Sentiment Analysis of Twitter Data in Malay Language (Bahasa Melayu)

NURNAJWA HAZWANI BT NAWI

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS

JUDUL: SENTIMENT ANALYSIS OF TWITTER DATA IN MALAY LANGUAGE (BAHASA MELAYU)

SESI PENGAJIAN : 2017

SAYA NURNAJWA HAZWANI BT NAWI (HURUF BESAR)

mengaku membenarkan tesis (PSM/SRFjaæLDekleF-Ea-lsa$ah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (O

________	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
________	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan Oleh organisasi/badan di mana penyelidikan dijalankan)
/	TIDAK TERHAD	

(TANDATANGAN PENULIS) (TANDATANGAN PENYELIA)

Alamat tetap: 252. Roundabout

Kota. Jalan Salor. 15100. Kota Bharus Kelantan.

DR ABDUL KARIM BIN MOHAMAD
Penyelia
Sarjana Muda (PSM)
29/8/2017.

_______Nama Penye 'a Tarikh.• 02/06/17

Tarikh:

CATATAN:*Tesis dimaksudkan sebagai Laporan Akhir Projek) **Jika tesis ini SULIT atau TERHAD,

Sila lampirkan surat daripada pihak berkuasa.

SENTIMENT ANALYSIS OF TWITTER DATA IN MALAY LANGUAGE (BM)

NURNAJWA HAZWANI BT NAWI

This report is submitted in partial fulfillment of the requirement for the Bachelor of Computer Science (Artificial Intelligence)

FACULTY OF INFORMATION AND COMMUNICATION AND TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I hereby declare that this project report entitled

SENTIMENT ANALYSIS OF TWITTER DATA IN MALAY LANGUAGE (BAHASA MELAYU)

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT: ______________ Date: 29/8/2017

(NURNAJWA HAZWANI BT NAWI)

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of Computer Science (Artificial Intelligence) With Honours.

SUPERVISOR:Date.

R: ______________ 29/8/2017

(DR. ABDUL KARIM BIN MOHAMAD)

DR. ABDUL KARIM BIN MOHAMAD

Timbalan Pengarah
Pusat Pengurusan Strategik, Kuallti & Risiko
Universiti Teknikal Malaysia Melaka

DEDICATION

This project is dedicated to my parents, Nawi bin Othman and Wan Rokiah bt Wan Abdul Kadir and my family, for your love, support and encouragement during the development of this project.

I would like to express my deepest gratitude to my supervisor, Dr. Abdul Karim bin Mohamad, for your guidance and critiques for this project.

Furthermore, I would also like to extend my thanks to my friends, UTeM staff and students who are willing to give their constructive recommendations and knowledge regarding this project.

ACKNOWLEDGEMENTS

I wish to thank Dr. Abdul Karim bin Mohamad for giving me the chance to complete this project. I appreciate your valuable suggestions during the planning and development of this project until the end. I would also like to thank Dr. Abdul Karim for your advice and assistance to keep my progress and schedule. Without your guidance and recommendation, maybe this project cannot finish on time. Besides, I would like to express my gratitude to Dr. Halizah binti Basiron for giving me the idea regarding sentiment analysis and providing a manually labelled training data for this project.

I would like to thank my parents and family as well for being very supportive and giving me the strength to complete this project.

I also would like to thank all my fellow friends that have directly or indirectly guide me in the development of this project.

ABSTRACT

The main motivation of this project is to identify the sentiment values of Twitter data whether it is positive, neutral or negative. Firstly, a set of tweets are labelled manually (using human interpretation) with their sentiments and considered as training data. Then, another set of tweets that is live streaming, are collected based on the text mining on Twitter Streaming API (Application Programming Interface) and python. The tweets are retrieved and saved as a text file and later will be used as a testing set. Testing data will learn from training data's calculation to predict sentiment value. The problem statements of this project are, there are no Twitter dataset corpus available in Malay language with labelled sentiment values. Next, finding a filter to search tweets in Malay language that is from Malaysia. Then, finding a classifier to categorize tweets into positive, neutral or negative. Major challenge of this project is to collect a labelled corpus as a training set. Since there is no labelled Twitter corpus available in Malay language, a database of sentences is manually labelled with sentiments using human interpretation and uses tweet's geo-location to search for tweets posted in Malaysia. At the end of this project, Twitter corpus using Twitter Streaming API able to be collected. Secondly, tweets from Malaysia collected by using tweet's geo location able to be obtained. Thirdly, there will be a Malay dataset, using the decision tree classifier can be categorized according to its sentiment value which are positive, neutral and negative.

ABSTRAK

Motivasi utama untuk menyiapkan projek ini ialah untuk mengenal pasti nilai sentimen ayat-ayat yang terdapat di dalam data Twitter sama ada ia positif, neutral atau negatif. Pertama sekali, satu set data Twitter dilabel terlebih dahulu dengan nilai sentimen secara manual (interpretasi manusia) dan dinamakan sebagai data latihan. Kemudian, satu set data Twitter yang lain, dikumpul dengan menggunakan perlombongan teks (data mining) berdasarkan "Twitter Streaming Application Programming Interface" (Aplikasi Pengaturcaraan Antaramuka) dan bahasa python. Data Twitter diperolehi dan disimpan sebagai fail teks dan kemudian digunakan sebagai data ujian. Data ujian akan mempelajari daripada data latihan untuk meramal nilai sentimen. Pernyataan masalah projek ini ialah ketiadaan set data Twitter dengan nilai sentimen yang sudah siap dilabel. Seterusnya, mencari penyaring (filter) untuk mengkategorikan data Twitter kepada positif, neutral atau negatif. Cabaran utama projek ini ialah mengumpul data Twitter yang berlabel sebagai data latihan. Disebabkan tiada data Twitter Bahasa Melayu yang sudah siap dilabel, satu pangkalan data dilabel secara manual dengan nilai sentimen dari interpretasi manusia dan menggunakan lokasi geografi data Twitter untuk mencari data Twitter dalam Bahasa Melayu dari Malaysia. Di akhir projek ini, data Twitter dapat dikumpul. Kedua, data Twitter dalam Bahasa Melayu dari Malaysia dapat diperolehi. Ketiga, dapat mencari sebuah pengkelas (classifier) yang dapat mengelaskan data Twitter kepada nilai sentimen iaitu positif, neutral dan negatif.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Internet today has no limitation to the users. Everything is only with a touch of a finger. The internet has proven to be useful and has come with a lot of advantages and a lot of disadvantages too. To determine the emotion of a person by their writings is a challenging task in developing sentiment analysis. Sentences are taken from Twitter live tweets using text mining of Twitter Streaming API. Text mining is the application of natural language processing techniques plus analytical method to collect meaningful tweets. An API is a tool to enable computer programs and web services to communicate. In this project, tweets undergo preprocessing phase where the tweets are filtered to remove unnecessary strings such as emoticons and http links. Besides, tokenization is performed on the tweets to divide the text by spaces and punctuation marks. The tweets will be broken down and each will be labelled as positive, neutral or negative based on the data dictionary which contains words with sentiment values. If there is no pairing match, the word will be considered as positive. Apart from that, stop words such as 'saya', 'ialah', 'yang' are removed from the tweets and considered as a positive word. In this project, decision tree classifier is used where information content, information gain and decision tree are first calculated manually and then RapidMiner tool is used to compare the results of using Decision Tree classifier. The program uses a data dictionary of Malay words labelled as positive, neutral and negative sentiment value. A Malay language dataset is manually labelled with their sentiment values using human interpretation to be used as a training set.

1.2 Problem Statements

The problem statements of this project are as follows:

- No labelled Twitter corpus available in Malay language.
- Finding a method to filter tweets in Malay language originated from Malaysia
- Finding a classifier to categorize tweets to positive, neutral and negative.

1.3 Objectives

The objectives of this project are as follows:

- To investigate how to collect tweets on Twitter using Twitter Streaming API.
- To study how to use tweet's geo location to search for tweets in Malay language originated from Malaysia.
- To identify how to categorize Twitter dataset into their sentiment values using Decision Tree.

1.4 Scope

The scope of this project is to find tweets with labelled sentiment to be used as training data. Apart from that, to find tweets in Malay language which is from Malaysia and lastly is to find classifier to calculate the sentiment values of the tweets. Developing this project must achieve the following scopes so that the project is a success.

1.5 Project Significance

This project study on the method to calculate the sentiment value of Twitter data in Malay language (Bahasa Melayu). The significance of sentiment analysis is that it can be a good source of information and can supply a model that is beneficial to companies such as improving the quality of a product or service. Besides, it can prove whether a campaign is a success or not plus improve strategy making.

1.6 Expected Outcome

The expected output for this project is live streaming Twitter dataset can be collected using Twitter Streaming API and processed to be used as testing set. Besides, manually labelled Malay language dataset can be created to be considered as training set. Apart from that, tweets can be collected in Malay language originated from Malaysia. Moreover, a classifier or Decision Tree will be able to execute in determining sentiment values of Twitter data.

1.7 Conclusion

All in all, project introduction, problem statements, objectives, scope, project significance and expected outcome have been stated clearly in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews studies on previous works of accredited researchers or scholars on topics such as Twitter, decision tree and supervised learning, that will be explained in the next subtopic. It also presents the idea on Machine Learning as in general, knowledge representation and ID3. This chapter is act as an improvement from the earlier research to produce better output and performance.

2.2 Facts and Finding

2.2.1 Twitter

Twitter is a social media application that has been launched since 2006. With social networking use surpassing web-based email use in February 2009 (Wilson, 2009), a few of the connection are just not being created by human. Businesses and organizations are also taking the opportunity to connect with their customers and other people online. Business are making advantage of the social networking medium to search for chances, promotions, workers, and details on how customers make use of their products and services (Wilson, 2009).

