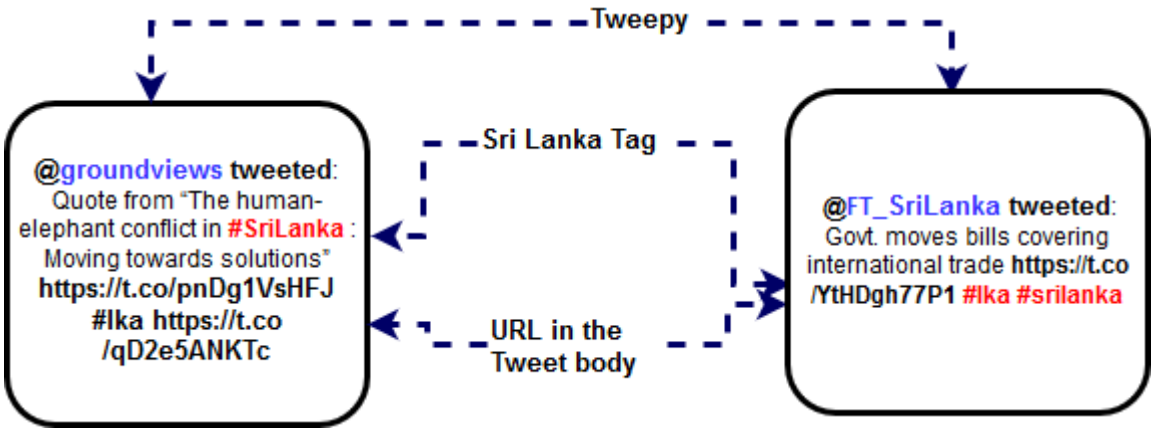

FIGURE 4.1: Examples of News Tweets

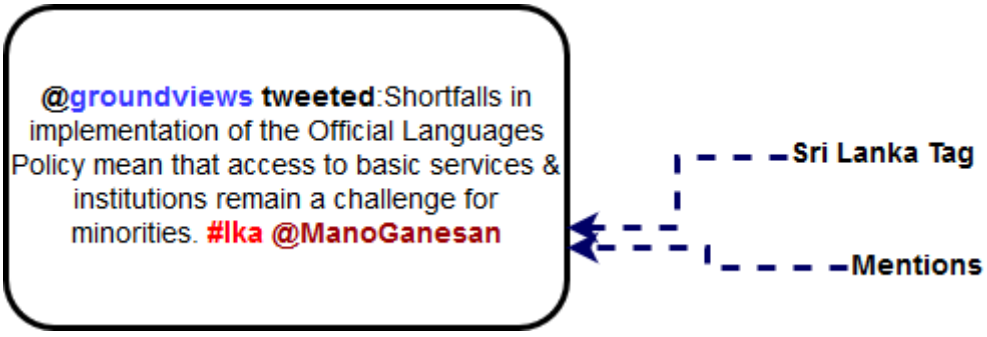

FIGURE 4.2: Example of a Tweet having mentions inside the tweet body

So as the first step of data filtering, Tweets are filtered which having "Sri Lanka" tags inside the tweet body. Those are considered as Tweets related to Sri Lankan

news stories. So at the first step, a dataset is generated with those news tweets. Secondly, those data is analysed in Model 02 (Mentions) and Model 03(URL Recommendations) in order to generate Credibility Network Models. In model 02, we analyse the factor "mentions". Therefore from the data, the user (who posted the tweet) and "mentions" are altered to generate the Credibility Network Model. In the Model 03, we analyse the factor "URL Recommendation". Therefore from the data, user(who posted the tweet) and posted "URLs" are altered to generate the Credibility Network Model. Figure 4.3 depicts the method of data filtering that is described above.

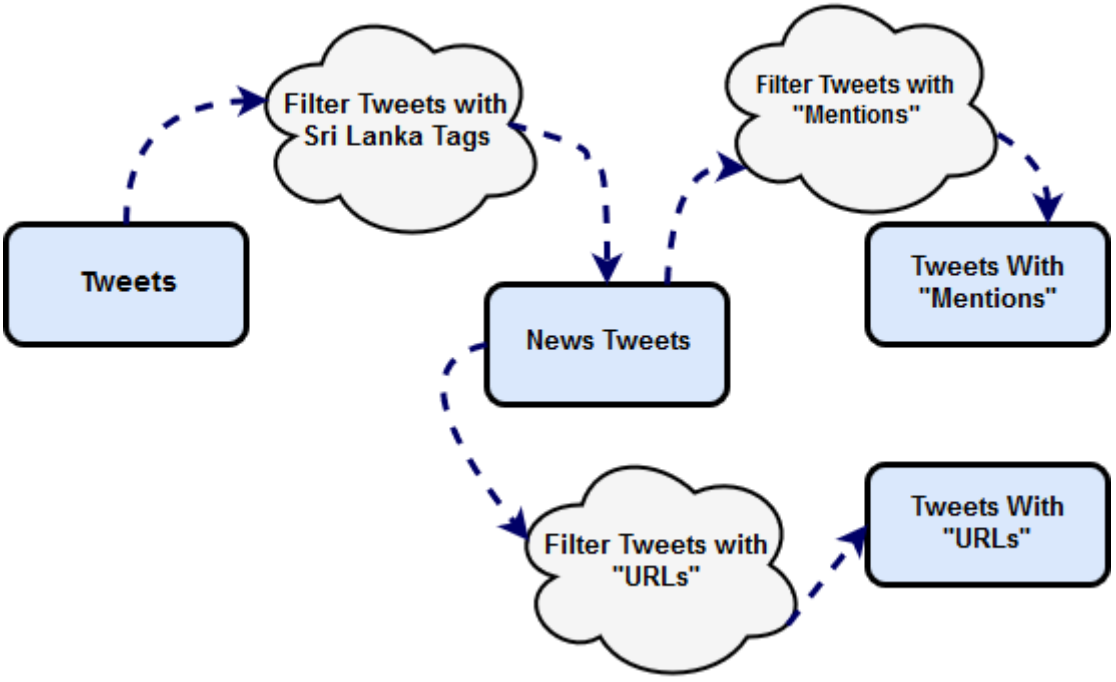

FIGURE 4.3: Data Filtering Method

4.2 Implementation of Rating Model

In section 3.2.2.1, the algorithms used to implement page ranking mechanism is described. This python implementation for the rating method is based on the page ranking algorithm. So the function definition for the page ranking is as following,

def pagerank(G, alpha=0.85, personalization=None, max_iter=100, tol=1.0e-6, nstart=None, weight='weight', dangling=None)

And the parameters described here are as following,

G : graph A NetworkX graph. Undirected graphs will be converted to a directed graph with two directed edges for each undirected edge.

alpha : float, optional Damping parameter for PageRank, default=0.85.

personalization : dict, optional The "personalization vector" consisting of a dictionary with a key for every graph node and nonzero personalization value for each node.By default, a uniform distribution is used.

max_iter : integer, optional Maximum number of iterations in power method eigenvalue solver.

tol : float, optional Error tolerance used to check convergence in power method solver.

nstart : dictionary, optional Starting value of PageRank iteration for each node.

weight : key, optional Edge data key to use as weight. If None weights are set to 1.

dangling : dict, optional

Table 4.1: Input matrix for the rating algorithm

	A	B	C	D
A	0	1	1	0
B	1	0	0	1
C	1	0	0	0
D	0	1	0	0

If there is no connection between two nodes X, Y the position of (X,Y) = 0
If there is connection between the two nodes X, Y then the position of (X,Y) = 1

If we consider Table 4.1 these A, B, C, D as websites (nodes), code 0 represents websites are not linked and code 1 represents websites are linked. For example, from the above matrix, A, B have a link(edge) and A, D dont have a link(edge). Therefore (A, B) position of the matrix is code 1 and (A, D) position of the matrix is code 0.

As described in section 03, there are three models considering factors: In-degree, Mentions and URL recommendation. So in each model, the method of generating this input matrix is unique to that model. The algorithms explained in section 03 clearly describes how the links(edges) are extracted between websites (nodes) by each model. And by considering those links(edges), a matrix is generated in every model according to the above matrix format. So all the matrices generated by each model are in the input format which is explained above. Therefore each model generates different Credibility Network Models because they produce different input matrices. But all the three models use the same ranking algorithm explained in section 3.2.2.1.

