Universiti Teknologi MARA

Sentiment Mining in Twitter for Early Depression Detection

Najihah Salsabila binti Ishak

Thesis submitted in fulfilment of the requirements for Bachelor of Computer Science (Hons.) Faculty of Computer and Mathematical Sciences

February 2021

ACKNOWLEDGMENT

Through the name of Allah, the most Merciful and the most Gracious.

Alhamdulillah, praises, and thanks to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given. First of all, I would like to express my sincere gratitude to my supervisor, Mrs Norulhidayah binti Isa, for her excitement, persistence, thoughtful comments, helpful knowledge, realistic advice and never-ending ideas that have supported me tremendously in my project and writing of this thesis. Thanks to her extensive knowledge, deep experience and technical skills in data science, I was able to successfully complete this project. Without its encouragement and direction, this idea would not have been feasible. I couldn't have dreamed that I had a better supervisor in my final year project research.

Besides, I would like to give a special thanks to my CSP650 lecturer, Mrs Norlina binti Mohd Sabri for tutoring me through two semesters, sharing all the knowledge to make sure that I complete this research on time and reminding me to ensure I am on the right path.

My special gratitude goes to my family member. Without the encouragement of them, it would not be possible to publish this thesis. Last but not least, I would like to express my appreciation to my dearest classmates and friends who has helped and gave me their support throughout this semester directly and indirectly.

May God shower the above-mentioned personalities with achievement and glory in their lives

ABSTRACT

Depression is a severe and pervasive threat to public health. These people like to express their thought, opinion and suggestion using social media network. Twitter is a popular microblogging site for users to post status updates (tweets). These tweets often reflect views on social issues, including psychological. Sentiment Analysis refers to natural language processing and text mining approaches to classify thoughts or sentiments from the tweet. Machine learning is an implementation of artificial intelligence (AI) that allows systems to learn and build on knowledge without being directly programmed automatically. This paper applies sentiment analysis, text mining, and machine learning to psychology to identify depression in Twitter user. The usefulness of using the user's tweet to measure depression studies using a literature review. The utility of current Python sentiment tools to a set of vocabulary used in microblogging is determined. The use of linguistic features to detect the sentiment in Twitter tweets are explored. A classifier model is developed using Naïve Bayes characteristics. A comparison between built-in Scikit Learn Naïve Bayes algorithm, and the scratch Naïve Bayes algorithm is used to measure its effectiveness in terms of accuracy. At the end of this project, a prototype that can classify tweet is developed and used to monitor the tweets' sentiment probability.

TABLE OF CONTENTS

CONTENT		PAGE
SUPERVISOR APPROVAL		i
STUDENT DECLARATION		ii
ACKNOWLEDGMENT		iii
ABSTRACT		iv
TABLE OF CONTENTS		v
LIST OF FIGURES		viii
LIST OF TABLES		ixx
CHAPTER ONE: INTRODUCTION		
1.1	Background of Study	1
1.2	Problem Statement	2
1.3	Objectives	3
1.4	Project Scope	4
1.5	Project Significance	4
1.6	Overview of Research Framework	4
1.7	Summary	6
CHAPTER TWO: LITERATURE REVIEW		
2.1	Mental Health	7
	2.1.1 Depression in Social Media	7
2.2	Web Scraping	9
	2.2.1 Twitter Scraping	9
2.3	Sentiment Analysis	10
	2.3.1 Stages in Sentiment Analysis	10

2.3.2 Advantages and Disadvantages of Sentiment Analysis 12
2.4 Naïve Bayes Algorithm 12
2.4.1 Naïve Bayes Classifier in Detecting Sentiment Analysis 13
2.5 Similar Research Article Comparison 14
2.6 Summary 22

CHAPTER THREE: METHODOLOGY

3.1 Background Study 29
3.1.1 Chapter 1 (Introduction) 29
3.1.2 Chapter 2 (Literature Review) 29
3.2 Data Collection 30
3.3 Training Data Labelling 31
3.3.1 Text Blob 31
3.3.2 Sentiment Intensity Analyzer 32
3.3.3 MPQA Subjectivity Lexicon Dictionary 33
3.4 Data Cleaning 35
3.4.1 Lowercase Conversion 35
3.4.2 Mention Removal 36
3.4.3 URL Removal 37
3.4.4 Punctuation, Numbers, Hashtag Symbols Removal 38
3.4.5 Retweet Removal 38
3.4.6 White Space Removal 39
3.4.7 Tokenization and Stop Word Removal 40
3.4.8 Lemmatization 41
3.5 Feature Extraction 43
3.6 Naïve Bayes Model Algorithm 44
3.6.1 Naïve Bayes Training Model 44
3.6.2 Naïve Bayes Testing and Evaluation 51
3.7 Prototype Development 53
3.7.1 User Interface 54
3.7.2 Prototype Requirement 54
3.8 Documentation 56
3.9 Conclusion 56