ISSN 2278-3091

Volume 9 No.2, March -April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse76922020.pdf

https://doi.org/10.30534/ijatcse/2020/76922020

Employ Twitter Data to Perform Sentiment Analysis in the Malay Language

Abdul Karim Mohamad[1], Mailasan Jayakrishnan[1], Nurnajwa Hazwani Nawi[2]

[1]Centre for Advanced Computing Technology, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, karim@utem.edu.my

[2]Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

ABSTRACT

The main intention of this research is to discover the significance of sentiment analysis of Twitter data whether it is positive, neutral or negative. The sentiment analysis is dependent mining about textual content which being extracted and identified as contextual and subjective knowledge in such perceptible origin of recent rapid expanding computer science research. We started with a systematic literature review, where we had adopted both qualitative coding and text mining by scrutinizing 3282 of input of textual data retrieved from Twitter Streaming API. We perceived the problem as the decision trees kind of sentiment analysis in learning and information gaining. Therefore, we showed how basic decision trees are built to calculate the sentiment values of Twitter data. Sentiment analysis has transformed from interpreting online textual output analysis into perceiving contextual social media texts for example from Twitter. Hence, two decision trees were built to observe the performance and information gaining of decision trees. Thus, the precision of both decision trees led to the precision percentage that will be respectively stated, and the best decision tree can be obtained.

Key words: Decision Tree, Malay Language, Precision, Sentiment Analysis, Text Mining, Twitter.

1. INTRODUCTION

The Internet is an influential automation about knowledge sharing and information maturity [1]. With the outbreak of the internet on the wireless connection in the new 21st century, the community is now directly united. The Internet today has no limitation to the users**.** Institutions, people and companies felt deeply about this technological diversity [2]. Everything is only with the touch of a finger [3]. Therefore, Twitter can also be used in more technically intriguing ways [4]. The inspired community has perceived to pursue and gain information from your tweets within their Twitter feeds [5]. Yet, a spreading figure of Twitter users has circulated practical content and that is the actual value about Twitter [6]. It derives the community to become an amateur writer who is writing sentences about life, sharing and describing something that they have had from engaging activities about their day [7].

Therefore, there are sentences are taken from Twitter live tweets using text mining of Twitter Streaming Application Programming Interface (API) for this research purpose. Text mining is the approach of the natural processing approach and a systematic method to collect meaningful tweets [8]. An API is a tool to enable computer programs and web services to communicate. The current scenario that we are facing now in Twitter data is that no labeled Twitter corpus available in the Malay language. We need to find a method to filter tweets in the Malay language originated from Malaysia and a classifier to categorize tweets to positive, negative or neutral. 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 labeled as the positive, negative or neutral hinge 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” is removed from the tweets and considered as a positive word. Furthermore, a decision tree classifier is used where information content, information gain, and decision tree are first calculated manually and then the RapidMiner tool is used to compare the results of using the Decision Tree classifier. The program uses a data dictionary of Malay words labeled as positive, negative or neutral sentiment value. A Malay language dataset is manually labeled with their sentiment values using human interpretation to be used as a training set.

2. LITERATURE REVIEW

Twitter has become one of the most utilized social media applications because it is both rapid and personal [9]. Organizations and businesses are also taking the opportunity of the social network to connect with their customers and other people online [10]. Twitter is a combination of texting, instant messaging and blogging but it has come up with concise content and a deep congregation [11]. The sentiment analysis approach is progressively leading toward classifying the sentiment text to one or extra predefined sentiment division for the automated maintenance and creation of analysis-aggregation websites.

Data will be divided into testing data and training data sets. The training data set is utilized for classifier learning action and the testing data set is used to test the classifier's behavior after learning action is done. The training of classifiers is based on both unigrams and bigrams features [12]. Results gained shows that training data size can be affecting the classifiers' performance [13].

In terms of the knowledge representation method, a comprehensive database is needed which contains labeled sentiment values to identify sentiments. By using these techniques, a hybrid model is adopted for performing sentiment analysis of almost any text given. The knowledge representation method is seemed difficult due to the need for a large lexical database. An ID3 is a commonly used decision tree [14]. Figure 1, 2 and 3 is the example of how ID3 is conducted.