Output of the Rating Model

As the output of this model it will generate ranking values for each node in the graph. So each node represent a website. Hence the ranking values are obtained for each news website.

4.3 Programming Languages and APIs

4.3.1 Python

Python is a widely used generalpurpose, high-level programming language developed by Guido van Rossum from National Research Institute for Mathematics and Computer Science in Netherlands. Due to its simplicity, large inbuilt libraries & functions; Python is considered as a stable programming language for file preprocessing purposes.

4.3.2 Twitter Search API

The Twitter Search API is part of Twitters REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days. This API is used to search tweets having Sri Lanka tags in its tweet body.

4.3.3 Twitter Steaming API

The streaming Twitter API is an example of a HTTP Streaming API. HTTP Streaming is a technique used to push updates to a web client. A persistent connection is held open between the web client and the web server so that when the server has new information it can push it to the client. This API is used to collect daily tweets.

4.4 Summary

This section discusses the data filtering techniques used by each model. Then discusses the code level implementation of the ranking algorithm. Furthermore discusses the input format of the ranking algorithm and how each model generates inputs to the ranking model. And furthermore describes the programming languages and Twitter APIs used in the research.

Chapter 5

Results and Evaluation

The purpose of this chapter is to evaluate the result of the generated models described by the previous chapters. We discussed three models based on the factors: Indgree, Mentions and URL recommendations. This chapter describes produced results by each model and the evaluation of the results. We followed a special procedure to do the evaluation of our models. That is based the feedbacks of few experienced and prominent journalists in Sri Lanka. The evaluation procedure is discussed in detail in this chapter. And finally based on the evaluation and literature we are going to decide the most appropriate model from the three generated models.

5.1 Data Collection

Based on System architecture described under section 3.2.1, data collection has two parts. The first part, Dataset 01, is the collection of Sri Lankan news websites with their features. The second part, Dataset 02, is the collection of followers details of each news website. The third part of the dataset is the collection of daily tweets about Sri Lankan news stories. Table 5.1 is consisting of few instances of the dataset 01 and Table 5.2 is consisting of few instances of followers details of the website Colombo Telegraph. Likewise, followers' details of all 85 news website are collected using Twitter APIs. The Dataset 01 and Dataset 02 is used in the Model 01(In-degree).

Table 5.1: Dataset of Sri Lankan news websites (Dataset 01)

Site Name	**Site URL**	**Alexa Rank**	**No of Foll--owers**	**No of Tweets**	**No Follo--wing**	**Google Plus Follo-wers**
Sri Lanka Guardian	srilankaguardian.org	621063	1797	4,482	427	5
Colombo Page	colombopage.com	283149	2044	18819	111	17
Lanka Truth	lankatruth.com/en	95123	191	44	137	1
Hiru News	hirunews.lk/	79673	145182	57911	59	211

Table 5.2: Few instances of the Dataset of Followers of Website- Colombo Telegraph

Screen ID	**Name**	**Loca--tion**	**Frie--nds count**	**Follo--wers count**	**Sta --tus count**	**Favo--rites count**
2226112430	azaff mohamed	Sri Lanka	68	8	75	235
830860539181961217	Samitha Sandaruwan	Sri Lanka	430	17	4	50
884626479891402753	Semondu	Sri Lanka	102	5	4	36
882899452423548928	Benson Morgan Ramiro	Indiana, USA	2472	29	29	9
2292089349	M. Abdul Hameed	Doha	173	55	14	25

As described in the section 3.2.2 and 3.2.3, daily tweets are collected for the model 02 and model 03. Table 5.3 is a sample of data collected to extract mentions and URLs in tweets (which are related to Sri Lankan News Stories) with the Twitter user. The data extraction process is described in sections 3.2.2 and 3.2.3.

5.1.1 Data Collection Results

Table 5.1 represents an instance of dataset 01 which consists of Sri Lankan news websites and five other features of them. Features which have been identified can be used to build credibility measurements for news websites. Above data is

Table 5.3: Set of tweets from dataset of tweets related to Sri Lankan news stories

User	**Tweets having Sri Lanka tags with URLs or Mentions**
groundviews	"Corruption is an on-going issue here. If someone hits a pedestrian on the road with their motorbike, he will go in https://t.co/KlGJqZdeSo
WedivistaraLK	https://t.co/FEtiCTiXTB #lka #Srilanka @MangalaLK
FT_SriLanka	GLOBAL ECONOMICS by Razeen Sally: Thoughts on the Budget and what comes next https://t.co/1Wxbp8qip1 #lka #srilanka
NalakaG	#Digital Transformation in #SriLanka: Opportunities and Challenges in Pursuit of Liberal Policies. Full report we launched ear

collected using python and some reliable known sources. Identified features are as following,

1. Alexa Rank of the website
2. Google plus connections per a website
3. Number of tweets a particular website has posted in its twitter account
4. Number of twitter accounts a news website follows
5. Number of followers per a news website

The dataset 01 described above were analyzed to derive credibility measurements. Those measurements are built upon special behaviors of features of websites. So from those features, Alexa Rank of the website and the total number of followers have considerable similarities. As described earlier in the section 3.2.1, No of followers means the In-degree of the website. The news websites which have good Alexa Rank (a high daily web traffic) seems to have a considerably high amount of twitter followers.

Figure 5.1 depicts how followers and Alexa Ranks behave on the websites which have more than 8000 Twitter followers. Those websites seem to have a considerable

traffic according to the facts of Alexa Rank. (If the Alexa Rank of particular website is less than 10000, that website has a considerable web traffic)

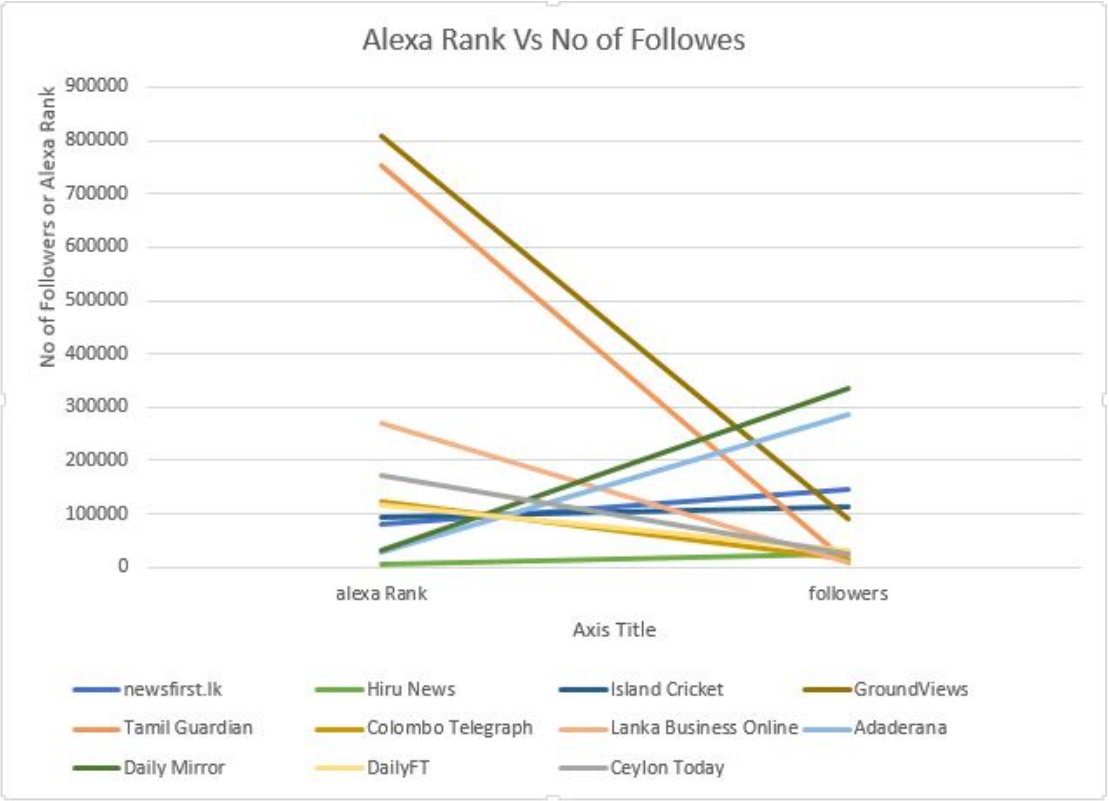

FIGURE 5.1: Similarities between Alexa Rank and No of Followers

According to this analysis, if the Alexa Rank is high (means that particular website has low web traffic) the total number of Twitter followers for a website is considerably low. Alternatively, if the Alexa Rank is a small value (means that particular website has high web traffic) the total number of Twitter followers of that website is considerably high.