2.2.2 Machine Learning – general, knowledge representation, focus on ID3 – decision tree, information content, information gain

i. Machine Learning

Supervised learning used existing machine learning methods to carry out sentiment analysis. It includes constructing classifiers from responses (movie reviews for instance) are used as training and testing set, (Pang, Lee & Vaithyanathan, 2002). Machine learning methods that usually used are Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME) and k-Nearest Neighbour (kNN). This method is initialized with data collection. Data will be divided into training data and testing data. The training data is used for classifier learning process and testing data is used to test the classifier's performance after learning process is done. Feature selection is the procedure to select a set of attributes or features that suits the process mining. (Samsudin, Puteh, Hamdan & Ahmad, 2013). According to (Pang, Lee & Vaithyanathan, 2002), they applied Support Vector Machine, Naïve Bayes and Maximum Entropy for movie reviews. Training of classifiers are based on unigrams and bigrams features. Results gained shows that training data size can affect the classifiers' performance. Naïve Bayes works best for smaller training data, however for larger training data, Support Vector Machine (SVM) have the best performance compared to Naïve Bayes and Maximum Entropy.

ii. Knowledge Representation

Knowledge representation and machine learning are the major contribution of tweet sentiment analysis. In knowledge representation method, a comprehensive database is needed which contains labelled sentiment values to identify sentiments. Machine learning method uses training set to classify the sentiment value of each words accurately. This method does not need a database the same as knowledge representation, which is a good thing. By using these techniques, hybrid model is produced which able to perform sentiment analysis of almost any text given. Knowledge representation method is quite difficult due to the need of large lexical database.

iii. ID3: decision tree, information content, information gain

ID3 is a commonly used decision tree. Below is the example on how ID3 is conducted.

Formula for Entropy:

$$\boldsymbol{Ent}\left(\frac{\boldsymbol{p}}{\boldsymbol{p}+\boldsymbol{n}},\frac{\boldsymbol{n}}{\boldsymbol{p}+\boldsymbol{n}}\right) = -\frac{\boldsymbol{p}}{\boldsymbol{p}+\boldsymbol{n}}\log_2\frac{\boldsymbol{p}}{\boldsymbol{p}+\boldsymbol{n}} - \frac{\boldsymbol{n}}{\boldsymbol{p}+\boldsymbol{n}}\log_2\frac{\boldsymbol{n}}{\boldsymbol{p}+\boldsymbol{n}}$$

Figure 2.1: Entropy Formula

Formula for Information Content:

$$\boldsymbol{I}\left(\frac{\boldsymbol{pos}}{\boldsymbol{pos}+\boldsymbol{neg}},\frac{\boldsymbol{neg}}{\boldsymbol{pos}+\boldsymbol{neg}}\right) = -\frac{\boldsymbol{pos}}{\boldsymbol{pos}+\boldsymbol{neg}}\log_2\frac{\boldsymbol{pos}}{\boldsymbol{neg}} - \frac{\boldsymbol{neg}}{\boldsymbol{pos}+\boldsymbol{neg}}\log_2\frac{\boldsymbol{neg}}{\boldsymbol{pos}+\boldsymbol{neg}}$$

Figure 2.2: Information Content Formula

Formula for Information Gain:

$$\textbf{Gain (A)} = \boldsymbol{I}\left(\frac{\boldsymbol{positive}}{\boldsymbol{negative}}\right) - \boldsymbol{Remainder}(\boldsymbol{A})$$

Figure 2.3: Information Gain Formula

Below is the sample token table to build a decision tree:

Table 2.1: Token Table

Sample	Token1	Token2	Token3	Token4	Sentiment
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1
X	1	1	1	1	?
Y	0	1	0	1	?

Z	1	1	0	0	?
Information Content					0.9544
Information Gain	0.0032	0.0032	0.0032	0.0488	

Based on the table 2.1, a set of sentences are broken down into five tokens where 0 is negative sentiment value and 1 is positive sentiment value. The information content of the Sentiment attribute is 0.9544. Token4 has the highest information gain, so it is chosen as the root node. Since the information gain of token1, token2 and token 3 are the same, either one of the three can be chosen. Token2 is the best in distinguishing class 0 and 1 among the three tokens thus is chosen as the next node.

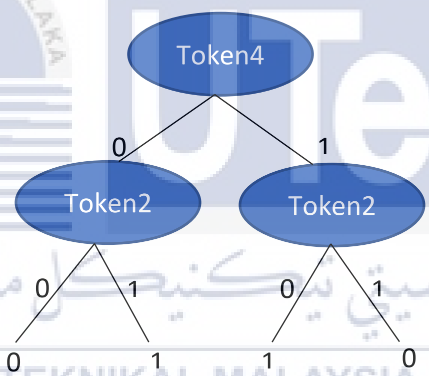


Figure 2.1: Decision Tree

Figure 2.1 shows the decision tree that will be used and referred as to predict the sentiment value of the last three unknown sentiment value.

Table 2.2: Sentiment Result

Token4	Token2	Sentiment
1	1	0
1	1	0
0	1	1

According to the decision tree constructed in Figure 2.1, when the value for the token4 is 1 and the value for token2 is 1, the sentiment value gained is 0. The result is the same for the second prediction. For the third prediction, when token4 has the value of 0 and token2 has the value of 1, the sentiment value gained is 1.

2.2.3 Tweet – Decision Tree

A Decision Tree is a tree where nodes are labelled based on the attributes, the edges leaving a node are labelled by tests on the attribute's weight, and the leaves are labelled by classes (Feldman & Sanger 2007: 72). It categorizes a document by initializing at the tree root and moving successfully downward through the branches (whose conditions are fulfilled by the document) until a leaf node is reached (Feldman & Sanger 2007: 72-73). The document is then classified in the class that labels the leaf node (Feldman & Sanger 2007: 73). Decision trees have already been used in various applications such as speech and language processing (Jurafky & Martin 2009: 247). For trees to have properties such as minimality, ID3 ranks the features of a training data according to the information gain.

2.2.4 Supervised learning (Relate ID3 with sentiment analysis)

Supervised learning one of the machine learning task of building a function from labeled training data. The training data includes a set of training samples. In supervised learning, each example is a pair consisting of an input object and a desired output value. A supervised learning algorithm perform analyzation o the training data and produces an inferred function, which can be used for mapping new examples. The main disadvantage of using supervised method is that the classifiers' performance depends on the training data. Larger and higher quality training data produce better classification. Information lacking and minimal information of training data can result in misclassification.

2.3 Conclusion

This chapter has stated the literature review such as Twitter, machine learning, transforming tweet to decision trees and supervised learning clearly in the project. Methodology will be discussed more in the next chapter.

CHAPTER 3

PROJECT METHODOLOGY AND DESIGN

3.1 Introduction

In this chapter, the methodology and design of the project is discussed. Methodology is the systematic steps and theoretical used in developing this project. Some of the methodology used are data collection, data analysis, experimental design, evaluation and testing, and lastly, conclusion.

Besides, design is the process flow of the project or the steps required for in developing a project. By having this process, it will keep the project on track to achieve high accuracy and quality.

3.2 Methodology

There are many of system development model that can be applied such as waterfall model, rapid prototyping and agile model. Waterfall model is selected for this project which contains five phases to be developed. The five phases involved are data collection, data analysis, experimental design, evaluation & testing and conclusion. Every stage is vital and must be done phase to phase to produce a high-quality result.

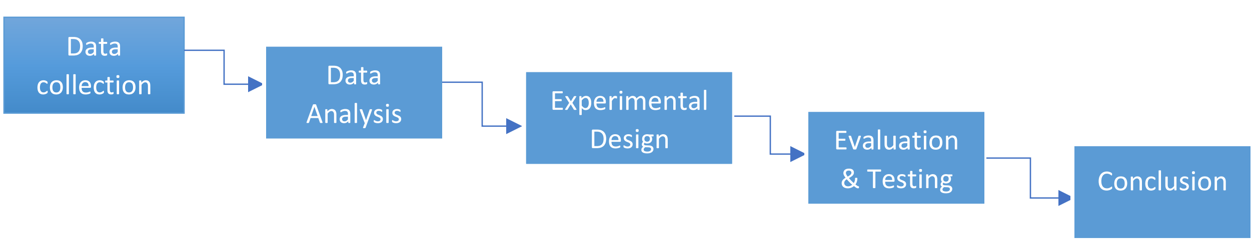


Figure 3.1: Waterfall model

3.2.1 Phase 1: Data collection

Phase 1 focused on collecting data which is needed to perform the sentiment analysis of Twitter data in Malay language (Bahasa Melayu). Methods need to be executed are as following:

- **Define the goal of the collecting data**
  To have a goal, researchers must always focus on their problem statements and objectives so that the project can be successful.
- **Declare principle**
  Create a principle before collecting data so that unnecessary procedures would not be taken and affect the quality and the outcome of the data.
- **Initialize collection of data**
  Start the process and follow the principle that have been stated. Every step are vital thus need to be recorded in case there is a need of reference and documentational purpose.
- **Observe the quality pattern of the data**
  Check the quality of data collected by referring to the problem statements and objectives. If the quality is not fulfilling the objectives of the project, the data collection phase must be repeated.

The first phase is about collecting testing data which is collecting Twitter data using Twitter Streaming API. The steps taken to complete this phase are:

i. **Sign up a Twitter account**

There are 4 keys needed to run Twitter Streaming API which are API key, API secret, access token and access token secret. To obtain the keys, a Twitter account must be created.

ii. **Insert keys obtained into the algorithm to download live streaming Tweets**

The keys obtained will be used in the coding to retrieve live streaming tweets.

The information details of the downloaded tweets are as following:

- *text*: Twitter text
- *created_at*: creation date
- *favorite_count, retweet_count*: amount of favourites and retweet
- *favorited, retweeted*: Boolean type stating whether the user of this Twitter account has favourite or retweet the tweets
- *lang:* language ancronym (e.g. “en” for english)
- *id:* identifier of the tweet
- *place, coordinates, geo*: geo-location details (where available)
- *user:* profile of the author
- *entities*: entities list such as URL, @-mentions, symbols and hashtags
- *in_reply_to_user_id:* user identifier if the tweet is a reply to a specific user
- *in_reply_to_status_id*: status identifier id the tweet is a reply to a specific status

The challenging part in this phase is to find the training data. Since there is no available Malay language dataset with labelled sentiment, a database of sentences is manually labelled with their sentiments using human interpretation.