Formula for Entropy:

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

Figure 1: Entropy Formula

Formula for Information Content:

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

Figure 2: Information Content Formula

Formula for Information Gain:

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

Figure 3: Information Gain Formula

Table 1 is the sample token table to build a decision tree.

Table 1: The Token Table

Sample	Token 1	Token 2	Token 3	Token 4	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	

Table 1 summarized a set of sentences are broken down into four 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 excessive information gain, so it is chosen as the origin 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. We have figured the decision tree that will be used and referred to as to predict sentiment value as shown in Figure 4.

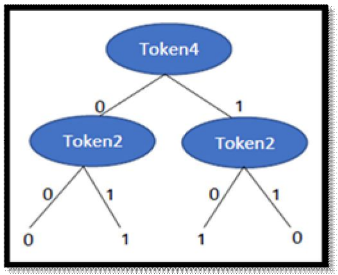


Figure 4: The Decision Tree

Figure 4 shows the decision tree that will be utilized and referred to as to predict the sentiment value about the last three samples of unknown sentiment value and sentiment results as shown in Table 2.

Table 2: The Sentiment Result

Samples	Token 4	Token 2	Sentiment
X	1	1	0
Y	1	1	0
Z	0	1	1

Table 2 summarizes the sentiment result based on the decision tree constructed in Figure 4 when the value for the Token 4 is 1 and the value for Token 2 is 1, the sentiment value gained is 0. The result is the same for the second prediction. For the third prediction, when Token 4 has a value of 0 and Token 2 has a value of 1, the sentiment value gained is 1. A Decision Tree is a tree where nodes are labeled based on the attributes, the edges split a node are labeled by tests on the attribute's weight, and the leaves are labeled by classes [15].

It segregates a script by initializing at the tree root and gripping strongly descending through the branches whose surroundings are fulfilled by the script until a leaf node arrives. The script is then categorized within the class that tags the leaf node. Decision trees have already been utilized within various applications such as language and speech processing [11]. For trees to have properties such as minimality, ID3 ranks the features of training data according to the information gain.

After performing a deep systematic literature review, we aim to find tweets with labeled sentiment to be used as training data. Apart from that, to find tweets in the Malay language which is from Malaysia and lastly is to find classifiers to calculate the sentiment values of the tweets. Developing this research must achieve the following objectives so that the research is a success.

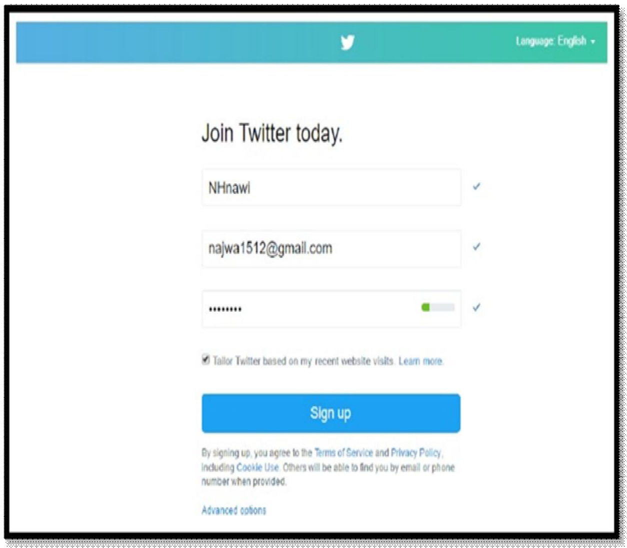
3. METHODOLOGY

A key part of the research is the methodology [16]. The methodology characterizes the deep philosophical base of the selected research methods, counting whether the study is utilizing quantitative or qualitative methods or a fusion about both methods [17]. The methodology can comprise of research on surveys, interviews and other research approaches and could combine both historical and present information [18]. A strong expanding and coherent methodology area will contribute a strong determination for the success of the full study and will edge toward a solid conclusion area [19].