So that means Total number of Twitter Followers of a website is good indicator for measure credibility. Therefore the model 01 described under section 3.2.1 is based on the Twitter followers of news websites.

5.2 Evaluation Procedure

5.2.1 Survey Results

To evaluate our model we conducted a survey with few selected reputed journalists in Sri Lanka. Most of them have more than ten years of experience in online journalism. They were selected based their experience in journalism and the reputation. On the survey, they were asked to rate 12 websites, which were randomly selected, based on online news credibility. And again asked them to weight the factors that influence the credibility of news such as Author, News source and News Medium. Then by considering their ratings and weighting factors, new credibility rating values for those 12 websites were generated in order to do the evaluation of Credibility Network Models. Credibility rating values are generated based on the equation (A). In later in this section, those generated rating values are used to evaluate the three models.

$$\mathbf{Rank\ \ of\ \ a\ \ website(A)} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i}}^{\mathbf{n}} (\mathbf{RA_i}) \times \mathbf{W_i} \qquad \Leftarrow \qquad \mathbf{(A)}$$

Where, n is the number of journalists. **(RAi)** is the rating value given to the website by a journalist(i). **Wi** is the weight given to the factor News Source by the journalist(i). Table 5.4 shows the ranking values obtained on the selected 12 websites by using the equation **(A)**.

TABLE 5.4: Ranks obtained from Survey Results

No	**Website**	**Ranks obtained from Survey Results**
1	Lanka Voice	1.68
2	Ground Views	2.96
3	Gossip Lanka	2.35
4	Virakesari	1.883333333
5	Hiru News	2
6	Tamil Guardian	2
7	Asian Mirror	3.966666666
8	BBC Sinhala	3.28
9	News First	4.833333334
10	Vikalpa Voices	2.85
11	Colombo Telegraph	5.21
12	Adaderana	5.05

5.2.2 Method of Evaluation

It is necessary to evaluate the algorithms and methods which are used by using an evaluation criteria. The developed three models produce different rating values for websites and users in the Credibility Network. The purpose of the evaluation is to identify the model which produces the most accurate credibility rankings that match with the survey results. Therefore a rank comparison method is used to compare ranking values produced by each model with the survey results. Hence in this research, Spearman's rank correlation coefficient is used to compare the models with survey results.

5.2.2.1 Spearman's rank correlation coefficient

Spearman's correlation determines the strength and direction of the monotonic relationship between the two variables rather than the strength and direction of the linear relationship between the two variables. A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases. Figure 5.2 represent examples of monotonic relationship and non-monotonic relationships.

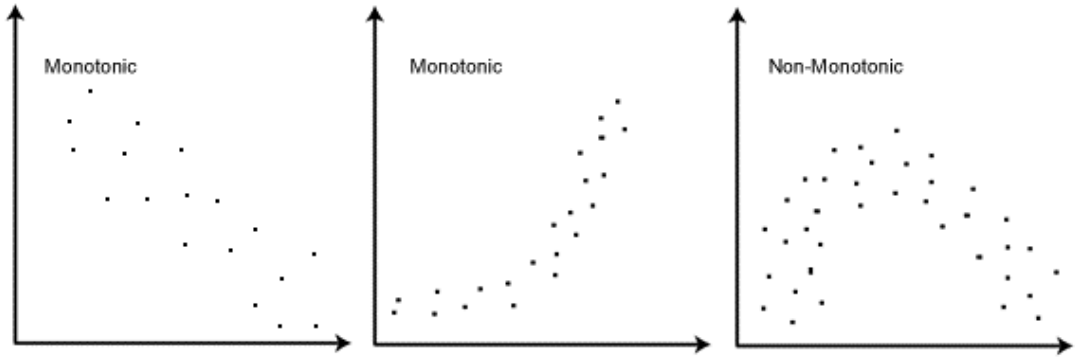

FIGURE 5.2: Monotonic and non-monotonic relationships

5.2.2.1.1 Definitions and Calculations

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size n, the n raw scores $\mathbf{X_i}, \mathbf{Y_i}$ are converted to rank $\mathbf{rg(X_i)}$, $\mathbf{rg(Y_i)}$. The formula of Spearman's Rank correlation coefficient, is given as,

$$\boldsymbol{\rho = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1))}} \qquad \Leftarrow \qquad \boldsymbol{(B)}$$

ρ : Coefficient of Rank correlation

$\mathbf{D_i}$: Difference in Ranks between paired values of X and Y

$$\mathbf{D_i = rg(X_i) - rg(Y_i)}$$

$\mathbf{n}$: Sample size

5.2.2.1.2 Values of Spearman's rank correlation coefficient

The Spearman correlation coefficient, ρ, can take values from +1 to -1.

A ρ of +1 indicates **a perfect association** of ranks,

A ρ of zero indicates **no association** between ranks and

A ρ of -1 indicates **a perfect negative association** of ranks.

The closer ρ is to zero, **the weaker the association** between the ranks.

5.3 Evaluation of the Models

As discussed in section 03, three Credibility Network Models are developed based on the factors: Indegree, Mentions, URL Recommendations. Each model produces different ranking values to news websites based on their implementations. It is necessary to identify which model is highly associated with the ranks obtained by the survey results. Therefore each model is evaluated with Spearman's rank correlation coefficient in order to identify strong associations with the ranks obtained by the survey results. So in this section, all the three models are evaluated in order to identify most accurate Credibility Network Model.

5.3.1 Result and Evaluation of Credibility Network Model Generated by Model 01 (In-Degree)

This model is implemented according to the algorithm explained under section 3.2.1. So in Table 5.2 consists of few instances of followers details of Colombo Telegraph News website. Likewise, followers details to all 85 websites were collected. But the network model only represents 40 websites. Those websites are randomly selected over having High, Middle, Low total number of Twitter followers. Because according to the analysis it is discovered that Number of Followers is a good credibility indicator. Resulted Credibility Network is depicted in Figure 5.3.

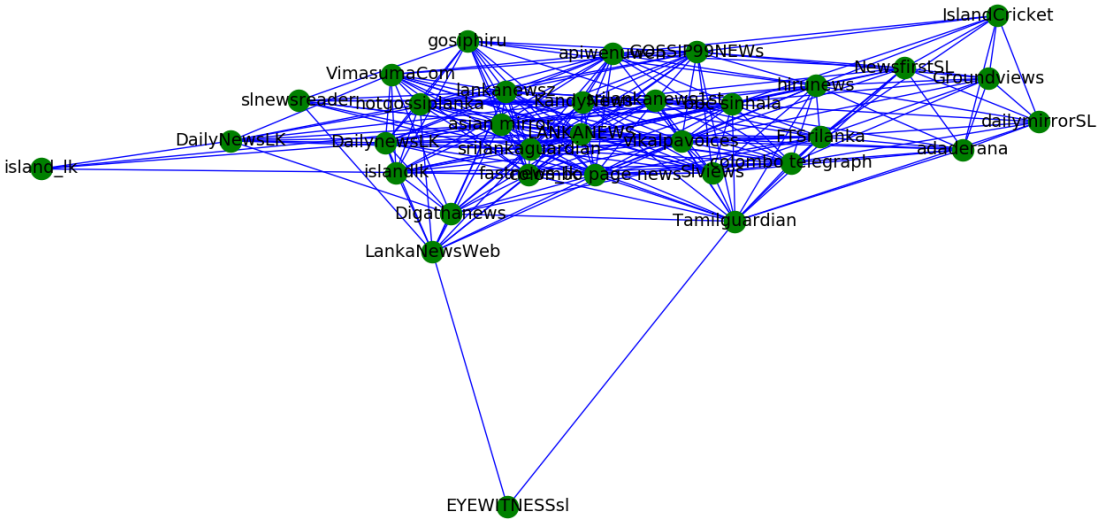

FIGURE 5.3: Resulted Credibility Network of Model 01

The ranking values produced by this model is evaluated with Spearman's rank correlation coefficient. Like explained in section 5.2.2.1.1, each rank in a sample is converted to a new rank based on the Spearman's rank correlation coefficient. In this analysis, the new ranks Xi are allocated according to the increasing order of ranks generated by the Model 01. So table 5.5 represents the ranks produced by the model 01 and the converted ranks for the same 12 websites shown in Table 5.4. So the ranks produced by the model 01 is compared with ranks obtained from the survey results for the same 12 websites.

