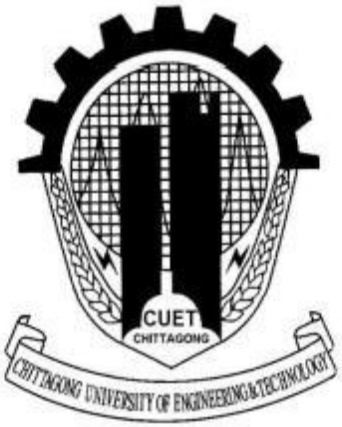
Conspiracy Detection by Real Time Email Analysis

Md. Ikramul Hoque

ID: 1304115

October, 2018

Conspiracy Detection by Real Time Email Analysis



This thesis is submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering.

Md. Ikramul Hoque

ID: 1304115

Supervised by

Abu Hasnat Mohammad Ashfak Habib

Associate Professor,

Department of Computer Science and Engineering (CSE)

Chittagong University of Engineering and Technology (CUET)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CHITTAGONG UNIVERSITY OF ENGINEERING AND TECHNOLOGY (CUET)

CHITTAGONG – 4349, BANGLADESH.

The thesis titled "**Conspiracy Detection by Real Time Email Analysis**" submitted by ID 1304115, Session 2016-2017 has been accepted as satisfactory in fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering(CSE) as B.Sc. Engineering to be awarded by Chittagong University of Engineering and Technology (CUET).

Board of Examiners

1. _____ Chairman

Abu Hasnat Mohammad Ashfak Habib
Associate Professor
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

2. _____ Member
(Ex-officio)

Dr. Mohammad Shamsul Arefin
Professor and Head
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

3. _____ Member
(External)

Dr. Asaduzzaman
Professor
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

Statement of Originality

It is hereby declared that the contents of this project is original and any part of it has not been submitted elsewhere for the award if any degree or diploma.

_____ _____

Signature of the Supervisor

Date:

Signature of the Candidate

Date:

Acknowledgement

Prima facie, I am grateful to the Almighty for giving me the strength for successful completion of this project. Then I would like to express my sincere gratitude to my honorable project supervisor Abu Hasnat Mohammad Ashfak Habib, Assistant Professor, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, for his valuable advices, constructive suggestions and sincere guidance with all the necessary facilities for assimilation, research and preparation for the project. I place on record, my sincere gratitude to Dr. Asaduzzaman, Professor, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, for his kind encouragement and cooperation. I would like to thank my family for their constant love and support. Finally, I would like to take this opportunity to express my gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

Abstract

Supervised vector-based methods to sentiment can design rich lexical meanings. This method for machine learning is largely used in present days. Sentiment analysis for text documents has been a growing field of text mining among researchers for the past few decades. Nevertheless, Email data sentiment analysis, a general means of social networking and communication, has been studied strongly.

Email has become the most popular communication tools for official purpose. Almost every private company uses their own mail server for exchanging their official mail. So, it has a great significance in terms of business and communication.

In the other hand conspiracy is a social concept that has also a great importance and impact over the working place. It is a pure psychological concept. It influences in the progress of any working place.

In this thesis, we have proposed a method to turn this psychological concept into a machine that can automatically detect the conspiracy among the employee by analyzing their email data in real time. Here we have proposed the design using vector based classification method for analyzing the text data. We have used TFIDF method to vectorization and prioritize the frequency of conspiracy related word and concept. And also we used Logistic Regression, a prediction based classifier to classify the text sentiment.