3.2.2 Phase 2: Data analysis

In analyzing the data, tweets undergo preprocessing phase where they followed the steps below:

i. **Filtering**

First, text attribute is extracted from the tweet. Other attributes such as *user* attribute, *created_at* attribute and *language* attribute is not needed in the process to determine the sentiment of the tweets. Then, tweets are filtered to remove unnecessary strings such as emoticons and http links.

ii. **Tokenization**

Tweets are tokenized, broken down into words to divide the text by spaces and punctuation marks. Each word will be labelled as positive, neutral or negative based on the data dictionary which contain words with their respective sentiment values. If there is no pairing match, the word will be considered as positive.

iii. Stemming

Stemming is removing the suffix and prefix such as '–lah', '-nya', '-kan', from a word, to gain only the root word. The reference of the dictionary of stemmer is based on (Muhamad, Fatimah, Ramlan & Tengku, 2006) which is also referred as Fatimah Stemmer.

iv. Removing stop words

Stopwords such as 'saya', 'dia', 'yang' are eliminated from the tweets and the space left is considered as positive word. The reference of the dictionary of stop words is from (Muhamad, Fatimah, Ramlan & Tengku, 2006) which is also referred as Fatimah's algorithm.

After preprocessing phase is done, tweets become more meaningful and ready to be experimented.

3.2.3 Phase 3: Experimental design

i. Building decision Tree

In this project, decision tree is chosen as the classifier to predict the sentiment values of testing data. To create a decision tree, information content or can be called as entropy, of the sentiment column must be calculated first. By having the entropy, information gain can be search for all the tokens. A decision tree can be formed after all information is calculated. By producing the decision tree, we can predict the sentiment values of tweets in the testing set.

3.2.4 Phase 4: Evaluation and testing

Based on the decision tree in phase 3, the sentiment values of the same testing data will be calculated by using RapidMiner Studio tool. This phase is to compare whether using decision tree classifier from human interpretation is as accurate as using a machine tool.

3.2.5 Phase 5: Conclusion

Conclusion will be made after the result is obtained. This includes the best method for this project besides revealing the outcome of the project.

3.3 Project Schedule and Milestone

Table shown below is the milestone of the project.

Table 3.1: Project Milestone for Final Year Project 1

Week	Activity	Note / Action
1 13 – 17 Feb **Meeting 1**	Proposal PSM: Discussion & Submission using PSM Online System	Deliverable – **Proposal** Action – Student
	Proposal assessment & verification	Action – Supervisor, Evaluator
2 20 – 24 Feb	Proposal Correction/Improvement	Action – Student
	List of supervisor/title	Action – PSM/PD Committee
3 27 Feb – 3 Mar **Meeting 2**	Proposal Presentation Chapter 1 (System Development Begins)	Deliverable – **Proposal Presentation (PP)** Action – Student
4 6 – 10 Mar	Chapter 1 Chapter 2	Deliverable – **Chapter 1** Action – Student, Supervisor
5 13 – 17 Mar	Chapter 2	Action – Student
6 20 – 24 Mar **Meeting 3**	Chapter 2 Chapter 3	Deliverable – **Chapter 2** **Progress Presentation 1 / Pembentangan Kemajuan 1 (PK 1)** Action – Student, Supervisor
	Student Status	Warning Letter 1 Action – Supervisor, PSM/PD Committee

7 27 – 31 Mar	Chapter 3 Chapter 4	Action – Student
8 3 – 7 Apr	**MID SEMESTER BREAK**	
9 10 – 14 Apr	Chapter 4 Project Demo	Deliverable: **Chapter 3** Action – Student, Supervisor
10 17 – 21 Apr **Meeting 4**	Chapter 4 Project Demo	Deliverable – **Progress Presentation 2 / Pembentangan Kemajuan 2 (PK 2)** Action – Student, Supervisor
	Student Status	Warning Letter 2 Action – Supervisor, PSM/PD Committee
11 24 – 28 Apr **Demonstration**	Project Demo	Action – Student
	Determination of student status (Continue/Withdraw)	Sumbit student status to Committee Action – Supervisor, PSM/PD Committee
12 1 – 5 May	Project Demo PSM 1 Report	Action – Student, Supervisor
13 8 – 12 May **Meeting 5**	Project Demo PSM 1 Report	Action – Student, Supervisor
	Presentation schedule	Action – PSM/PD Committee
14 15 – 19 May	Project Demo PSM 1 Report	Deliverable – **Complete PSM 1 Draft Report** Action – Student, Supervisor
15 22 – 26 May **Final Presentation**	**FINAL PRESENTATION & PROJECT DEMO**	Action – Student, Supervisor, Evaluator
16 29 May – 2 Jun	**REVISION WEEK** Correction on the draft report based on the comments by the Supervisor and Evaluator during the final presentation session Submit PSM 1 Logbooks to PSM Online System	Deliverable – **Complete PSM 1 Logbooks** Action – Student, Supervisor
	Submission of overall marks to PSM/PD committee	Deliverable: **Overall PSM 1 score sheet** Action – Supervisor, Evaluator, PSM/PD Committee
17 & 18 5 – 18 Jun	**FINAL EXAMINATION WEEKS**	

Table 3.2 Project Milestone for Final Year Project 2

Week	**Activity**	**Note / Action**
1 **Meeting 1**	Chapter 4 Chapter 5	Deliverable – **Chapter 4** Action – Student, Supervisor
2 **Meeting 2**	Chapter 5 Project Demo	Deliverable – **Progress Presentation 1 / Pembentangan Kemajuan 1 (PK 1)** Action – Student, Supervisor
3	Chapter 5 Chapter 6	Deliverable – **Chapter 5** Action – Student
	Student Status	Warning Letter 1 Action – Supervisor, PSM/PD Committee
4 **Meeting 3**	Chapter 6 Project Demo	Deliverable – **Progress Presentation 2 / Pembentangan Kemajuan 2 (PK 2)** Action – Student, Supervisor
5 **Meeting 4**	Chapter 6 Chapter 7	Deliverable – **Chapter 6** Action – Student, Supervisor
	Presentation Schedule	Action – PSM / PD Committee
	Student Status	Warning Letter 2 Action – Supervisor, PSM / PD Committee
6 11 **Meeting 5**	Chapter 7 Project Demo PSM2 Report	Deliverable – **Chapter 7 & Complete PSM2 Draft Report** Action – Student, Supervisor
	Determination of student status (Continue / Withdraw)	Submit student status to committee Action – Supervisor, PSM / PD Committee
7 **Final Presentation**	**FINAL PRESENTATION & PROJECT DEMO**	Action – Student, Supervisor, Evaluator & PSM / PD Committee

8	**FINAL EXAMINATION WEEK** Correction on the draft report based on the comments by the Supervisor and Evaluator during the final presentation session Submit PSM2 Logbooks to PSM Online System	Deliverable – **Complete PSM2 Logbooks** Action – Student, Supervisor
	Submission of overall marks to PSM / PD Committee	Deliverable – **Overall PSM2 Score Sheet** Action – Supervisor, Evaluator, PSM / PD Committee
9	**INTER-SEMESTER BREAK** Submission of the final complete report, which is the updated & corrected PSM2 report, onto the PSM e-Repository online system	Deliverable: **Complete Final PSM Report** Action – Student, Supervisor

3.4 Design on Sentiment Analysis using Twitter Data in Malay language (Bahasa Melayu)

Figure shown below is the flowchart of the project:

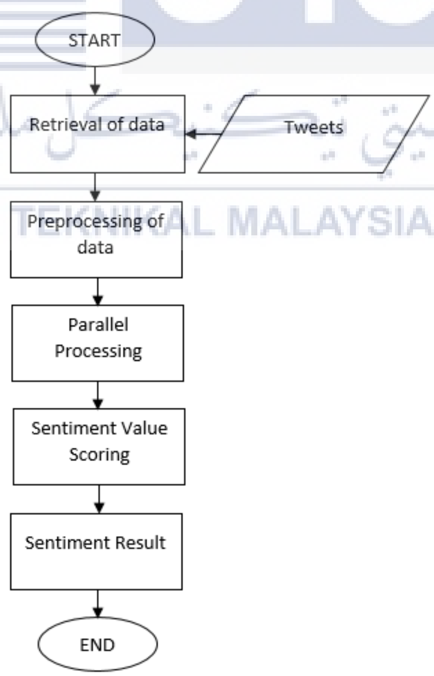


Figure 3.2: Project Flowchart

Based on Figure 3.2, the project is initialized with retrieving the data from Twitter that collect tweets as a major input. After gaining the dataset, it is divided into training and testing data. Training data is the data with manually labelled sentiment values using human interpretation while testing data learns from the training data to predict the sentiment values of new test set of tweets. Both data undergo preprocessing phase to make the data more meaningful. Steps required in preprocessing phase is filtering, tokenization, stemming and removing stop words. After that, data undergo parallel processing, which apply classifier. In this project case, classifier used is Decision Tree. From the decision tree, the sentiment value scoring is acquired and can be referred for testing set. Based on scores, the sentiment result of the testing set can be predicted.

CHAPTER 4

IMPLEMENTATION

4.1 Introduction

This chapter explains on procedures to be taken to produce the required outcome. The procedure must be followed step by step carefully. This is to prevent from receiving inaccurate and imprecise result. This chapter is also a demonstration on how decision tree is built to calculate the sentiment values of Twitter data. In this project, two decision trees were built to observe the performance of decision trees. The precision of both decision trees is also stated in this chapter. From the precision percentage, the best decision tree can be proved.

4.2 Project Requirements

For this experiment to be a success, both hardware and software are used.