Table 5.5: Ranks obtained by Model 01 and New Ranks

Website Name	**Generated Rank by Model 01**	New Rank (Xi)
Lanka Voice	1.102529836	1
Ground Views	2.091371957	2
Adaderana	2.411397462	3
News First	2.573328486	4
Virakesari	2.959844717	5
BBC Sinhala	3.278557778	6
Gossip Lanka	3.280423463	7
Colombo Telegraph	3.472654746	8
Hiru News	3.478766828	9
Tamil Guardian	3.504891444	10
Vikalpa Voices	4.09695303	11
Asian Mirror	4.501379723	12

In the same way, ranks obtained by survey results also converted to news ranks to the scale of 1 to 12. Table 5.6 shows the ranks obtained by survey results and newly converted ranks (Yi). And again the new ranks (Yi) are allocated according to the increasing order of ranks obtained from the survey results.

Table 5.6: Ranks obtained by Survey Results and allocated New Ranks

Website Name	**Ranks Obtained By Survey Results**	**New Rank (Yi)**
Lanka Voice	1.68	1
Virakesari	1.883333333	2
Hiru News	2	3
Tamil Guardian	2	4
Gossip Lanka	2.35	5
Vikalpa Voices	2.85	6
Ground Views	2.96	7
BBC Sinhala	3.28	8
Asian Mirror	3.966666666	9
News First	4.833333334	10
Adaderana	5.05	11
Colombo Telegraph	5.21	12

Now according to the Table 5.7 $\mathbf{D_i}$ (Difference in Ranks between paired values of X and Y) and $\mathbf{D_i^2}$ are calculated in order to determine the Coefficient of Rank correlation (ρ).

TABLE 5.7: Determination of Coefficient of Rank correlation (ρ) for the Model 01

Website	**Model 01 Rank (Scale 1-10)**	**Ranks obtained from survey (Scale 1-10)**	**New Rank (X_i)**	**New Rank (Y_i)**	**Di**	**D_i^2**
Lanka Voice	1.102529836	1.68	1	1	0	0
Ground Views	2.091371957	2.96	2	7	-5	25
Adaderana	2.411397462	5.05	3	11	-9	81
News First	2.573328486	4.83334	4	10	-6	36
Virakesari	2.959844717	1.883333	5	2	3	9
BBC Sinhala	3.278557778	3.28	6	8	-2	4
Gossip Lanka	3.280423463	2.35	7	5	2	4
Colombo Telegraph	3.472654746	5.21	8	12	-4	16
Hiru News	3.478766828	2	9	3	6	36
Tamil Guardian	3.504891444	2	10	4	6	36
Vikalpa Voices	4.09695303	2.85	11	6	5	25
Asian Mirror	4.501379723	3.966666	12	9	3	9
						$\sum D_i^2 = 281$

By using the Equation **(B)**,

$$\boldsymbol{\rho = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1))}}$$

$\mathbf{n = 12}$ (there are 12 websites) and $\sum \mathbf{D_i^2} = 281$

Then $\rho = \mathbf{0.018}$

By using equation (B), calculated Coefficient of Rank correlation (ρ) for the Model 01 is 0.018. According to 5.2.2.1.2, the closer (ρ) is to zero, the weaker the association between the ranks. Therefore association between the ranks produced by the model 01 and ranks obtained by the survey results have a weak relationship

because 0.018 is closer to zero.

5.3.2 Result and Evaluation of Credibility Network Model Generated by Model 02 (Mentions)

This model is implemented according to the algorithm explained under section 3.2.2. So in Table 5.3 consists of few instances of tweets, which have mentions in the tweet body, used to develop this Credibility Network Model. Figure 5.4 represents the Credibility Network obtained through this Model and each node in the network represents a website or an active Twitter User. The ranking values produced by this model and the ranking values obtained by the survey results are evaluated with Spearman's Coefficient of Rank correlation in order to determine the association between them. Further, this model is evaluated in the same way as the Model 01.

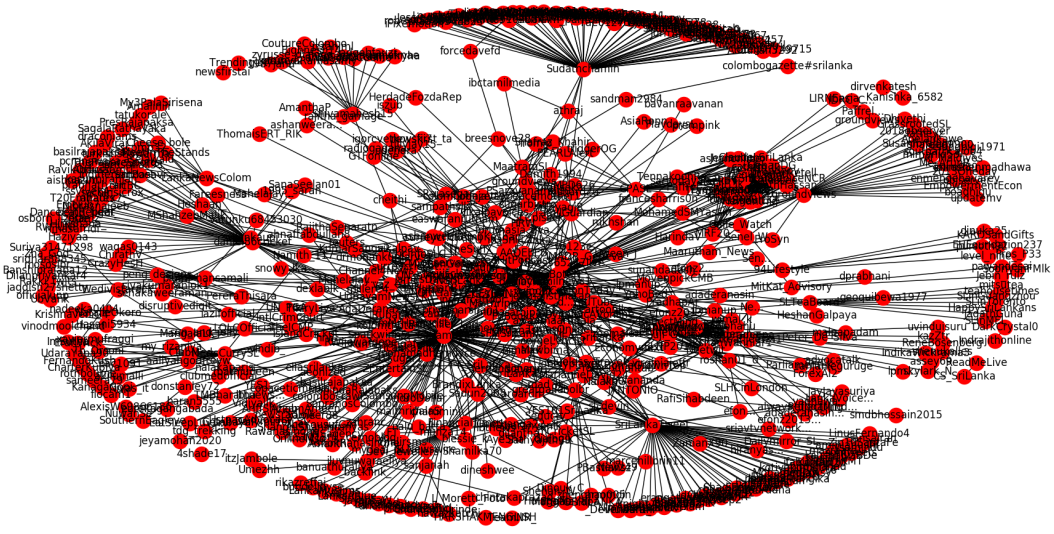

FIGURE 5.4: Credibility Network Obtained by Model 02

The ranking values produced by this model is evaluated with Spearman's rank correlation coefficient. Like explained in section 5.2.2.1.1, each rank in a sample is converted to a new rank. In this analysis, the new ranks (Xi) are allocated according to the increasing order of ranks generated by the Model 02. And Table 5.6, which represent the allocation of new ranks (Yi) to obtained ranks of survey

results, is used for this calculation as well. Further calculations are same as section 5.3.1. Then Coefficient of Rank correlation (ρ) is determined for Model 02 as well. Now according to the Table 5.8 Di (Difference in Ranks between paired values of X and Y) and D_i^2 are calculated in order to determine the Coefficient of Rank correlation (ρ).