Table of Contents

Chapter 1

Introduction ---------- **1**

1.1 Introduction---------- 1

1.2 Previous Works ---------- 2

1.3 Present State and Contribution---------- 3

1.4 Motivation---------- 3

1.5 Prospects ----------4

1.6 Organization of the Project----------4

Chapter 2

Literature Review ---------- **6**

2.1 Introduction ---------- 6

2.2 Conspiracy---------- 6

2.3 Conspiracy Theory ---------- 7

2.4 Psychology of Conspiracy Theories ----------9

2.5 Organizational Conspiracy Theories---------- 9

2.5.1 Organizational Identification----------10

2.5.2 Organizational Commitment----------11

2.5.3 Job Satisfaction----------11

2.5.4 Implications----------12

2.6 Machine Learning----------12

2.6.1 Supervised Learning----------13

2.6.2 Unsupervised Learning----------13

2.7 Text Classifier-The Basic Building Blocks----------13

2.8 Sentiment Analysis---------- 15

2.9 Dataset----------16

2.10 Features----------17

2.11 Data Processing ----------18

2.12 Python----------19

2.12.1 Advantages----------19
2.12.2 Uses----------20
2.12.3 Matplotlib---------- 20
2.12.4 Numpy----------20
2.12.5 Scikit-Learn----------21
2.12.5.1 History----------21
2.12.6 Pandas----------22
2.12.6.1 Usages of Pandas----------22
2.12.7 PyMySQL----------22
2.12.8 wordCloud----------23
2.12.9 BeautifulSoup----------24
2.12.10 CountVectorzer and LagisticRegression----------24
2.12.11 TF-IDF----------24
2.13 Private Mail Server----------25

Chapter 3

Conspiracy Detection Methodology ----------26
3.1 System Architecture ---------- 26
3.1.1 Data Acquisition and Refining----------28
3.1.1.1 Data Refining----------30
3.1.2 Data Processing Module---------- 32
3.1.2.1 Tokenization----------32
3.1.2.2 Feature Vector----------32
3.1.3 Training Module----------34
3.1.4 Testing in Real-time----------36
3.2 Analytical Representation of the Architecture----------37
3.2.1 Labelling the Email Data----------37
3.2.2 Clean the Data----------37
3.2.3 Process the Data----------38
3.2.4 Train the Model----------39
3.2.5 Collecting the Mail Data in Real Time----------40

3.2.6 Generating Output----------40
3.3 Complexity Analysis----------42
Chapter 4
Implementation of Conspiracy Detection Framework----------43
4.1 Experimental Setup----------43
4.2 Email Exchanging System----------43
4.3 Detection System Implementation----------48
4.4 Conclusion ----------49

Chapter 5
Experimental Results---------- **50**
5.1 Data Collection ---------- 50
5.1.1 Green Data Collection----------50
5.1.2 Red Data Collection----------51
5.1.2.1 Financial Conspiracy----------52
5.1.2.2 Organizational Conspiracy----------52
5.1.2.3 Reputational Conspiracy----------52
5.2 Evolution of the System----------54
5.2.1 Evaluates from the Mail Dataset----------54
5.2.2 Evaluates from the Real Time Mail----------55
Chapter 6
Conclusion and Future Recommendation ---------- **57**
6.1 Conclusion ---------- 57
6.2 Limitations and Suggestions for Future ---------- 57

References ---------- **59**

List of Figures

Figure 3.1: Conspiracy Predictor Model Architecture----------27
Figure 3.2: Green Mail Content----------28
Figure 3.3: Red Mail Content----------29
Figure 3.4: A Peek into the Dataset----------29
Figure 4.1: Login Interface of CUET Mail----------44
Figure 4.2: Inbox of the System----------45
Figure 4.3: Send Box Interface----------45
Figure 4.4: Compose Mail Interface----------46
Figure 4.5: 'Email Data' Table Interface---------- 47
Figure 4.6: Interfaced 'Result' Database Table----------47
Figure 4.7: User Verification Table Interface----------48
Figure 4.8: Email Sending System----------48
Figure 4.9: Detection Illustration----------49

Figure 5.1: Green Data CSV File----------51
Figure 5.2: Red Dataset ----------53
Figure 5.3: Percentage of Error in Green Data----------54
Figure 5.4: Percentage of Error in Red Data----------55

List of Tables

Table 3.1: Contracted Word and Long Form----------31
Table 5.1: Real time Accuracy of Mail Data----------56