Table 4.1: Project Requirements

Hardware	Software
Personal computer with minimum specification of 4GB RAM and intel i5 processor	- Microsoft Excel - RapidMiner Studio

4.3 Project Results

There are 3282 of input data retrieved from Twitter by using Twitter Streaming API. However, after the preprocessing phase, only 1000 meaningful tweets are chosen, 500 is treated as training data and another 500 is treated as testing data. The first step to retrieve the data tweets is signing up for a Twitter account to receive required keys. 4 keys that are needed in downloading live streaming tweets are *API key, API secret, access token and access token secret.*

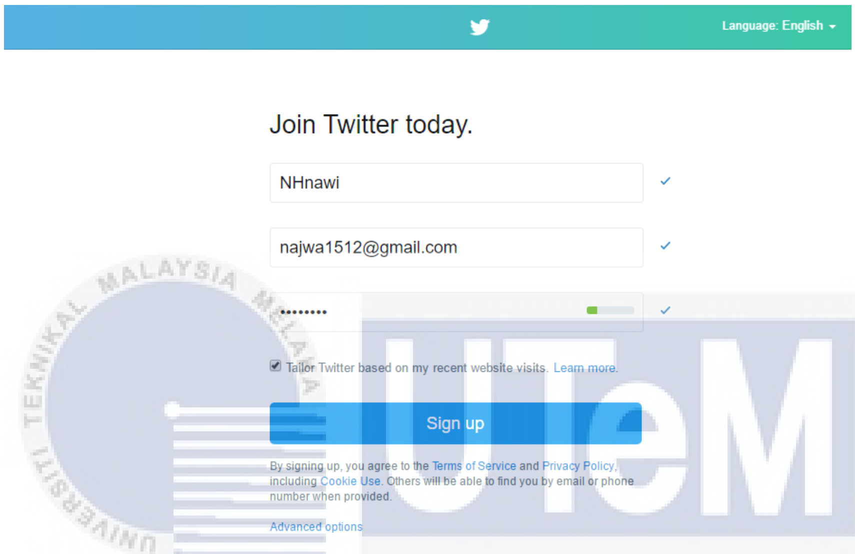


Figure 4.1: Twitter Sign-Up

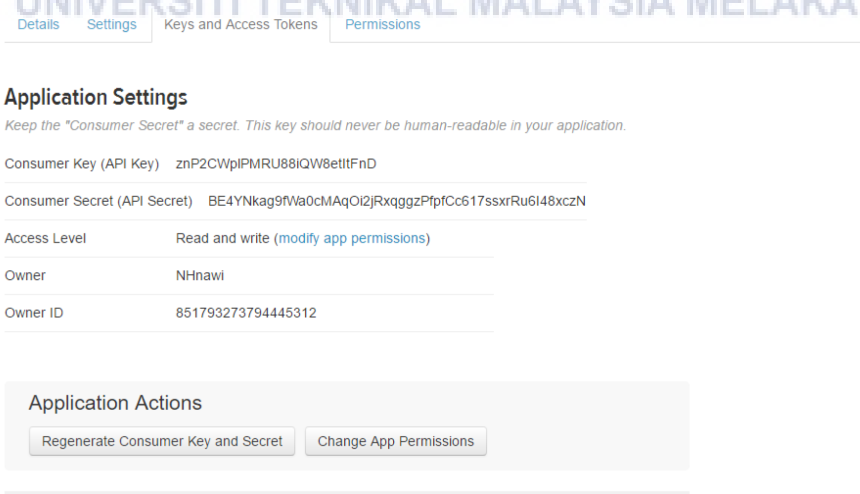


Figure 4.2: Consumer Key (API key) & Consumer Secret (API Secret)

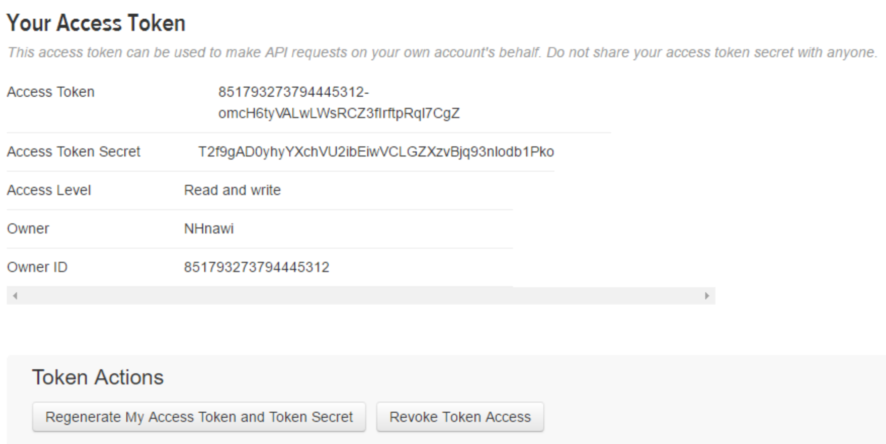


Figure 4.3: Access Token & Access Token Secret

ii. Below is the coding to download the tweets, that is saved as *tweet_streaming.py*.

Table 4.2: Live Streaming Tweets Algorithm

```
#Import the necessary methods from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import json
import re
import pandas as pd
import matplotlib.pyplot as plt

#Variables that contains the user credentials to access Twitter API
access_token = "851793273794445312-omcH6tyVALwLWsRCZ3fIrftpRql7CgZ"
access_token_secret = "T2f9gAD0yhyYXchVU2ibEiwVCLGZXzvBjq93nlodb1Pko"
consumer_key = "znP2CWplPMRU88iQW8etItFnD"
consumer_secret = "BE4YNkag9fWa0cMAqOi2jRxqggzPfpfCc617ssxrRu6I48xczN"
```

```
#This is a basic listener that just prints received tweets to stdout.
class StdOutListener(StreamListener):

    def on_data(self, data):
        json_load = json.loads(data)
            texts = json_load['text']
            coded = texts.encode('utf-8')
            s = str(coded)
            print s
            #print(s[2:-1])
            return True

    def on_error(self, status):
        print status


if __name__ == '__main__':

    #This handles Twitter authetification and the connection to Twitter Streaming API
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    #This line filter Twitter Streams to capture data by location: Malaysia
    stream.filter(locations=[98.94,0.85,119.4,7.52],languages=['in'])

tweets_data_path = stream.filter

tweets_data = []
```

```
tweets_file = open(tweets_data_path,"r")
for line in tweets_file:
        try:
            tweet = json.loads(line)
            tweets_data.append(tweet)
        except:
            continue
print len(tweets_data)

tweets['text']= map(lambda tweet: tweet['text'], tweets_data)
wiki = TextBlob(tweets['text'])
r = wiki.sentiment.polarity

print r
```

iii. **Enter *python tweet_streaming.py* to run the coding in the terminal. After that, it will produce data such as the figure below.**

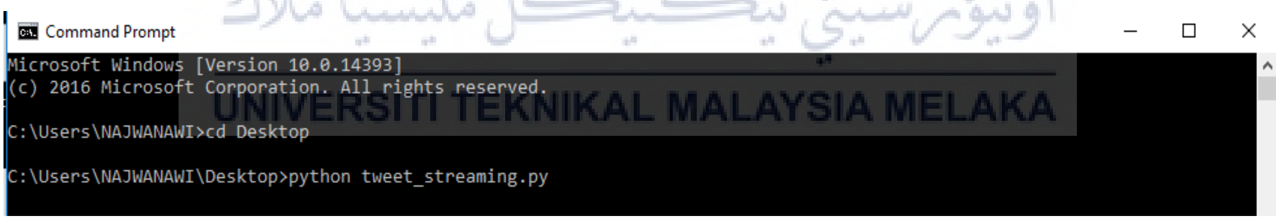


Figure 4.4: Command using Terminal

```
{"created_at":"Tue May 23 14:16:15 +0000 2017","id":867021360500887552,"id_str":"867021360500887552","text":"Cintaaa bukan hanya harapan","source":"\u003ca href=\"http:
\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":
null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":579071173,"id_str":"579071173","name":"JOM KURUS!","screen_na
me":"akrnjw","location":"Johor Bahru, Johor","url":null,"description":"Safwan \ud83d\udc9e\u2728","protected":false,"verified":false,"followers_count":2114,"friends_cou
nt":701,"listed_count":1,"favourites_count":5151,"statuses_count":54108,"created_at":"Sun May 13 16:00:52 +0000 2012","utc_offset":28800,"time_zone":"Beijing","geo_enab
led":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"070808","profile_background_image_url":"http:\/\/pbs.twimg.com\/pro
file_background_images\/485072537101946880\/Yc8nhXMF.jpeg","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/485072537101946880\
/Yc8nhXMF.jpeg","profile_background_tile":true,"profile_link_color":"0D0C0C","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"FFFFFF","profile_text
_color":"000000","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/866607374613835776\/PtIggsAZ_normal.jpg","profile_imag
e_url_https":"https:\/\/pbs.twimg.com\/profile_images\/866607374613835776\/PtIggsAZ_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/57907117
3\/1495148396","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"pl
ace":{"id":"59e757bb93090b8c","url":"https:\/\/api.twitter.com\/1.1\/geo\/id\/59e757bb93090b8c.json","place_type":"city","name":"Plentong","full_name":"Plentong, Johor"
,"country_code":"MY","country":"Malaysia","bounding_box":{"type":"Polygon","coordinates":[[[103.701431,1.426407],[103.701431,1.624400],[103.974757,1.624400],[103.974757
,1.426407]]]},"attributes":{}},"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[],"
symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"in","timestamp_ms":"1495548975478"}
```


Figure 4.5: Live Stream Data Output

The output which is in json format and is saved as text file and undergo preprocessing phase. They are filtered to get only the *text* attribute which is needed for sentiment analysis. Other attributes are not needed to calculate sentiment value. After retrieving the tweets from Malaysia, there is also no need to collect location attribute to perform sentiment analysis. For example, the text attribute in this output is "*Cintaaa bukan hanya harapan*".

Training data sample

Table 4.3: Training Data Sample

Token1	**Token2**	**Token3**	**Token4**	**Token5**	**Sentiment**
Saya	Di	Rumah	Seri	Kenangan	**1**
Amek	cik	Binik	Lagi	Kerja	**1**
Mati	Sangat	Keraskah	Kena	Buat	**0**
Majlis	Daerah	Hulu	Langat	Selangor	**1**
Apa	Itu	Majlis	Daerah	0	**1**
Duit	Hadiah	Yang	Diambil	Dari	**1**
aku	Ada	Dengar	Dua	Ipoh	**1**
Liverbird	Tirf	Apa	Lagi	Yang	**1**
Aku	Rasa	Aku	Nak	Perkhidmatan	**1**
Tapi	Tapi	0	0	0	**1**
Aktif	Ciri	Itu	Kena	Iaitu	**1**
Dan	Satu	Perkara	Yang	Aku	**1**

Aku	Tidak	Akan	Ada	Lagi	**0**
Jangan	Takut	Jatuh	Hati	Mesti	**0**
ini	lawak	jangan	Marah-marah	0	**0**

Table 4.3 shows the training data consisting of 5 tokens and its overall sentiment value. Sentiment with the value 1 is a positive sentiment value while sentiment with the value 0 is a negative sentiment value.