TABLE 5.8: Determination of Coefficient of Rank correlation (ρ) for the Model 02 (Mentions)

Website	**Model 01 Rank (Scale 1-10)**	**Ranks obtained from survey (Scale 1-10)**	**New Rank (X_i)**	**New Rank (Y_i)**	**D_i**	**D_i^2**
Gossip Lanka	0.594749787	2.35	1	5	-4	16
Lanka Voice	0.739616146	1.68	2	1	1	1
Asian Mirror	0.803498245	3.9666	3	9	-6	36
Hiru News	0.878867689	2	4	3	1	1
Virakesari	0.921997851	1.8833	5	2	3	9
BBC Sinhala	1.051830689	3.28	6	8	-2	4
News First	0.997735084	4.8334	7	10	-3	9
Vikalpa Voices	1.116409782	2.85	8	6	2	4
Adaderana	1.213161691	5.05	9	11	-2	4
Ground Views	1.824795222	2.96	10	7	3	9
Colombo Telegraph	1.832414654	5.21	11	12	-1	1
Tamil Guardian	1.999297365	2	12	4	8	64
						$\sum D_i^2 = 158$

By using the Equation **(B)**,

$$\boldsymbol{\rho = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1))}}$$

$\mathbf{n = 12}$ (there are 12 websites) and $\sum \mathbf{D_i^2} = 158$

Then $\rho = \mathbf{0.448}$

By using equation (B), calculated Coefficient of Rank correlation (ρ) for the Model 01 is 0.448. According to 5.2.2.1.2, the (ρ) is between 0 and +1. The significant Spearman correlation coefficient value of 0.448 confirms a positive correlation between the Ranks produced by the Model 02 and the ranks obtained by the survey results.

5.3.3 Result and Evaluation of Credibility Network Model Generated by Model 03 (URL Recommendations)

This model is implemented according to the algorithm explained under section 3.2.3. So in Table 5.3 consists of few instances of tweets, which have "URLs" posted in the tweet body, used to develop this Credibility Network Model. Figure 5.5 represents the Credibility Network obtained through this Model and each node in the network represents a website or an active Twitter User. The ranking values produced by this model and the ranking values obtained by the survey results are evaluated with Coefficient of Rank correlation in order to determine the association between them. Further, this model is evaluated in the same way as the Model 01, Model 02.

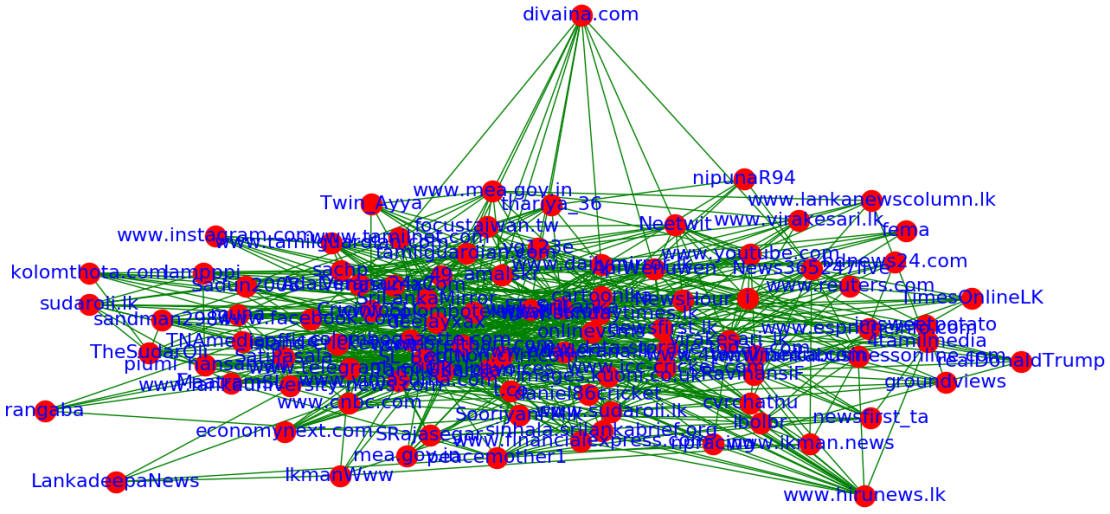

FIGURE 5.5: Credibility Network Obtained by Model 03

The ranking values produced by this model is evaluated with Spearman's rank correlation coefficient. Like explained in section 5.2.2.1.1, each rank in the sample is converted to a new rank based on the Spearman's rank correlation coefficient. In this analysis, the new ranks (Xi) are allocated according to the increasing order of generated ranks by Model 03. And Table 5.6, which represent the allocation of new ranks (Yi), to the obtained ranks of survey results, is used for this calculation as well. Further calculations are same as section 5.3.1 and 5.3.2. Then Coefficient of Rank correlation (ρ) is determined for the Model 03 as well. Now according to the Table 5.9 Di (Difference in Ranks between paired values of X and Y) and D_i^2 are calculated in order to determine the Coefficient of Rank correlation (ρ).

TABLE 5.9: Determination of Coefficient of Rank correlation (ρ) for the Model 03 (URL Recommendations)

Website	**Model 01 Rank (Scale 1-10)**	**Ranks obtained from survey (Scale 1- 10)**	**New Rank (X_i)**	**New Rank (Y_i)**	**D_i**	**D_i^2**
Lanka Voice	1.1672796	1.68	1	1	1	0
Ground Views	1.2092166	2.96	2	7	-5	25
Gossip Lanka	1.7237396	2.35	3	5	-2	4
Virakesari	1.791969	1.883333	5	2	3	9
Hiru News	1.8415426	2	4	3	1	1
Tamil Guardian	1.9135218	2	6	4	2	4
Asian Mirror	1.950952	3.966666	7	9	-2	4
BBC Sinhala	2.5703016	3.28	8	8	0	0
News First	3.3192166	4.833334	9	10	-1	1
Vikalpa Voices	3.8940924	2.85	10	6	4	16
Colombo Telegraph	4.7825482	5.21	11	12	-1	1
Adaderana	5.2464346	5.05	12	11	1	1
						$\sum D_i^2 = 66$

By using the Equation **(B)**,

$$\boldsymbol{\rho = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1))}}$$

$\mathbf{n = 12}$ (there are 12 websites) and $\sum \mathbf{D_i^2} = 66$

Then $\rho = \mathbf{0.768}$

By using equation (B), calculated Coefficient of Rank correlation (ρ) for the Model 03 is 0.768. According to 5.2.2.1.2, the (ρ) is between 0 and +1. The significant Spearman correlation coefficient value of 0.768 confirms a strong positive correlation between the Ranks produced by the Model 03 and the survey results.

5.4 Comparison of the Models

As discussed in the section 5.3, association of each model with the survey results is different. Table 5.10 represent obtained Coefficient of Rank correlation (ρ) values for each model.

TABLE 5.10: Coefficient of Rank correlation ρ of each model

	Model 01 (In-degree)	**Model 02 (Mentions)**	**Model 03 (URL Recommendations)**
Coefficient of Rank correlation (ρ)	0.018	0.448	0.768

All correlation analyses express the strength of linkage or cooccurrence between to variables (here variables are ranking methods) in a single value between -1 and +1. This value is called the correlation coefficient (ρ). A positive correlation coefficient indicates a positive relationship between the two variables (the larger A, the larger B) while a negative correlation coefficients expresses a negative relationship (the larger A, the smaller B). A correlation coefficient of 0 indicates that no relationship between the variables exists at all.

Based on the ρ values obtained, Coefficient of Rank correlation of Model 03 (URL Recommendations) is the highest and it is closer to +1. The value of Model 01(In-degree) is closer to zero, means association between Model 01 and the survey results is weak and the value of the Model 02 is comparatively low than the Model 03.Model 02 has a positive correlation with survey results but it is not as strong as Model 03. According to the Spearman correlation, Model 03 and the survey results have a strong positive correlations. Hence ranks produced by model 03 is highly associated with ranks obtained by survey results. Therefore Model 03 is the acceptable and most accurate model over the three models.

5.5 Summary

This chapter discusses the evaluation of the three models of In-degree, Mentions and URL Recommendations. The Spearman's rank correlation coefficient is used to determine the association of each ranking model with the ranks obtained by the survey result. According to that evaluation, Model 03 (URL Recommendations) has the highest Spearman's rank correlation coefficient. Therefore Model 03 has a strong positive correlation with the survey results. The purpose of the evaluation is succeeded by identifying a Credibility Network Model that produce accurate credibility ranking values. This chapter elaborate how these calculations are made and how the Model 03 is selected as the best model.