The waterfall model is a linear, analytical progression of tread be taken throughout the Software Development Life Cycle (SDLC) that is famous within Software Engineering (SE) and yet it is a product evolution. Moreover, in a waterfall model, each stage must be concluded before the next stage can initiate and there are no extending features within the stages [20].

The conclusion about one stage action as the input for the next stage is most required. We have designed and developed the waterfall model for this research that consists of five (5) phases: (1) Data Collection (Retrieval of Tweets data), (2) Data Analysis (Preprocessing of data), (3) Experimental Design (Parallel Processing), (4) Evaluation and Testing (Sentiment Value Scoring) and (5) Conclusion (Sentiment Result).

The first phase is about collecting testing data which is collecting Twitter data using Twitter Streaming API by (1) Sign up a Twitter account as shown in Figure 5 for gaining 4 keys that needed to run Twitter Streaming API which is (1) API key and (2) API secret as shown in Figure 6, (3) access token and (4) access token secret as shown in Figure 7 and (2) Insert Keys obtained into the algorithm to download live streaming Tweets as shown in Figure 8.



NHnawl ✓

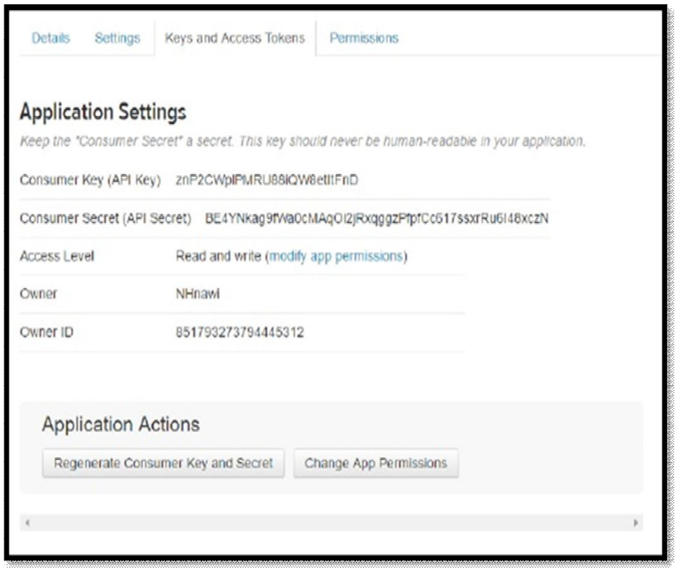
najwa1512@gmail.com ✓

•••••••• ✓

☑ Tailor Twitter based on my recent website visits. Learn more.

By signing up, you agree to the Terms of Service and Privacy Policy, including Cookie Use. Others will be able to find you by email or phone number when provided.

Figure 5: The Twitter Sign-Up



Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) znP2CWplPMRU88lQW8etltFnD

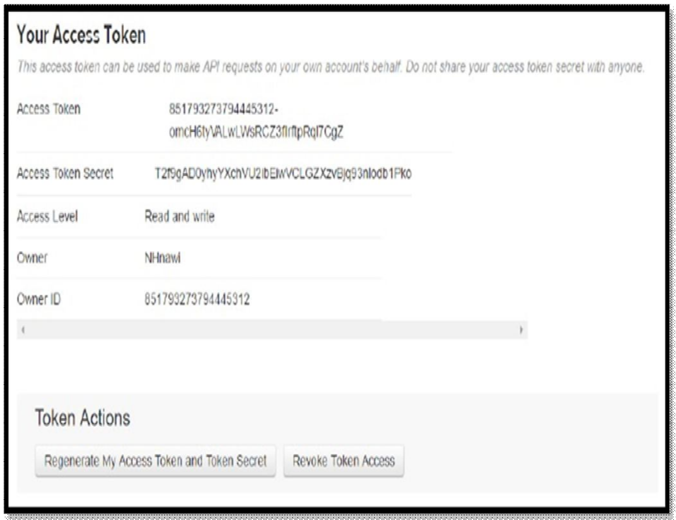
Consumer Secret (API Secret) BE4YNkag9fWa0cMAqOl2jRxqggzPfpfCc617ssxrRu6l48xczN