Testing data sample

Table 4.4: Testing data sample

Token1	**Token2**	**Token3**	**Token4**	**Token5**
Selebriti	Yang	Menyokong	Liverpool	daniel
Aku	dah	Nampak	Dah	Bayangan
Selamat	Hari	Jadi	Lejen	Terima
Baru	Ikut	Instagram	Ustaz	Awan
Tweepy	Yang	Sudah	Bungkus	Tidak
Tweepy	Sikit	0	0	0
Ayam	Sudah	0	0	0
Percayalah	sayang	Ijat	Ketat-ketat	0
Makan	Nasi	Kandang	0	0
Tengah	bungkus	Nasi	Kandang	Beratu
Naik	raya	kenalah	Rambut	Baru
Saya	Hanya	Dikeluarkan	Nora	Azlina
Masukkan	Mylfc	Ini	Sentiasa	Kemas
Kirim	Salam	Imam	Sahak	Anak-anak
Sudah	di	polowin	0	0

Table 4.4 shows the testing data consisting of 5 tokens without the sentiment values. Testing data will learn from training data to predict sentiment value.

Training Data-1:

Table 4.5: Training Data-1

Tweets	Token1	Token2	Token3	Token4	Sentiment
1	0	0	0	0	**0**
2	0	0	1	0	**0**
3	1	1	0	1	**0**
4	1	0	0	1	**1**
5	0	1	1	0	**1**
6	0	0	1	1	**1**
7	0	0	0	1	**1**
8	1	1	0	0	**1**
9	1	1	1	1	**0**
10	0	1	0	1	**0**
11	1	1	0	0	**1**
12	0	0	1	1	**1**
13	1	1	1	1	**1**
14	1	0	0	1	**1**
15	0	1	1	1	**1**
16	0	1	1	0	**1**
17	0	1	0	1	**1**
18	1	0	0	0	**0**
19	1	1	1	0	**1**
20	0	1	0	1	**0**
21	1	0	1	1	**1**
22	1	1	1	0	**1**
23	0	1	0	1	**0**
24	1	1	1	0	**1**
25	1	1	0	1	**1**
26	1	0	1	0	**0**
27	0	1	1	0	**1**
28	1	0	0	1	**0**
29	1	0	1	1	**1**
30	0	1	1	1	**1**
31	1	1	1	0	**1**
32	1	1	0	1	**1**
33	1	0	1	1	**1**

34	1	1	0	0	**0**
35	0	1	1	0	**0**
36	1	1	0	1	**1**
37	1	1	1	1	**1**
38	1	1	1	0	**1**
39	1	1	0	1	**1**
40	0	0	1	1	**1**

Table 4.5 show the Training Data–1 which includes 4 tokens and its overall sentiment value. Sentiment having value 1 is a positive sentiment value, sentiment having value 0 is a negative sentiment value.

Decision Tree ID3-1

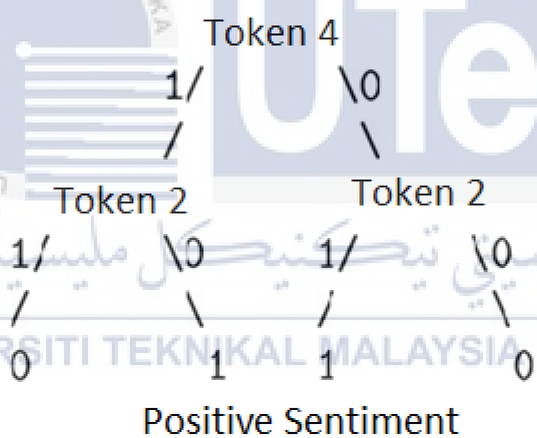


Figure 4.6: Decision Tree ID3-1

Figure 4.6 shows the decision tree of Training Data-1. This decision tree will be tested on testing data.

Training Data-2:

Table 4.6: Training Data-2

Tweets	Token1	Token2	Token3	Token4	Sentiment
1	1	-1	-1	-1	**-1**
2	1	0	-1	0	**-1**
3	0	0	-1	0	**0**
4	0	-1	-1	-1	**-1**
5	1	0	-1	0	**-1**
6	1	-1	0	0	**0**
7	0	0	0	0	**0**
8	0	0	1	1	**1**
9	1	1	1	0	**1**
10	0	1	1	1	**1**
11	1	-1	1	0	**1**
12	1	0	1	1	**1**
13	0	1	0	0	**0**
14	-1	0	-1	1	**-1**
15	1	0	-1	1	**-1**
16	0	1	0	1	**0**
17	1	0	-1	0	**-1**
18	-1	1	-1	1	**-1**
19	0	0	-1	1	**-1**
20	0	-1	-1	0	**-1**

Table 4.6 shows the Training Data-2 which includes 4 tokens and its overall sentiment value. Sentiment with value 1 is a positive sentiment value while sentiment with value 0 is a neutral sentiment value and -1 is a negative sentiment value.

Decision Tree ID3-2:

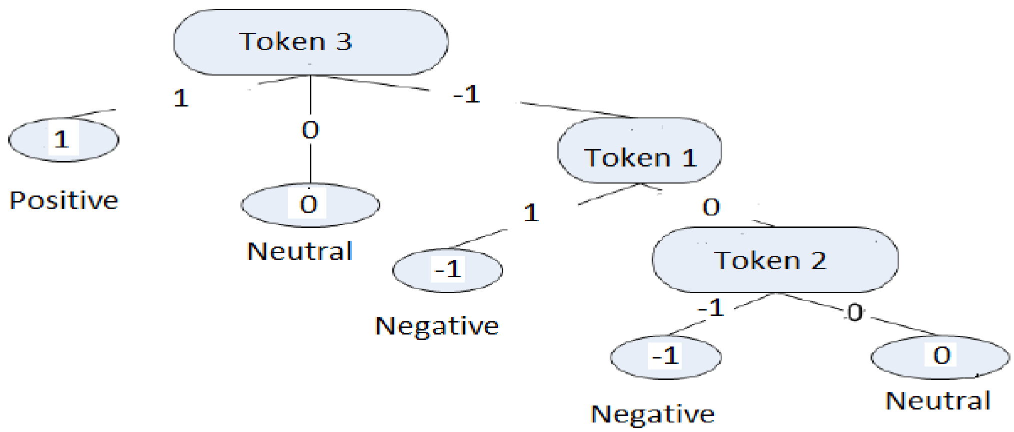


Figure 4.7: Decision Tree ID3-2

After constructing the two decision trees, the training data is tested using the calculation of two decision tree stated. It is to compare the original sentiment value and the sentiment value both decision trees predict.

Table 4.7: Sentiment Calculation

												Precision Check		
No. of Tweets	T1	T2	T3	T4	T5	Sentiment	ID3-1	ID3-2	SIM SUM	$\frac{sim\ sum}{15}$	Value	ID3-1	ID3-2	SIM SUM
1	0	0	1	1	1	1	1	1	3	0.6	1	T	T	T
2	0	0	0	1	1	0	1	0	2	0.4	0	F	T	T
3	1	0	0	0	0	0	0	0	1	0.2	0	T	T	T
4	1	1	1	1	1	1	0	1	5	1	1	F	T	T
5	0	1	1	1	1	1	0	1	4	0.8	1	F	T	T
6	1	1	1	1	1	1	0	1	5	1	1	F	T	T
7	1	0	1	1	0	0	1	1	3	0.6	1	F	F	F
8	0	0	1	1	1	1	1	1	3	0.6	1	T	T	T
9	1	1	1	0	1	1	1	1	4	0.8	1	T	T	T
10	0	0	1	1	1	0	1	1	3	0.6	1	F	F	F
11	1	1	1	1	1	1	0	1	5	1	1	F	T	T