Chapter 6

Conclusions

In this chapter, we summarize the methods we followed in order to build credibility network model and the rating mechanism. Apart from that the challenges and problems encountered while doing the research are discussed. Furthermore, discusses the applicability of these models to obtain credibility ratings of news websites.

6.1 Introduction

The main aim of this research work was to propose a mechanism to rate the credibility of online news stories. Therefore this research focus on determining news credibility based on the credibility of the news source. In the thesis, the concepts of Network and linguistic approaches and the nature of user influence in Twitter and a ranking algorithm were presented thoroughly. A survey of the existing approaches which perform credibility analysis online has been conducted. After that, an overview of the conceptual approach developed has been given in the thesis. The approach namely rating the credibility of online news stories based on the news source credibility has been designed and developed. A proper evaluation based on a survey and evaluating method was carried out to show the effectiveness of our Credibility Network Model. Finally, various interesting results about online news credibility based on influential factors (In-Degree, Mentions, URL Recommendations) is discovered.

6.2 Conclusions about research Objectives

n this research, we have developed three models that determine the credibility of the news source based on the influential Twitter factors: In-degree, Mentions and URL Recommendations [28]. So all the three factors are analyzed in three different Credibility Network Models and evaluated to identify best Credibility Network Model over the three models. When developing these models, different data collecting methods were applied. That process is described under section 04. Furthermore, different algorithms are developed in each model to extract connections among news websites. Section 4.2 elaborate the process of developing each model to generate Credibility Ranking values for news websites.

As discussed in section 5.2.1, to do the evaluation, first a survey was conducted with the reputed and experienced journalists in Sri Lanka. On the survey, they were asked to rate 12 different news websites and some other questions related to online news credibility. Then by considering their feedbacks ranking values were calculated for the previous 12 websites. And the each Credibility Network Model (In-degree, Mentions, URL Recommendations) is evaluated with the survey results. The evaluation procedure is explained in section 5.2. For the evaluation, Spearman's rank correlation coefficient is used. Spearman's correlation coefficient (ρ) measures the strength and direction of the association between two ranked variables. By using Spearman's rank correlation, the association between each Credibility Network Model (In-degree, Mentions, URL Recommendations) and the survey results were determined in section 5.3. And according to that evaluations Model with highest (closer to +1) Spearman's rank correlation coefficient is identified and that model is the Model 03 (URL Recommendations). Therefore it is identified that Model 03 produces credibility ranking which is highly associated with journalists credibility ranking values. In other world Model 03 matches with the perspective of journalists. According to the studies of Jilin, Rowan[29], URL of the content recommendation on Twitter is a better way of getting users attention. In the perspective of credibility of news, results of our research show that URL recommendation for news is a better way of identifying credible information.

That is a new contribution made by our research towards the area of the research.

When focusing on other two models, Model 01 (In-degree) and Model 02 (Mentions), they have comparatively low Spearman's rank correlation with the survey results. As described in section 5.4, Coefficient of Rank correlation (ρ) of Model 01 (In-Degree) is closer to zero. That indicates a weak relationship between the Model 01(In-Degree) and the survey results. And again according to the literature, popular users who have high In-degree are not necessarily influential in terms of spawning retweets or mentions [28]. On our perspective based on the analysis conducted, points that In-degree of the news website may not necessarily influence the news credibility online. That is again a contribution to this research area.

The Model 02 (Mentions) has a positive Coefficient of Rank correlation (ρ) between 0 and +1, but as not as high as the value of Model 03 (URL recommendations). Therefore model 02 has an average relationship with survey results.

According to this studies finally, a Credibility Network Model is introduced with proper evaluation. Therefore the problem addressed in the beginning, that is determining the credibility of online news website/source can be answered based on the model 03 (URL Recommendations).

The two research questions we discussed can be answered based on our model. Because the Credibility Network Model we developed considering the URL Recommendation produces fine credibility rating values as described above. Therefore credibility of an online news website/source can be determined based on our Credibility Network Model (Model 03). And again based on the news source credibility, humans can decide the credibility of a news story posted in that particular news source. In that way the research questions can be justified based on the developed Credibility Network Model.

6.3 Conclusions about research problem

In research problem we discussed in section 2.1, we discuss the importance of having credible information and the bad influence that arise when the information we get is not credible. One way of evaluating the credibility of the news story is evaluating the credibility of the news website/source. So our Credibility Network Model introduces a new mechanism for evaluating the credibility of news websites. The problem we discuss is addressed by this model because it ranks the credibility of news websites based on their network behavior on Twitter. Therefore our contribution to this problem is introducing a new model that ranks news websites based on their credibility.

6.4 Limitations

As described in section 1.6, we consider Twitter as the social media platform and collect Twitter data for this research. Currently, we consider only the Sri Lankan news environment, therefore, all the collected data related is to Sri Lanka news stories. The Credibility Network Model represents the news websites which are currently active. Therefore we need to update the Model with current data from time to time in order to be updated with newly arriving news websites.

6.5 Implications for further research

As for further implications, global data can be collected and improve the model with global data. In this research, we evaluate the credibility of news website/-source in order to determine the credibility. Therefore this can be further improved to examine the credibility of Authors of new stories etc. Therefore this model can be improved to determine the credibility of the news story based on the author of the news story combining with the source credibility.

Bibliography

[1]Conroy, Niall J., Victoria L. Rubin, and Yimin Chen. "Automatic Deception Detection: Methods For Finding Fake News". Proceedings of the Association for Information Science and Technology 52.1 (2015): 1-4. Web.

[2] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A Stylometric Inquiry into Hyperpartisan and Fake News. arXiv preprint arXiv:1702.05638.

[3] Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012, February). Tweeting is believing?: understanding microblog credibility perceptions. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 441-450). ACM.

[4]Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. Spread of (Mis)Information in Social Networks. Games and Economic Behavior, 70(2):194227.

[5] Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (pp. 675-684). ACM.

[6] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent Features of Rumor Propagation in Online Social Media. In Data Mining (ICDM), 2013 IEEE 13th International Conference on, pages 11031108. IEEE.

[7] Hardalov, M., Koychev, I., & Nakov, P. (2016, September). In Search of Credible News. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications (pp. 172-180). Springer International Publishing.

[8]Chen, Y., Conroy, N. J., & Rubin, V. L. (2015, November). Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (pp. 15-19). ACM.

[9] Borah, P. (2014). The hyperlinked world: A look at how the interactions of news frames and hyperlinks influence news credibility and willingness to seek information. Journal of ComputerMediated Communication, 19(3), 576-590.

[10]Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In

Proceedings of the Second Workshop on Computational Approaches to Deception Detection, pages 717, San Diego, California, June. Association for Computational Linguistics.

[11]Feng, S., Banerjee, R. & Choi, Y. (2012). Syntactic Stylometry for Deception Detection. 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 171175.

[12]Rubin, V. & Lukoianova, T. (2014). Truth and deception at the rhetorical structure level. Journal of the American Society for Information Science and Technology, 66(5).DOI: 10.1002/asi. 23216

[13] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. News in an Online World: The Need for an "Automatic Crap Detector". In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, ASIST 15, pages 81:181:4, Silver Springs, MD, USA. American Society for Information Science. 21

[14]Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open Information Extraction from the Web. Commun. ACM, 51(12):6874, December.

[15]Amr Magdy and Nayer Wanas. 2010. Web-based Statistical Fact Checking of Textual Documents. In Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents, SMUC 10, pages 103110, New York, NY, USA. ACM.

[16]Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. 2007. TextRunner: Open information extraction on the web. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 2526, Rochester, New York, USA, April. Association for Computational Linguistics.

[17]Ciampaglia, G., Shiralkar, P., Rocha, L., Bollen, J. Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks.

[18]Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. Spread of (Mis)Information in Social Networks. Games and Economic Behavior, 70(2):194227.

[19]Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. 2015. Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks. In Proceedings of the 24th International Conference on World Wide Web, WWW 15 Companion, pages 977982, New York, NY, USA. ACM.

[20]Aditi Gupta, Ponnurangam Kumaraguru . Credibility Raning of Tweets during High Impact Events. In ACM 2012.

