



Sentiment Based Approach to Analyze Fake News Related to COVID-19 in Sri Lanka

K.I. Somasinghe

(Reg. No.: MS19816500)

M.Sc. in Information Systems

Supervisor: Mr. Prasanna S. Haddela

September 2021

**Department of Computer Systems Engineering
Faculty of Computing
Sri Lanka Institute of Information Technology**

Declaration

I certify that this report does not contain any material previously submitted for a degree or diploma at any university without acknowledgement, and that it does not contain any material previously published or written by another person to the best of my knowledge and belief, except where due reference is made in text.



.....
K.I. Somasinghe (MS19816500)

19.09.2021

.....
Date



.....
Supervisor: Mr. Prasanna S. Haddela

..... 22.10.2021

Date

Abstract

Generation and spread of fake news have drastically increased with the growth of technology and advancement of online media platforms. Today, rather than using traditional resources to get information, most people rely on the internet and it has become a part of every individual's life since this is a one of the simplest methods to acquire information on almost everything. This internet based media has become a source of sharing news and these sources are used by companies, political parties as well as social influencers etc.

Fake news changes the perception of the viewers and diverts them from the reality. By analyzing fake news in Sinhala Language related to COVID-19 which is a disastrous situation to not only Sri Lanka but the whole world it would be a great advantage to notify people regarding fake news and the resources they use to spread fake news and reduce unethical sharing of news, to protect the authenticity of news that reaches people and the authenticity of the journalism field.

This research presents an approach to analyze the effect of polarity in the sentiment of the news data and analyze how it affects towards fake news in Sinhala language using textual data. The proposed model uses natural language processing techniques such as sentiment analysis and machine learning algorithms such as Logistic Regression, Support Vector Machine, Naïve Bayes.

Acknowledgements

First of all, I would like to thank my supervisor, Mr. Prasanna S. Haddela, for providing an opportunity and continuous guidance for me, and for the patience, encouragement and expertise of my research and review of the thesis. The encouragement he gave helped me to constantly plan and write the thesis.

I want to thank Dr. Anuradha Jayakody, the coordinator of the IS program, for his guidance and the support given me during the hardest times.

Last but not least, I want to thank my family and friends for their support throughout my life.

Table of Contents

| | |
|---|-----|
| Table of Contents | iv |
| List of Figures | vi |
| List of Tables | vii |
| Chapter 1 : Introduction | 1 |
| 1.1 Context | 1 |
| 1.2 Social Media: Definition | 2 |
| 1.3 Fake News: Definition | 2 |
| 1.4 Sentiment Analysis: Definition | 2 |
| 1.5 Motivation | 2 |
| Chapter 2 : Research Design | 4 |
| 2.1 Problem Statement | 4 |
| 2.2 Goals | 5 |
| 2.3 Research Objectives | 5 |
| 2.4 Research Questions | 6 |
| Chapter 3 : Theoretical Background | 7 |
| 3.1 Related Work | 7 |
| 3.2 Characteristics of Fake News | 7 |
| 3.3 Manual Classification, Fact Checking | 8 |
| 3.4 Automatic Fact Checking and Fake News Detection | 9 |
| 3.5 Sentiment Analysis vs Fake News Detection | 11 |
| 3.6 Sentiment Analysis in Sinhala Language | 12 |
| Chapter 4 : Conceptualization and Methodology | 14 |
| 4.1 Proposed Architecture | 14 |
| 4.2 Model Overview | 15 |
| 4.3 Dataset | 16 |
| 4.4 Data preprocessing | 17 |
| 4.5 Feature Extraction | 20 |
| Vectorizing Data: Bag-Of-Words | 21 |
| Vectorizing Data: TF-IDF | 22 |
| 4.6 Sentiment Analysis | 22 |
| 4.7 Challenges of Sentiment Analysis | 24 |
| 4.8 Fake News Detection | 25 |
| Chapter 5 : Results & Performances | 26 |

| | |
|---|----|
| 5.1 Evaluation Metrics | 26 |
| 5.2 Comparison of Preprocessing Techniques between Algorithms | 28 |
| 5.3 Evaluating Sentiment Analysis | 29 |
| 5.4 Evaluating Fake News Analysis..... | 32 |
| 5.5 Sentiment vs Fake News Comparison..... | 34 |
| Chapter 6 : Conclusion..... | 36 |
| References..... | 38 |

List of Figures

| | |
|--|----|
| Figure 4.1 Methodology..... | 15 |
| Figure 4.2 Word Frequency Distribution..... | 18 |
| Figure 4.3 Word Cloud | 19 |
| Figure 4.4 Stop words | 19 |
| Figure 4.5 Feature extraction | 21 |
| Figure 4.6 Term frequency | 22 |
| Figure 5.1 Comparison between preprocessing techniques | 29 |
| Figure 5.2 Sentiment Analysis | 30 |
| Figure 5.3 Comparison of algorithms: Sentiment analysis | 31 |
| Figure 5.4 Fake news detection | 33 |
| Figure 5.5 Comparison between algorithms: Fake news analysis | 34 |
| Figure 5.6 Sentiment analysis vs Fake news detection..... | 35 |

List of Tables

| | |
|--|----|
| Table 5.1 Performance evaluation parameters | 26 |
| Table 5.2 Comparison of preprocessing techniques | 28 |
| Table 5.3 Confusion matrix: Naive Bayes | 29 |
| Table 5.4 Confusion matrix: Logistic Regression | 30 |
| Table 5.5 Confusion matrix: SVM..... | 30 |
| Table 5.6 Performance metrics: Sentiment analysis | 31 |
| Table 5.7 Confusion matrix: Naive Bayes | 32 |
| Table 5.8 Confusion matrix: Logistic Regression | 32 |
| Table 5.9 Confusion matrix: SVM..... | 32 |
| Table 5.10 Performance metrics: Fake news analysis | 33 |