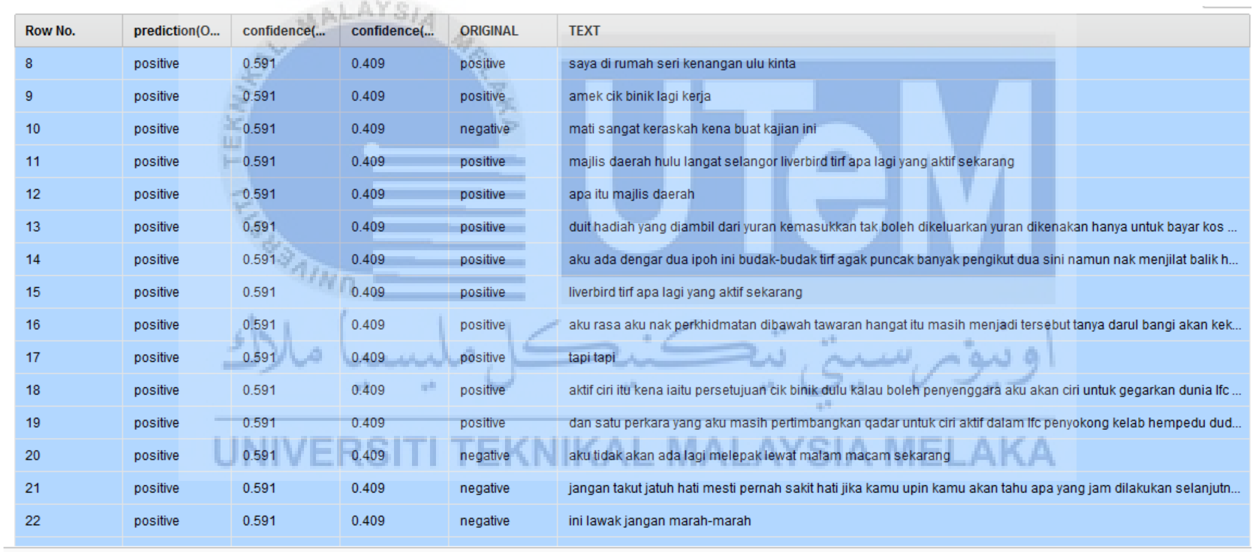
12	0	0	1	1	1	1	1	1	3	0.6	1	T	T	T
13	1	1	0	1	0	1	0	0	3	0.6	1	F	F	T
14	0	0	1	1	1	1	1	1	3	0.6	1	T	T	T
15	1	1	0	0	0	1	1	0	2	0.4	0	T	F	F
16	1	1	1	1	1	1	0	1	5	1	1	F	T	T
17	1	0	0	1	1	1	1	0	3	0.6	1	T	F	T
18	1	-1	-1	1	0	0	1	-1	0	0	0	F	F	T
19	-1	-1	-1	0	1	0	1	-1	-2	-0.4	0	F	F	T
20	1	-1	-1	-1	0	0	1	-1	-2	-0.4	0	F	F	T
21	1	1	0	-1	1	1	1	0	2	1	1	T	F	T
22	1	1	1	1	-1	1	0	1	3	1	1	F	T	T
23	1	0	1	1	1	1	1	1	4	1	1	T	T	T
24	1	1	1	1	1	1	0	1	5	1	1	F	T	T
25	-1	1	1	-1	-1	0	1	1	-1	-1	1	F	F	F
26	-1	1	0	0	0	1	1	0	0	0	1	T	F	T
27	1	1	0	0	0	1	1	0	2	1	1	T	F	T
28	-1	1	1	-1	0	1	1	1	0	0	1	T	T	T
29	1	-1	-1	0	0	1	1	-1	-1	-1	1	T	F	T
30	0	-1	-1	-1	-1	1	1	-1	-4	-1	1	T	F	T

Based on Table 4.7 above is the result of sentiment value for original, ID3-1, ID3-2 and SIM-SUM (simple summation). Simple summation is adding the value of all tokens. Then, divide with number of Tweets to get the final value. In the precision check column, T is True and F is false, which means whether the original sentiment and the other sentiment values are same (true) or not (false).

ID3-1	ID3-2	SIM-SUM
42%	50%	66%
52%	48%	52%
42%	35%	35%

36%	32%	31%
33%	29%	28%
33%	27%	27%

Based on the result of precision check, the precision percentage is calculated and it is proven that ID3-2 is the best decision tree to be used. ID3-2 has higher percentage of precision value which is 53%, while ID3-2 has a lower percentage of precision value, which is 50%. This is maybe because ID3-1 uses two attributes, 1 (positive) and 0 (neutral), therefore the result is not too accurate since the testing set consists of 3 attributes which are 1 (positive), 0 (neutral) and -1 (negative). ID3-2 produce better results due to having 3 attributes, that is the same as testing data.



Row No.	prediction(O...	confidence(...	confidence(...	ORIGINAL	TEXT
8	positive	0.591	0.409	positive	saya di rumah seri kenangan ulu kinta
9	positive	0.591	0.409	positive	amek cik binik lagi kerja
10	positive	0.591	0.409	negative	mati sangat keraskah kena buat kajian ini
11	positive	0.591	0.409	positive	majlis daerah hulu langat selangor liverbird tirf apa lagi yang aktif sekarang
12	positive	0.591	0.409	positive	apa itu majlis daerah
13	positive	0.591	0.409	positive	duit hadiah yang diambil dari yuran kemasukkan tak boleh dikeluarkan yuran dikenakan hanya untuk bayar kos ...
14	positive	0.591	0.409	positive	aku ada dengar dua ipoh ini budak-budak tirf agak puncak banyak pengikut dua sini namun nak menjilat balik h...
15	positive	0.591	0.409	positive	liverbird tirf apa lagi yang aktif sekarang
16	positive	0.591	0.409	positive	aku rasa aku nak perkhidmatan dibawah tawaran hangat itu masih menjadi tersebut tanya darul bangi akan kek...
17	positive	0.591	0.409	positive	tapi tapi
18	positive	0.591	0.409	positive	aktif ciri itu kena iaitu persetujuan cik binik dulu kalau boleh penyenggara aku akan ciri untuk gegarkan dunia lfc ...
19	positive	0.591	0.409	positive	dan satu perkara yang aku masih pertimbangkan qadar untuk ciri aktif dalam lfc penyokong kelab hempedu dud...
20	positive	0.591	0.409	negative	aku tidak akan ada lagi melepak lewat malam macam sekarang
21	positive	0.591	0.409	negative	jangan takut jatuh hati mesti pernah sakit hati jika kamu upin kamu akan tahu apa yang jam dilakukan selanjutn...
22	positive	0.591	0.409	negative	ini lawak jangan marah-marah

Figure 4.8: RapidMiner Studio output

Figure 4.8 shows the result of using Decision Tree classifier in RapidMiner Studio. It produces inaccurate results, where all the predicted sentiment is classified to positive. It is maybe because the calculation algorithm is incorrect such as operators used in the tool.

4.4 Conclusion

The project was a success by following the methodology stated in Chapter 3. Based on the results, it can be concluded that ID3-2 is more suitable for data testing compared to ID3-1.

CHAPTER 5

ANALYSIS

5.1 Introduction

Analysis is done to make sure the information is meaningful and can be used as reference in developing strategies in assuring the success or failure of an experiment. This chapter discuss on assessing the accuracy of the result of this project. This is an important process to prove that this project is functional and practical. Data need to be analyze so that to ensure whether the outcomes are high in quality and consistent or not. According to the previous chapter, there are 3 methods to identify the sentiment values, which are decision tree (ID3-1, ID3-2) and Simple Summation which is calculated manually. It is proven that all three methods are not consistent in terms of precision thus making it difficult to identify the best method to calculate sentiment analysis.

5.2 Accuracy Assessment / Testing

In this chapter, the same data will be inserted into a tool which is JCreator LE, to calculate the precision of sentiment analysis of the data by using Multilayer Perceptron (MLP) method. JCreator LE uses Java programming language. Multilayer Perceptron source code is obtained from the internet which is freely used and is modified to enable the coding to calculate sentiment analysis using Artificial Neural Network. The author for the coding is Phil Brierley. Details on data analysis is explained below.

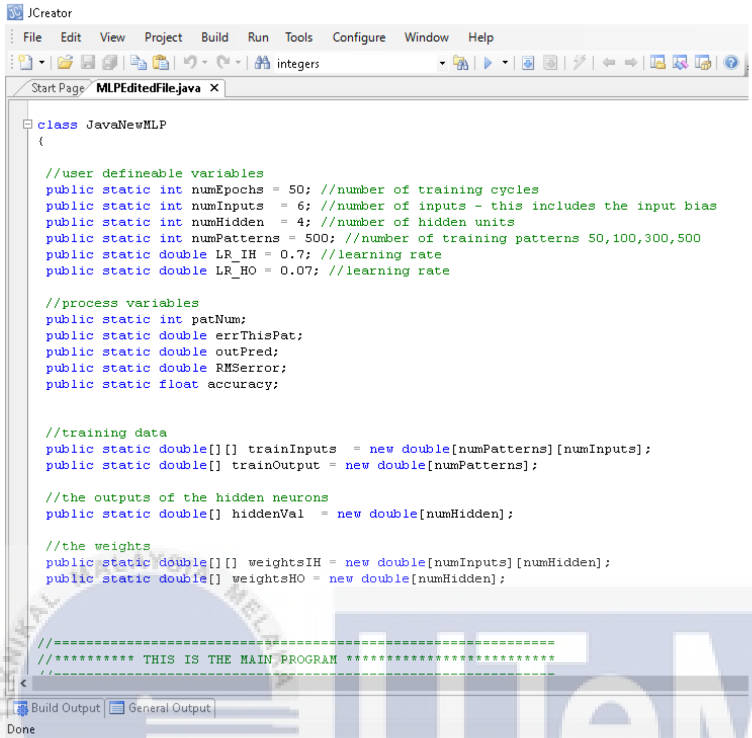


Figure 5.2.1 MLP source code in JCreator LE

Some of the variables that are used in this code are *numEpochs*, *numInputs*, *numHidden*, *numPatterns*, *LR_IH*, and *LR_HO*. The first term, *numEpochs*, is defined as how many times data are trained. In neural network, one epoch means one forward pass and one backward pass for the whole set of data. It cannot be sure whether 10 epochs or 100 epochs is sufficient for the data to be well-trained. In this code, the author set the number of training cycles as 50 at first. Then, the number of training patterns are raised to observe the performance. For *numInputs*, it means how many inputs that is inserted. For this project, there are 5 tokens, or also known as inputs. In neural network, bias (which is always 1) in inserted in the number of inputs to create a hyperbolic tangent (tanh) curve with the range from -1 to 1, so that it can handle the sentiments value of the data which are -1, 0 and 1. Therefore, number of inputs in this project is 6, which includes bias. Next is the *numHidden*, explained as number of hidden units that have been defined by the original author. Apart from that is the number of training patterns, which is *numPatterns*. The significant of the term is the number of patterns that can be trained, which can be any number until 1000 because there are only 1000 data. Plus, the meaning for the term *LR_IH* is the learning rate from the input layer to the hidden layer while LR_HO is the learning rate from the hidden layer to the output layer. The learning rate set by the author for *LR_IH* is 0.7 and the learning rate for *LR_HO* is 0.07. Learning rate is also known as the weight in the neural network diagram.