[21]Zhiwei Jin, Juan Cao, Yongdong Zhang, & Jiebo Luo.News Verification by Exploiting Conflicting Social Viewpoints in Microblogs.Association for the Advancement of Artificial Intelligence(2016)

[22]Anwar us Saeed, Zareen Sharf. Twitter News Credibility Meter.International Journal of Computer Applications (0975 8887)

[23] Sagui, Fernando M.; Maguitman, Ana G.; Chesnevar, Carlos I.; Simari, Guillermo R. Modeling News Trust: A Defeasible Logic Programming Approach-Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, Vol. 12, Nm. 40, 2008, pp. 63-72Asociacin Espaola para la Inteligencia Artificial Espana

[24] A. Garca and G. Simari. Defeasible Logic Programming: An Argumentative Approach. Theory and Practice of Logic Programming, 4(1):95138, 2004.

[25]Ryosuke Nagura, Yohei Seki, Noriko Kando, and Masaki Aono. A method of rating the credibility of news documents on the web. In SIGIR 06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 683684, New York, NY, USA, 2006. ACM Press.

[26]Greenwood, D.N., 2013. Fame, facebook, and twitter: How attitudes about fame predict frequency and nature of social media use. Psychology of Popular Media Culture, 2(4): 222-236.

[27]Barry Fox. Google searches for quality not quantity. New Scientist magazine, 2497:24. 30 April2005.

[28]Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17), 30.

[29]O'Donovan, J., Kang, B., Meyer, G., Hollerer, T., & Adalii, S. (2012, September). Credibility in context: An analysis of feature distributions in twitter. In

Privacy, Security, Risk and Trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom) (pp. 293-301). IEEE.

[30]Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election (No. w23089). National Bureau of Economic Research.

[31]Teevan, J., Ramage, D., & Morris, M. R. (2011, February). # TwitterSearch: a comparison of microblog search and web search. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 35-44). ACM.

[32]Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1079-1088). ACM.

[33]Sullivan, D. (2009). Twitters Real Time Spam Problem. Search Engine Land.

[34]Corcoran, M. (2009). Death by cliff plunge, with a push from twitter. The New York Times.

[35] Gupta, A., & Kumaraguru, P. (2012, April). Credibility ranking of tweets during high impact events. In Proceedings of the 1st workshop on privacy and security in online social media (p. 2). ACM.

[36] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM.

[37]M. Mendoza, B. Poblete, and C. Castillo, Twitter Under Crisis: Can we trust what we RT? in 1st Workshop on Social MediaAnalytics (SOMA 10). ACM Press, Jul. 2010.

[38] Y. Ding, E. Yan, A. Frazho, and J. Caverlee, Pagerank for ranking authors in co-citation networks," Journal of the American Society for Information Science and Technology, vol. 60, no. 11, pp. 22292243, 2009.

[39] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine," Computer networks and ISDN systems, vol. 30, no. 1, pp.107117, 1998.

[40] L. Page, S. Brin, R. Motwani, and T. Winograd, The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

Chapter 7

Appendices

7.1 Appendix A

7.1.1 Survey Questions

1. Your experience in online journalism

   - More than 5 years
   - More than 10 years
   - Other:

2. Which of the following sites do you visit for International news?

   i BBC

   ii The Guardian

   iii Asia Times

   iv AdaDerana24x7

   v DailyFT

   vi Daily News

   vii Other:

3. Which of the following sites do you visit for Local news? *

i Colombo Telegrap

ii Ground Views

iii DailyFT

iv News First

v Lanka Voice

vi SriLanka Mirror

vii Vikalpa

viii Ceylon Today

ix Daily News

x Lanka C News

xi Other:

4. Which factors would you consider most important in determining the credibility of a news item? Give your answer by weighting the following factors * 0,1.5,2.5,3.0,4.5,5,6.5,7.5,8,9.5

1. Author

2. News Source/website

3. News Medium (Traditional news papers/Online news))

5. Rate your credibility ranking of the Following News Websites (1-HighlyCredible, 2-Good, 3-Average, 4-Poor, 5-VeryPoor) *