Access Level Read and write (modify app permissions)

Owner NHnawl

Owner ID 851793273794445312

Figure 6: The Consumer Key (API Key) and Consumer Secret (API Secret)



This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token 851793273794445312-omcH6tyVALwLWsRCZ3flrftpRql7CgZ

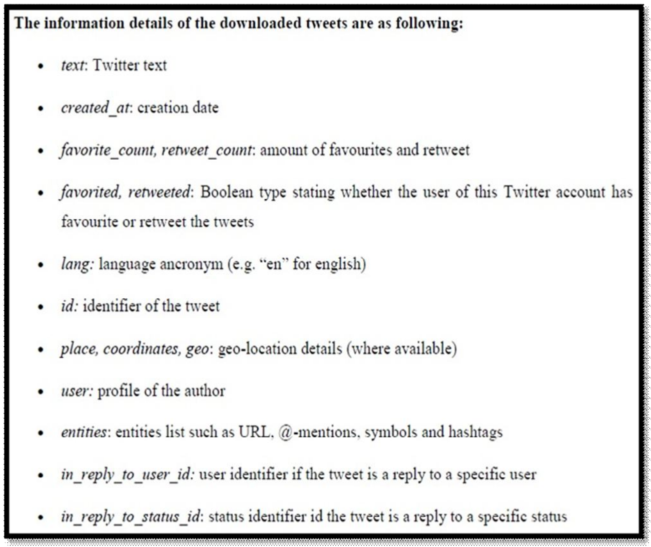
Access Token Secret T2f9gAD0yhyYXchVU2lbEIwVCLGZXzvBjq93nlodb1Pko

Access Level Read and write

Owner NHnawl

Owner ID 851793273794445312

Figure 7: The Access Token and Access Token Secret



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

Figure 8: Insert Keys Downloading Live Streaming Tweets

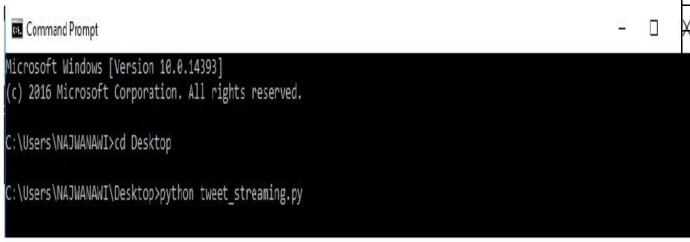
Figure 8 shows the insert keys downloading live streaming Tweets. Next, it is used to find the training data set. The training data set is the data with manually labeled sentiment values using human interpretation while testing data set learns from the training data to predict the sentiment values of a new test set of tweets. Since there is no available Malay language dataset with the labeled sentiment, a database of sentences is manually labeled with their sentiments using human interpretation.

Both data will undergo the preprocessing phase to make the data more meaningful. Several steps required in the preprocessing phase are filtering, tokenizing, stemming and removing stop words. After preprocessing, Tweets will become more meaningful and now ready for the experiment. We have coded programs to download the tweets, that will be saved as tweet_streaming.py as shown in Table 3, followed by entering python tweet_streaming.py to run the coding in the terminal as shown in Figure 9. After that, it will produce data such as shown in Figure 10.