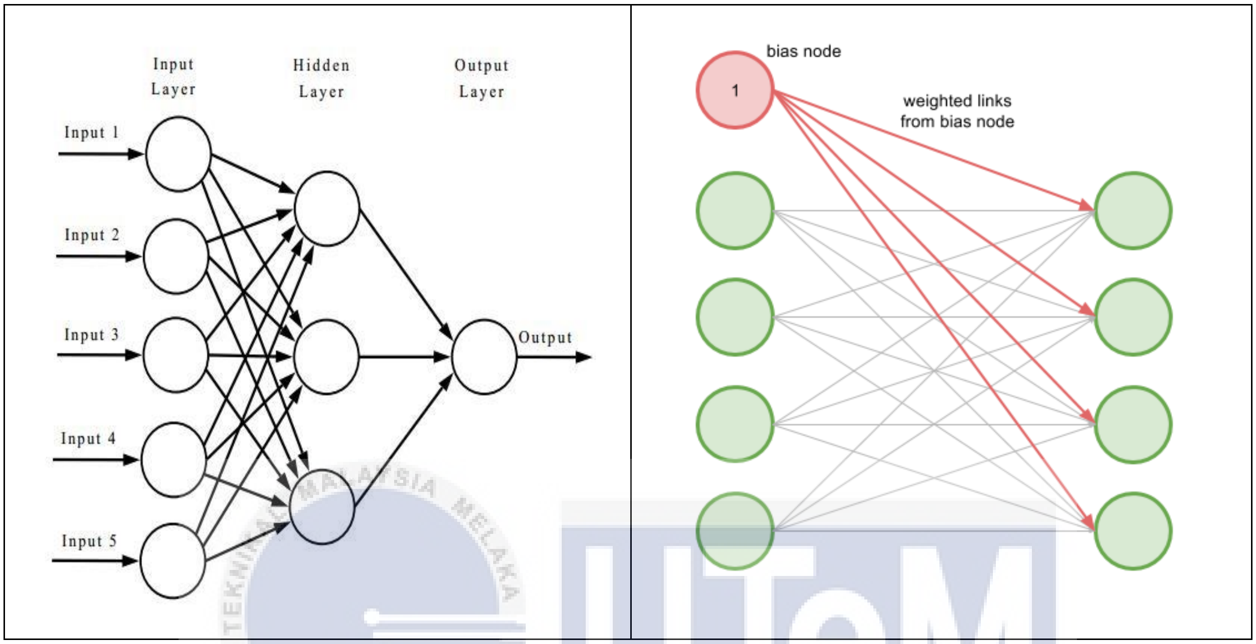


Figure 5.2.2: Artificial Neural Network Diagram with Bias node sample

The performance of this method can be observed by looking at the different number of training patterns. For this project, the number of training patterns chosen are 50, 100, 200, 300, 400 and 500. Below are the Artificial Neural Network Model Prediction table for 50 patterns, their actual sentiment value and the Artificial Neural Network Prediction value and the status whether it is Correct or Wrong.

When ***numPatterns*** **= 50:**

Table 5.2.1 ANN Model Prediction for first 50 Patterns

PATTERN	ACTUAL	ANN MODEL PREDICTION	FINAL	STATUS
1	-1	-0.877928593	-1	Correct
2	0	-0.46871618	0	Correct
3	-1	-0.962273327	-1	Correct
4	1	1.026502288	1	Correct
5	-1	-0.853502349	-1	Correct
6	1	1.026488646	1	Correct

7	0	0.825133267	1	WRONG
8	1	0.875166093	1	Correct
9	1	0.940172608	1	Correct
10	-1	-0.849268001	-1	Correct
11	1	1.026805407	1	Correct
12	1	1.025979735	1	Correct
13	1	0.809613603	1	Correct
14	1	1.026769593	1	Correct
15	1	0.608767767	1	Correct
16	1	1.026835039	1	Correct
17	1	0.333243477	0	WRONG
18	1	1.026839489	1	Correct
19	1	1.026242709	1	Correct
20	0	0.355435571	0	Correct
21	-1	-0.421700227	0	WRONG
22	-1	-0.134637251	0	WRONG
23	1	1.026769593	1	Correct
24	1	1.026488646	1	Correct
25	1	1.026688511	1	Correct
26	1	1.026839489	1	Correct
27	0	0.176021747	0	Correct
28	1	0.986868751	1	Correct
29	1	1.026502288	1	Correct
30	1	1.001415662	1	Correct
31	1	0.338600892	0	WRONG
32	-1	-0.929553852	-1	Correct
33	1	1.026839489	1	Correct
34	1	0.868098436	1	Correct
35	1	0.901873212	1	Correct
36	1	1.026834078	1	Correct
37	1	1.010272152	1	Correct
38	1	1.022762247	1	Correct
39	1	0.995194735	1	Correct
40	1	1.006886955	1	Correct
41	1	1.008863713	1	Correct
42	1	0.987463394	1	Correct
43	1	0.914498317	1	Correct
44	-1	-0.48910277	0	WRONG

45	1	0.836121501	1	Correct
46	-1	-0.424680518	0	WRONG
47	1	0.995194735	1	Correct
48	-1	-0.123619005	0	WRONG
49	1	1.022944187	1	Correct
50	0	1.026823769	1	WRONG

According to the table above, there are 9 incorrect predictions of sentiment value for 50 patterns. This proves that the method is showing a good performance although having a few mistakes in precision.

Below are the Artificial Neural Network Model Prediction table for 100 patterns, their actual sentiment value and the Artificial Neural Network Prediction value and the status whether it is Correct or Wrong.

When *numPatterns* **= 100:**

Table 5.2.2 ANN Model Prediction for first 100 Patterns

PATTERN	ACTUAL	ANN MODEL PREDICTION	FINAL	STATUS
1	-1	-0.647112108	-1	Correct
2	0	-0.437728575	0	Correct
3	-1	-1.10199158	-1	Correct
4	1	0.980036544	1	Correct
5	-1	-1.096011071	-1	Correct
6	1	0.979950878	1	Correct
7	0	0.889833041	1	Wrong
8	1	0.97988345	1	Correct
9	1	0.980450126	1	Correct
10	-1	-0.640969349	-1	Correct
11	1	0.979903475	1	Correct
12	1	0.97991431	1	Correct
13	1	0.791245013	1	Correct
14	1	0.979914687	1	Correct
15	1	0.904597236	1	Correct
16	1	0.979903681	1	Correct
17	1	0.406006526	0	Wrong
18	1	0.979903313	1	Correct

19	1	0.979902883	1	Correct
20	0	0.661714718	1	Wrong
21	-1	-0.682780205	-1	Correct
22	-1	-1.074875115	-1	Correct
23	1	0.979914687	1	Correct
24	1	0.979950878	1	Correct
25	1	0.979903303	1	Correct
26	1	0.979903313	1	Correct
27	0	0.983469859	1	Wrong
28	1	0.976664288	1	Correct
29	1	0.980036544	1	Correct
30	1	0.979918377	1	Correct
31	1	0.637090384	1	Correct
32	-1	-1.146961964	-1	Correct
33	1	0.979903313	1	Correct
34	1	0.854541647	1	Correct
35	1	0.979500416	1	Correct
36	1	0.979903342	1	Correct
37	1	0.983941466	1	Correct
38	1	0.979946598	1	Correct
39	1	0.979907008	1	Correct
40	1	0.966745736	1	Correct
41	1	0.979333553	1	Correct
42	1	0.979717145	1	Correct
43	1	0.978988242	1	Correct
44	-1	-0.724602228	-1	Correct
45	1	0.971124312	1	Correct
46	-1	-0.240048058	0	Wrong
47	1	0.979907008	1	Correct
48	-1	-0.930924969	-1	Correct
49	1	0.979903193	1	Correct
50	0	0.979903602	1	Wrong
51	-1	-0.759480771	-1	Correct
52	0	-0.317123536	0	Correct
53	0	0.979903313	1	Wrong
54	0	0.979937426	1	Wrong
55	1	0.406006526	0	Wrong
56	1	0.979673878	1	Correct
57	0	0.665352562	1	Wrong
58	1	0.955374105	1	Correct
59	1	0.980036544	1	Correct

60	1	0.979903207	1	Correct
61	1	0.979903681	1	Correct
62	1	0.979907575	1	Correct
63	1	0.406006526	0	Wrong
64	1	0.980263482	1	Correct
65	0	0.689803612	1	Wrong
66	0	0.979717145	1	Wrong
67	1	0.978856775	1	Correct
68	1	0.979906116	1	Correct
69	1	0.979630655	1	Correct
70	1	-0.194234196	0	Wrong
71	1	0.979896373	1	Correct
72	1	0.947694177	1	Correct
73	-1	-1.114371383	-1	Correct
74	0	0.979892841	1	Wrong
75	1	0.653437441	1	Correct
76	1	0.979903342	1	Correct
77	1	0.979903321	1	Correct
78	1	-0.240048058	0	Wrong
79	1	0.979903313	1	Correct
80	1	1.023167432	1	Correct
81	0	0.97990222	1	Wrong
82	1	0.979903303	1	Correct
83	1	1.023440299	1	Correct
84	1	0.97990222	1	Correct
85	1	0.623256503	1	Correct
86	1	0.980036544	1	Correct
87	1	0.980036544	1	Correct
88	1	0.979903207	1	Correct
89	1	0.974085559	1	Correct
90	1	0.979903313	1	Correct
91	1	0.979950878	1	Correct
92	1	0.979903602	1	Correct
93	1	0.979900288	1	Correct
94	1	0.931460419	1	Correct
95	1	0.617797026	1	Correct
96	0	-0.365901885	0	Correct
97	0	0.995400908	1	Wrong
98	1	0.668326211	1	Correct
99	1	0.979907008	1	Correct
100	1	0.109519176	0	Wrong

Based on the table above, there are 19 incorrect predictions of sentiment value for 100 patterns. This proves that the method is good in terms of performance although not precisely accurate.