i Colombo Telegrap

ii Vikalpa

iii Asian Mirror

iv Adaderana

v News First

vi Gossip Lanka News

vii Ground Views

viii BBC

ix Lanka Voice

x Tamil Guardian

xi Hirugossip

xii Virakesari

xiii Colombo Telegrap

xiv Vikalpa

xv Asian Mirror

xvi Adaderana

xvii News First

xviii Gossip Lanka News

xix Ground Views

xx BBC

xxi Lanka Voice

xxii Tamil Guardian

xxiii Hirugossip

xxiv Virakesari

6. Have you ever experienced fake news from any of the sites you ranked 1 or 2 above?

- Never
- Almost Never
- Rarely
- Sometimes
- Frequently

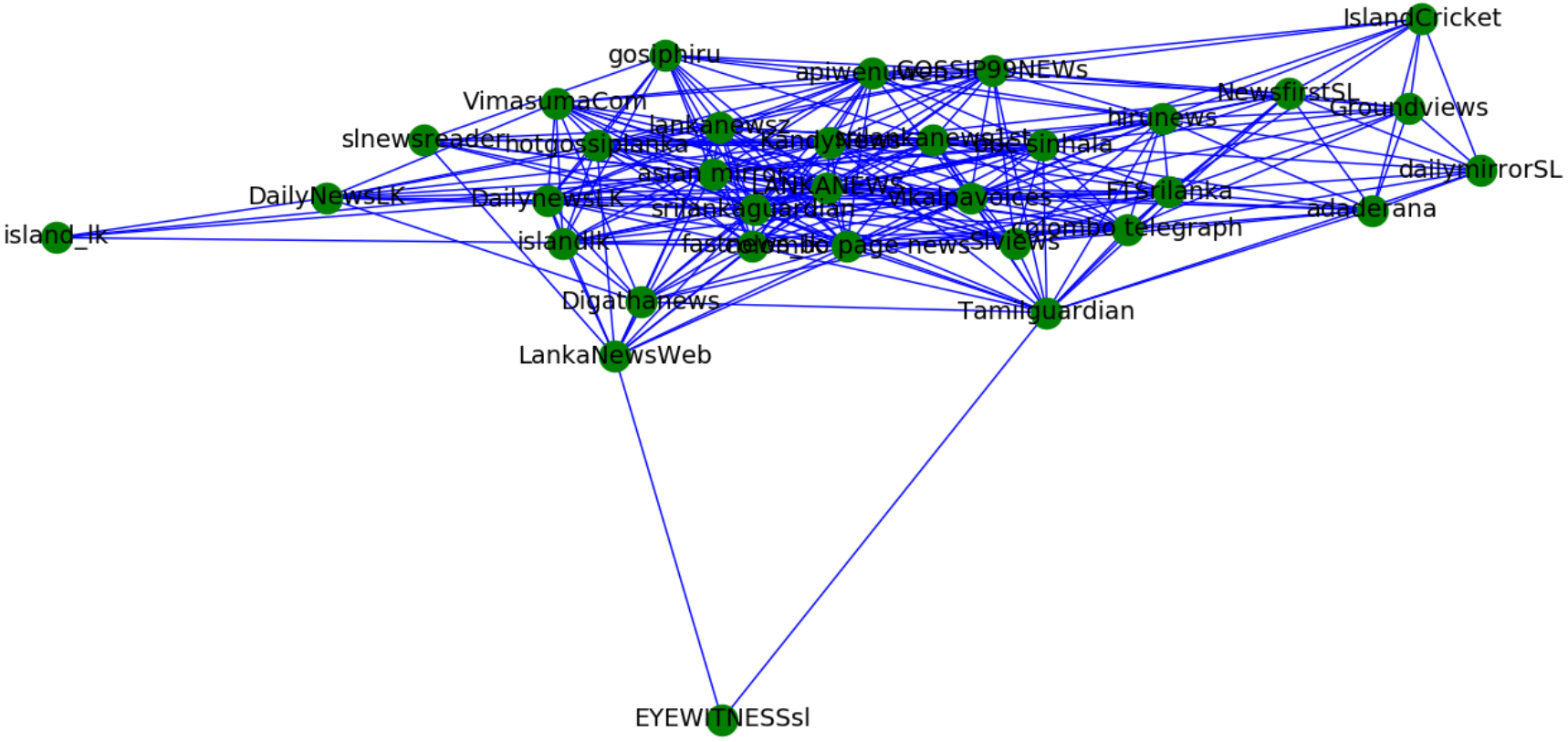

FIGURE 7.1: Resulted Credibility Network of Model 01

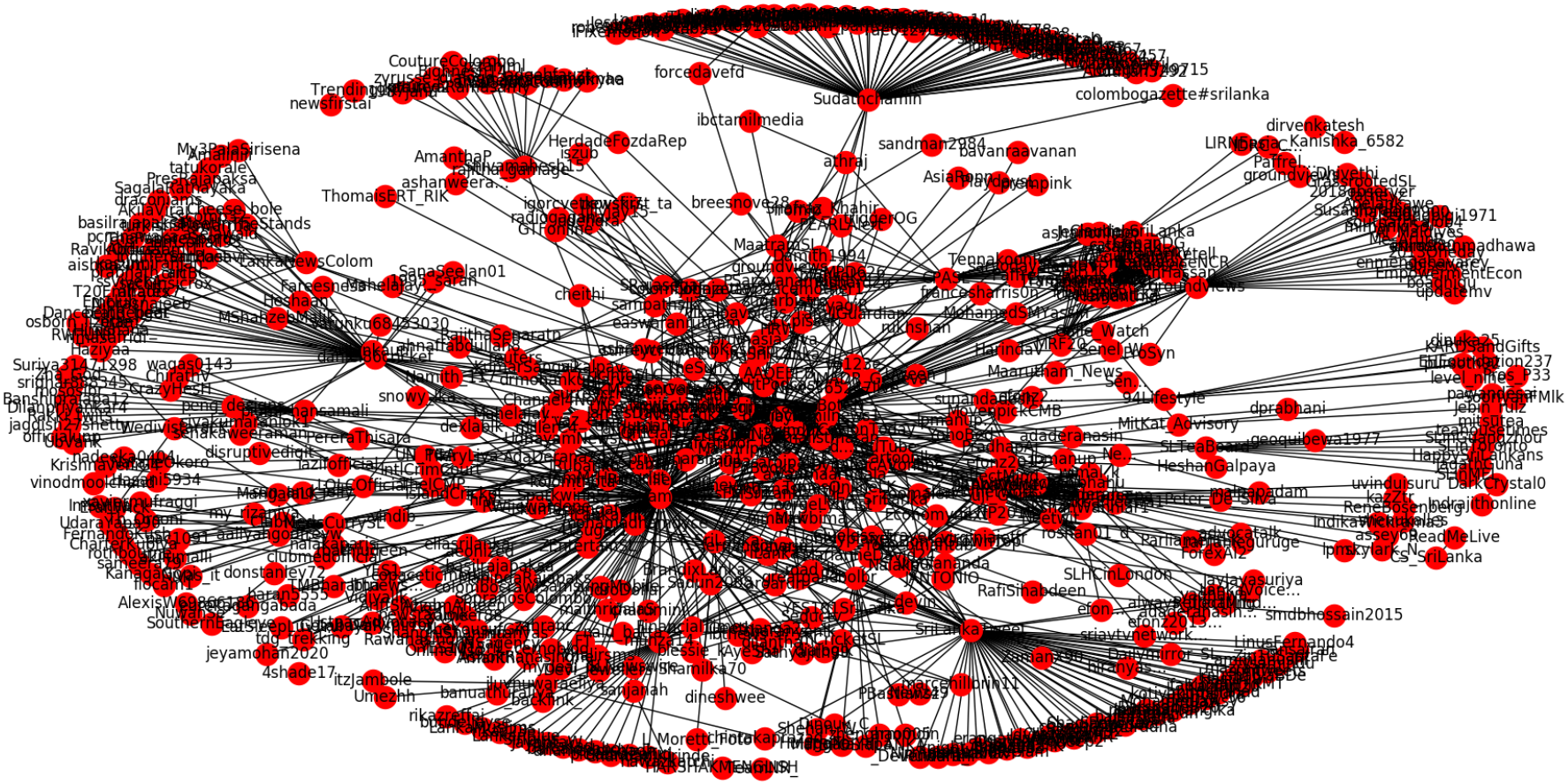

FIGURE 7.2: Resulted Credibility Network of Model 02

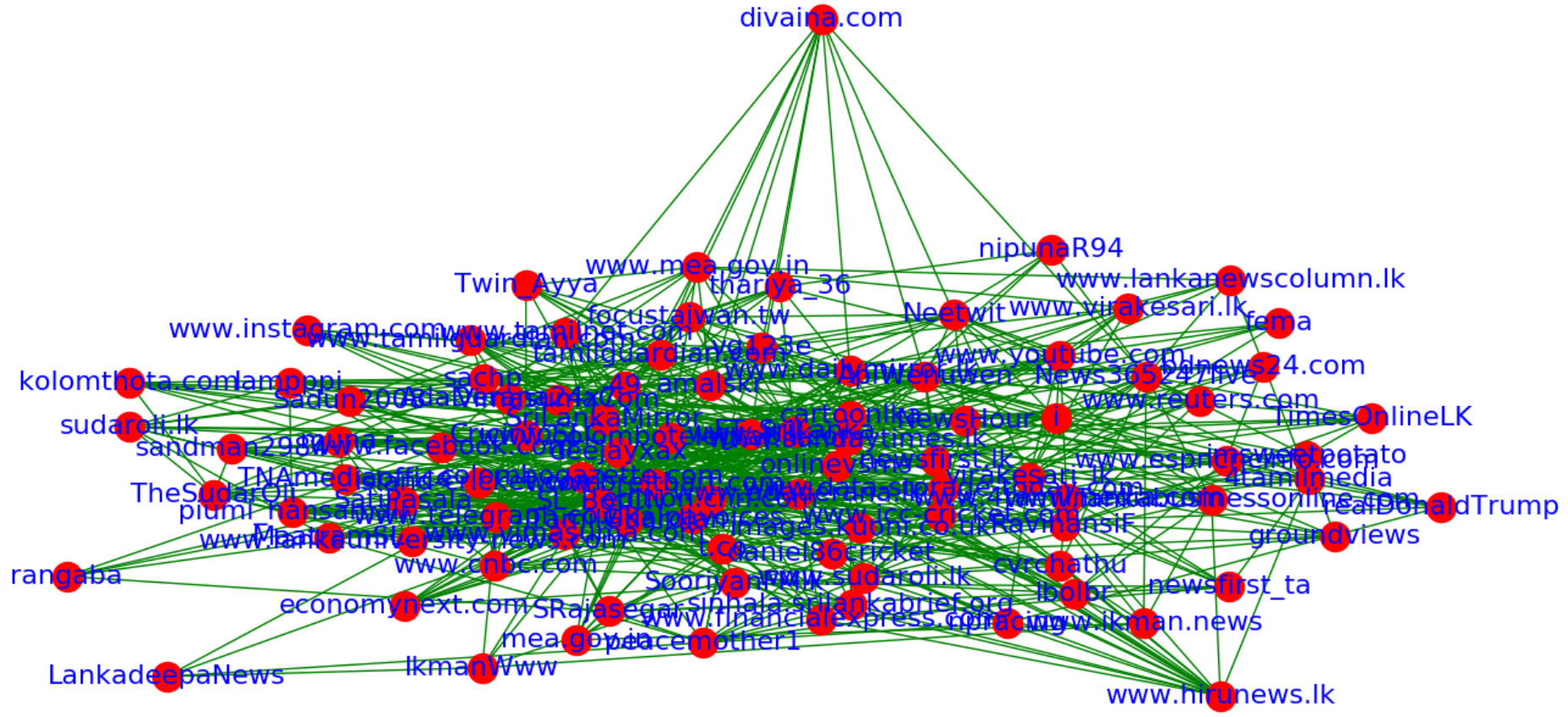

FIGURE 7.3: Credibility Network Obtained by Model 03