

Sentiment indicator for e-mails

K.D.Chandima

169304M

Faculty of Information Technology

University of Moratuwa

2019

LB/DC/192/2019
IT 01/232

Sentiment indicator for e-mails

K.D.Chandima

169304M

Supervisor: Dr. Lochandaka Ranathunga

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Honours Degree of Bachelor of Science in Information Technology.

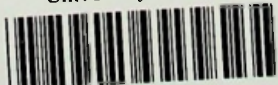
LIBRARY
UNIVERSITY OF MORATUWA, SRI LANKA
MORATUWA

004 "19"
004(043)

2019

TH3884

University of Moratuwa



TH3884

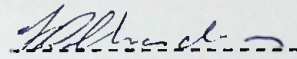
+
CO-2010

TH3884

Declaration

I hereby declare that this project report entitled "Sentiment indicator for e-mails" contains my own work and has not been submitted and will not be submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student: K. D. Chandima

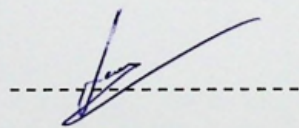


Signature of Student

Date: 25/04/2019

Supervised By

Dr. Lochandaka Ranathunga



Dedication

I would like to dedicate my project, "Sentiment indicator for e-mails";

To my project supervisor Dr. Lochandaka Ranathunga,

To my Father and sister who have given support for proof reading project documents,

To the Instructors of Faculty of Information Technology at University of Moratuwa who have supported me to find requirements for the project.

Acknowledgments

I wish to express my sincere thanks to my supervisor Dr. Lochandaka Ranathunga for guiding me to have a successful outcome in this project work.

I would like to acknowledge and extend my heartfelt gratitude to the following people who have made the completion of this project possible –

My father and sister for the help given to proof read all the project documents by spending their valuable time,

Instructors of Faculty of Information Technology at University of Moratuwa who have supported me to find resources and requirements for the project.

Academic, non-academic staff members who helped me throughout this project,

Batch mates who motivated in doing the project, and friends.

Abstract

People receive large amount of e-mails from friends, relatives, companies, institutions and known and unknown people. It is not possible the user to read all the received e-mails during a busy day. Normally, people avoid reading most of the received e-mails by giving priority to the important e-mails. E-mails may contain happy news or sad news or defamatory contents or obscene contents. If there is a way for a user to get an idea of what kind of news would be there in just before he or she opens e-mails, then he or she can choose which e-mails should be opened first and find out which e-mails would make him happy. If the user can find out any obscene contents in an e-mail just before he opens it, then he or she can avoid the embarrassment of opening the email in front of a stranger and then the user will be able to delete it without simply opening it. A proper e-mail categorization is required and this attempt is to categorize e-mails according to the sentiment and the ingredient of the e-mail content, by indicating emoticons on the subject of e-mails. Text Preprocessing, Feature Extraction, Sentiment Classification methods are used to identify sentiments and proper emoticons are used to indicate sentiments on the subjects of e-mails. Machine learning and lexicon based approaches are used to predict the sentiment of emails. A better accuracy level is expected from machine learning in the process of sentiment extraction.

Table of Contents

Declaration	i
Dedication	ii
Acknowledgments	iii
Abstract	iv
Chapter 1	1
1.1 Introduction	1
1.2 Background	1
1.3 Aim and Specific Objectives	2
Chapter 2	3
Literature Review	3
2.1 Categorize e-mails by filtering Spam e-mails	3
2.2 E-mail classification	4
2.3 Categorize e-mails from Sentiment Analysis	4
2.4 Using Natural-Language-Processing (NLP)	5
2.5 Lexicon based sentiment analysis	6
2.5 Applications of Sentiment Analysis	6
Chapter 3	7
Technology adapted	7
3.1 Text preprocessing	7
3.2 Feature Extraction	7
3.3 Sentiment Classification	7
3.4 Natural-Language-Processing (NLP)	7
3.5 Graphical User Interface Creation	8
3.6 Reading E-mails from Gmail account	8
3.7 Lemmatization	8
3.8 Implementation of Neural Networks	8
3.9 Tools, Technologies and Libraries	9
3.10 Dataset	9
Chapter 4	10
Methodology	10
4.1 Statement of Research Problem	10
4.2 Expected outcome/alternative approaches	10
4.3 Reading E-mails	11
4.4 Text Preprocessing	11

4.4.1 Data Cleaning	12
4.4.2 Tokenization	13
4.4.3 Idiom handling.....	13
4.4.4 Negation handling.....	14
4.4.5 Part of Speech Tagging.....	14
4.4.6 Lemmatization	14
4.4.7 Stop word removal.....	14
4.5 Sentiment analysis using Lexicon approach	14
4.6 Machine Learning/ Deep Learning approach	15
4.6.1 Features selection	15
4.6.2 Neural Network	16
4.6.3 Training and testing datasets	16
4.7 Obscene image detection	16
Chapter 5	18
Analysis and Design	18
5.1 Text preprocessing	19
5.2 Feature Extraction	19
5.3 Sentiment Classification	19
5.4 System training and evaluation	19
5.5 Graphical User Interface (GUI) design.....	20
5.6 Neural Network.....	21
Chapter 6	22
Implementation	22
6.1 Reading emails from Gmail accounts	22
6.2 Creation of GUI	23
6.3 Data Preprocessing.....	24
6.3.1 Remove previously forwarded information from email body.	24
6.3.2 Removing images/links from the e-mail body	25
6.3.3 Remove punctuations	25
6.4 Detect obscene images from the body part of the image	26
6.5 Remove extra white spaces from email body	26
6.6 Tokenization	27
6.7 Lemmatization	27
6.8 Rcpetition handling.....	28
6.9 Idiom handling	28

Chapter 7.....	29
Evaluation and Results.....	29
7.1 Remove forwarded information method.....	29
7.2 Tokenize sentences method.....	30
7.3 Punctuation removing method.....	30
7.4 Repetition removal method.....	31
7.5 Sentiment extraction from Lexicon based approach.....	32
7.6 Sentiment extraction from Machine Learning Approach.....	33
7.7 Classifier Model.....	34
7.8 Accuracy.....	34
Chapter 8.....	35
Discussion.....	35
8.1 Discussion.....	35
8.2 Conclusion.....	35
8.3 Future works.....	36
References.....	38

List of Figures

Figure 4 1: emoticons.....	10
Figure 4 2: Previously forwarded details on emails.....	12
Figure 5 1: System Diagram.....	18
Figure 5 2: Graphical user interface with output.....	20
Figure 5 3: Neural Network Design.....	21
Figure 6 1: Source code to access Gmail API.....	22
Figure 6 2: Graphical User Interface implementation.....	23
Figure 6 3: Main emoticons.....	23
Figure 6 4: Forwarded e-mail 1.....	24
Figure 6 5: Forwarded e-mail 2.....	24
Figure 6 6: Source code for detect previously forwarded information.....	24
Figure 6 7: Source code for remove URLs from email body.....	25
Figure 6 8: Detecting obscene images from image URL.....	26
Figure 6 9: Source code for remove extra white spaces.....	26
Figure 6 10: Source code for tokenizing text into sentences and words.....	27
Figure 6 11: Lemmatization based on POS tags.....	27
Figure 7 1: E-mail body with previously forwarded information.....	29

Figure 7 2: After removing forwarded information.....29
Figure 7 3: Tokenized sentences with line numbers30
Figure 7 4: Punctuation, Symbols, Numbers removed output.....30
Figure 7 5: Confusion Matrix for Machine Learning Approach.....34