Table 5.2.3 RMS Error Based on Default Epoch = 50

epoch	RMS ERROR
0	0.232256681
1	0.169777041
2	0.136815503
3	0.142983149
4	0.121719607
5	0.127634747
6	0.13245719
7	0.119434683
8	0.116522973
9	0.12212544
10	0.11380472
11	0.118205043
12	0.113363844
13	0.11512509
14	0.113989982
15	0.105713225
16	0.105983094
17	0.107277329
18	0.105867267
19	0.10567526
20	0.105668706
21	0.107487643
22	0.115689916
23	0.123468761
24	0.12601211
25	0.124700187
26	0.132661519
27	0.133530619
28	0.113846899
29	0.110621334
30	0.122383744
31	0.106180001
32	0.108704947
33	0.104217983

34	0.103365898
35	0.130305335
36	0.104847937
37	0.122885893
38	0.123343549
39	0.116702608
40	0.121160942
41	0.108620215
42	0.106167749
43	0.108385342
44	0.101760424
45	0.100882879
46	0.101595852
47	0.12950327
48	0.129572958
49	0.100759955
50	0.10108852

Based on the Table 5.2.3, the Root Mean Square Error or RMS Error for the first 50 epoch are 0.10108852. RMS Error is to measure the difference between fitted line to data points. In this project scope, RMS Error is the difference between the Artificial Neural Network model prediction on sentiment value and the actual sentiment value over the number of training patterns. The best performance is where the RMS Error is the most minimal.

Overall, table below is the RMS Error based on different number of epochs.

Table 5.2.4 RMS Error based on Epochs

Number of Patterns	**Number of Epochs**	**Precision Percentage**	**RMS Error**
500	**50**	74%	0.45654375746099235
	100	76%	0.45929934502840647
	200	76%	0.4773323848486768
	300	72%	0.4450825489116451
	400	74%	0.4517595588134799
	500	77%	0.47461938184957225

Based on the table above, it is proven that the RMS Error and the precision percentage are inconsistent. It has its ups and down that should have supposed to be decreasing but it is not the same as expected.

In conclusion, below is the comparison table between the methods, which are decision tree, simple summation and artificial neural network. Selected epochs are set as 50 and the number of patterns are 50, 100, 200, 300, 400 and 500.

Table 5.2.5 Precision Percentage Comparison Table

Number of Epochs	**Number of Patterns**	**ID3-1**	**ID3-2**	**Simple Summation**	**Artificial Neural Network**
50	**50**	42%	52%	66%	86.0%
	100	52%	48%	52%	88.0%
	200	42%	35%	35%	72.0%
	300	36%	32%	31%	73.0%
	400	33%	29%	28%	73.5%
	500	33%	27%	27%	73.4%

Based on Table 5.2.5, it can be concluded that the best method for sentiment analysis by using Artificial Neural Network where when the first 500 patterns is calculated its sentiment analysis, the ID3-1 precision percentage is 33%, ID3-2 is 27%, simple summation is 27% and Artificial Neural Network is 73.4%. It is clearly proven that Artificial Neural Network is the overall best method to predict sentiment analysis.

5.3 Conclusion

In this chapter, full description and details on the analysis for this project have been discussed. It aims to describe the outcome on using a tool which in this project is JCreator LE by using the Multi-Layer Perceptron coding. Project conclusion will be discussed in the next chapter.

CHAPTER 6

CONCLUSION

6.1 Introduction

This is the last chapter that describes the result and the outcome by focusing on the strength and weakness plus suggestion on how to improve it in the future to ensure the development of the system is smooth and efficient.

6.2 Project Weakness

Since the system is a new idea, it is still not yet perfect as it requires more time to be improved and developed. Truthfully, the system is not yet precise since only the first 5 token from a long sentence is taken to calculate the sentiment analysis. Hopefully, in the future, more tokens can be collected because with huge amount of information, it is easier to gain good performance. Apart from that, there is not defined scope for the tweets that is retrieved. Twitter sentences are taken randomly based on Malay language and originated from Malaysia only. There is no defined scope such as movie review, product trend or service satisfaction, to name a few. Therefore, the training and testing of the data are most probably affected since their type of sentences are not the same. Apart from that, by using the multilayer perceptron coding, it only calculates the precision of training data instead of the testing data. Thus, the precision percentage is higher than expected result.

6.3 Project Strength

Although having several weaknesses stated as above, there are also strengths which are the contribution of this project. Firstly, the project can identify the most suitable method that can be done to calculate sentiment analysis between decision tree, simple summation and artificial neural network. Artificial neural network is the highest according to the precision percentage. Hopefully in the future, a coding that can calculate the precision percentage of testing data can be created.

This project can be developed more to be a system with high precision if the weaknesses can be overcome. Improvement and adding new features such as a website system that can predict Malay sentences sentiment analysis just by typing in can increase the value of this project and is not impossible to proceed with.

6.4 Suggestion for Project Improvement

Improvement for a project is very much needed so that the project can be better in terms of precision. It will ensure that the system is more reliable and can be used and high in user satisfaction.

For this project, there are a few improvements that can be done according to the project weaknesses that has been stated earlier. Firstly, increase the number of tokens. It is known that several tokens will build up a sentence. For this project, only the first 5 tokens of training data are taken for the sentiment analysis to be calculated. Thus, it will not necessarily accurate in terms of to train the testing data due to insufficient amount of training data.

Apart from that, to define a scope for tweets data that are retrieved from Twitter. As for now, there is still no coding for the scope to be fetch together with Malay language and Malaysia as the location in one single source code. Sentences should be in a scope so that the training and the testing will be easier and more accurate. Sentences discussing about different topics are more difficult to handle because the testing data is not compatible with the training data.

6.5 Lessons learnt

As a student, there are many things that I have learnt along in completing this final year project. It is vital to save a backup whenever I have done any tasks to avoid from losing it permanently. Besides, it is important to refer to supervisor whenever I am facing any problem. For example, I was facing a dead end when the RapidMiner cannot calculate the sentiment values correctly. Instead of using RapidMiner, my supervisor encourages me to use another alternative which is JCreator LE. Fortunately, the tool work out fine although with a few modifications to make it able to calculate the precision of Malay tweets sentiment analysis. Apart from that, I also learnt to handle stress and manage time wisely. It is very stressing to face a problem like I stated earlier and to face it a few weeks before presentation week. Luckily, my supervisor advises me to proceed with the result although it is incorrect because having incorrect result does not mean failure, it just means that there are maybe steps that are skipped or mistaken somewhere. The mistake is difficult to find since RapidMiner does not show what happen to data when we insert each operator.

6.6 Conclusion

In this chapter, project strengths and weaknesses are discussed for the system to be more perfect in the future. Nevertheless, the objectives have been achieved.

REFERENCES

B. Hawelka, I. Sitko, E. Beinat, S.Sobolevsky, P. Kazakopoulos, C.Ratti. (2014) Geo-located Twitter as proxy for global mobility patterns *Cartography and Geographic Information Science, 4*(3), 260-271. Available at: https://www.researchgate.net/publication/258247194_Geo-Located_Twitter_as_Proxy_for_Global_Mobility_Patterns

L. C. Leong, S. Basri, R. Alfred (2012) Enhancing Malay Stemming Algorithm with Background Knowledge *PRICAI 2012: Trends in Artificial Intelligence, Vol 7458,* 753-758. Available at: https://link.springer.com/chapter/10.1007/978-3-642-32695-0_68

Adriani, M., Trisedya, B.D., Vania, C., & Wicaksono, A.F. (2014). Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets. *PACLIC*. Available at: https://www.semanticscholar.org/paper/Automatically-Building-a-Corpus-for-Sentiment-Anal-Wicaksono-Vania/cf10fc1504bd3d261192def17a66dc5b1dd9b821

Mohammad Darwich, and Shahrul Azman Mohd Noah, and Nazlia Omar, (2016) Automatically generating a sentiment lexicon for the Malay language. *Asia-Pacific Journal of Information Technology and Multimedia*, 5 (1). pp. 49-59. ISSN 2289-2192. Available at: journalarticle.ukm.my/10056/1/11736-37831-1-PB.pdf

Wattana Punlumjeak, Nachirat Rachburee, "A comparative study of feature selection techniques for classify student performance", *Information Technology and Electrical Engineering (ICITEE) 2015 7th International Conference on*, pp. 425-429, 2015. Available at: ieeexplore.ieee.org/document/7408984/

A. Alsaffar, N. Omar, Integrating a Lexicon Based Approach and K Nearest Neighbour for Malay Sentiment Analysis, *Journal of Computer Science*, *11*(4), 639-644. **DOI:** 10.3844/jcssp.2015.639.644. Available at: thescipub.com/abstract/10.3844/jcssp.2015.639.644

Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, Volume 5, Issue 4, 2014, Pages 1093-1113, ISSN 2090-4479, http://dx.doi.org/10.1016/j.asej.2014.04.011.
(http://www.sciencedirect.com/science/article/pii/S2090447914000550)

Hansen Wijaya, Alva Erwin, Amin Soetomo, Maulahikmah Galinium (Department of Information Technology Faculty of Engineering and Information Technology Swiss German University). Twitter Sentiment Analysis and Insight for Indonesian Mobile Operators. *2 Dec 2013, Volume 2013.* Available at: is.its.ac.id/pubs/oajis/index.php/file/download_file/1221

Pang, Bo & Lee, Lillian. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Computing Research Repository - CORR. 271-278. 271-278. 10.3115/1218955.1218990. Available at: www.cs.cornell.edu/home/llee/papers/cutsent.pdf

Prabowo, Rudy & Thelwall, Mike. (2013). Sentiment analysis: A combined approach. Journal of Informetrics. . 143-157. 10.1016/j.joi.2009.01.003. Available at: www.cyberemotions.eu/rudy-sentiment-preprint.pdf

Shamsudin, Nurul & Basiron, Halizah & Sa'aya, Z. (2016). Lexical based sentiment analysis - Verb, adverb & negation. 8. 161-166. Available at: https://www.researchgate.net/publication/308052502_Lexical_based_sentiment_analysis_-_Verb_adverb_negation

Puteh, Mazidah & Isa, Norulhidayah & Puteh, Sayani & Amalina Redzuan, Nur. (2013). Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System. Lecture Notes in Engineering and Computer Science. 3. Available at: https://www.researchgate.net/publication/260427410_Sentiment_Mining_of_Malay_Newspaper_SAMNews_Using_Artificial_Immune_System

Singh, Vikram & Midha, Neha. (2015). A Survey on Classification Techniques in Data Mining. International Journal of Computer Science & Management Studies 2231–5268. 16. Available at: https://www.researchgate.net/publication/280737179_A_Survey_on_Classification_Techniques_in_Data_Mining

Kouloumpis, Efthymios & Wilson, Theresa & Moore, Johanna. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG!. ICWSM. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2857