Table 3: The Live Streaming Tweets Algorithm

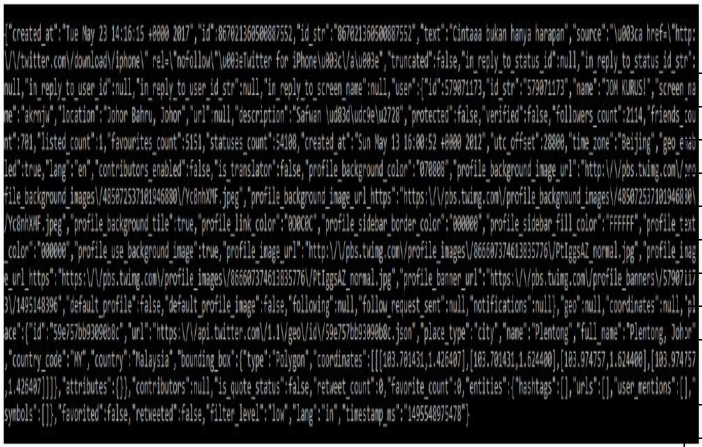
```
#Import the necessary methods from the tweepy library from
tweepy. streaming import StreamListener from tweepy import
OAuthHandler from tweepy import Stream import json import
re import pandas as pd import matplotlib.pyplot as plt

# Variables that contains the user credentials to access Twitter
API
access token =
“851793273794445312-omcH6tyVALwLWsRCZ3fIrftpRql7
CgZ”
access_token_secret =
“T2f9gAD0yhyYXchVU2ibEiwVCLGZXzvBjq93nlodb1Pko”
consumer_key = “znP2CWplPMRU88iQW8etItFnD”
consumer_secret =
“BE4YNkag9fWa0cMAqOi2jRxqggzPfpfCc617ssxrRu6I48xc
zN”

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

def on_date(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 1 = StdOutListener()
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
stream = Stream(auth, 1)

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



```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\NAJWANAWI>cd Desktop

C:\Users\NAJWANAWI\Desktop>python tweet_streaming.py
```

Figure 9: The 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\/5790711 7
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 10: The Live Stream Data Output

The data will undergo parallel processing, which applies a classifier used as a Decision Tree. From the decision tree, the sentiment value scoring is acquired and can be referred to for the testing data set. Based on the scores, the sentiment result of the testing set can be predicted.

4. IMPLEMENTATION ANALYSIS

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 output is in JSON format and is saved as a 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 attributes to perform sentiment analysis. For example, the text attribute in this output is *"Cinta Bukan Hanya Harapan"*. We have tabulated the training data as shown in Table 4.

Table 4: The Training Data Sample

Token 1	Token 2	Token 3	Token 4	Token 5	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	Dearah	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 summarized the training data sample that consists of five (5) tokens and their overall sentiment value. The sentiment with value 1 is a positive sentiment value while sentiment with the value 0 is a negative sentiment value. Furthermore, we have tabulated the testing data as shown in Table 5.

Table 5: The Testing Data Sample

Token 1	Token 2	Token 3	Token 4	Token 5
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 5 summarized the testing data sample that consists of five (5) tokens without the sentiment values. Testing data will learn from training data to predict sentiment value. Therefore, we have tabulated the training data-1 as shown in Table 6.

Table 6: The Training Data-1

Tweets	Token 1	Token 2	Token 3	Token 4	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 6 summarized the training data-1 that consists of four (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. Furthermore, we have designed and developed the decision tree ID3-1 based on Table 6, as shown in Figure 11.

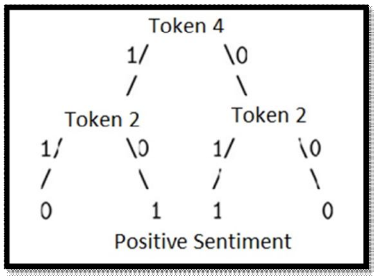


Figure 11: Decision Tree ID3-1

Figure 11 shows the decision tree ID3-1 of the training data-1. Therefore, we will use this decision tree ID3-1 for testing data. As we mentioned earlier two (2) decision trees will be built to observe the performance of decision trees. Therefore, we have tabulated the training data-2 as shown in Table 7.

Table 7: The Training Data-2

Tweets	Token 1	Token 2	Token 3	Token 4	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 7 summarized the training data-2 that consists of four (4) tokens and its overall sentiment value. Sentiment having value 1 is a positive sentiment value, while sentiment with value 0 is a neutral sentiment value and -1 is a negative sentiment value. Furthermore, we have designed and developed the decision tree ID3-2 based on Table 7, as shown in Figure 12.

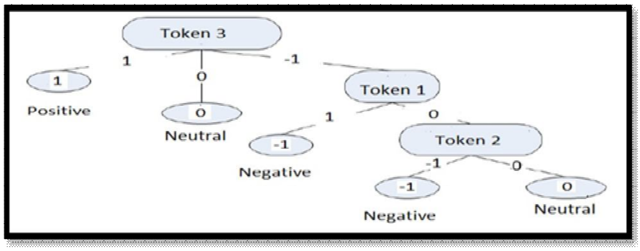


Figure 12: Decision Tree ID3-2

Figure 12 shows the decision tree ID3-2 of the training data-2. Therefore, we will use this decision tree ID3-2 for testing data. 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. Therefore, we have tabulated the sentiment calculation as shown in Table 8.

Table 8: The Sentiment Calculation

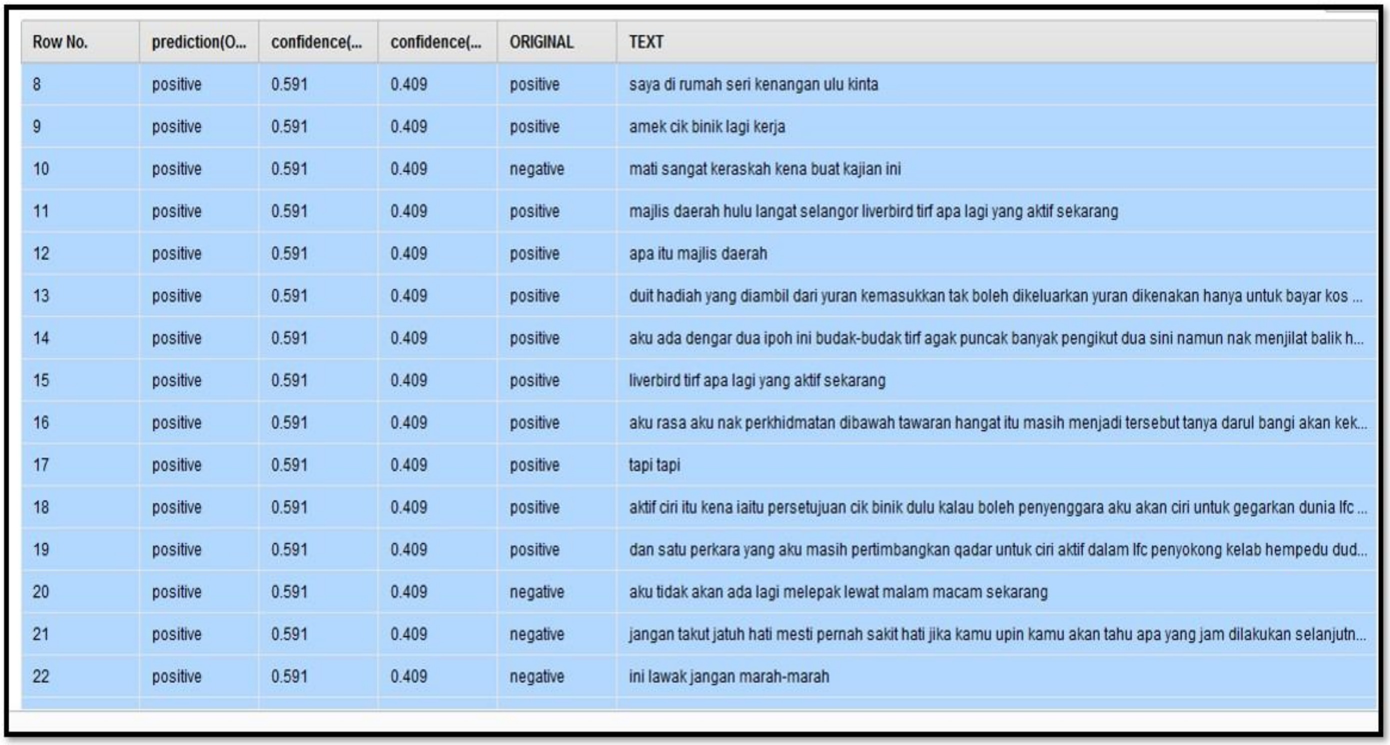
												Precision Check		
No of Tweets	T1	T2	T3	T4	T5	Sentiment	ID3-1	ID3-2	SIM SUM	sim sum **15**	Value	ID3-1	ID3-2	SIM SUM
1	0	0	1	1	1	1	1	1	3	0.6	1	T	T	T
2	0	0	0	1	1	0	1	0	2	0.4	0	F	T	T
3	1	0	0	0	0	0	0	0	1	0.2	0	T	T	T
4	1	1	1	1	1	1	0	1	5	1	1	F	T	T
5	0	1	1	1	1	1	0	1	4	0.8	1	F	T	T
6	1	1	1	1	1	1	0	1	5	1	1	F	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

Table 8 summarized the sentiment calculation of the result of sentiment value for original, ID3-1, ID3-2 and SIM-SUM (simple summation). The simple summation is adding the value of all the tokens. Then, divide with several 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 the same (true) or not (false). Therefore, we have summarized and tabulated the precision percentage as shown in Table 9.

Table 9: The Precision Percentage

ID3-1	ID3-2	SIM-SUM
42%	50%	66%
52%	48%	52%
42%	35%	35%
36%	32%	31%
33%	29%	28%
33%	27%	27%

Table 9 summarizes the precision percentage. Based on the result of a 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 a higher percentage of precision value which is 50%, while ID3-2 has a lower percentage of the precision value, which is 27%. 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 produces better results due to having 3 attributes, which is the same as testing data. The result of using the Decision Tree classifier in RapidMiner Studio, as shown in Figure 13.



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 13: The RapidMiner Studio Output

Figure 13 shows the RapidMiner Studio output that produces accurate results, where all the predicted sentiment is classified as positive. It is maybe because the calculation algorithm is correct such as operators used in the tool.

5. CONCLUSION

We can conclude that the live streaming Twitter dataset can be composed utilizing Twitter Streaming API and processed to be used as a testing set. Besides, a manually labeled Malay language dataset can be created to be considered as the training set. Apart from that, tweets can be collected in the Malay language originated from Malaysia. Moreover, a classifier or Decision Tree will be useful to execute in determining sentiment values of Twitter data.

The initial stage was to identify the most suitable method that can be done to calculate sentiment analysis between decision tree and simple summation. We can increase the number of tokens. It is known that several tokens will build up a sentence. We only took the first 3 to 5 tokens of training data for the sentiment analysis to be calculated. Therefore, it will not necessarily accurate in terms of training the testing data due to an insufficient amount of training data. We can conclude that ID3-2 is more suitable for data testing compared to ID3-1. For future work, we believe that some artificial neural network approaches and techniques could be used for this sentiment analysis.

ACKNOWLEDGMENT

The authors would like to thank the reviewers and editor for their recommendation to strengthen the standard of this research paper. Our thanks are also forwarded to the Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, UTeM for the support and encouragement for the work and this publication.

REFERENCES

[1] W. Halim, “Enterprise Information Technology Strategic Plan (EITSP) delivers Indonesian Bank Performance,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 383–402, Feb. 2020. https://doi.org/10.30534/ijatcse/2020/56912020

[2] J. Zhu, Y. Wang, and C. Wang, “A comparative study

of the effects of different factors on firm technological innovation performance in different high-tech industries," *Chinese Manag. Stud.*, vol. 13, no. 1, pp. 2–25, Apr. 2019. https://doi.org/10.1108/CMS-10-2017-0287

[3] M. Shenify, "Understanding User's Behavior by Social Media Data Clustering," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 167–170, Feb. 2020. https://doi.org/10.30534/ijatcse/2020/25912020

[4] J. P. Pinto, "Twitter Sentiment Analysis: A Political View," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 723–729, Feb. 2020. https://doi.org/10.30534/ijatcse/2020/103912020

[5] I. Xie and J. A. Stevenson, "@Digital libraries: harnessing Twitter to build online communities," *Online Inf. Rev.*, vol. 43, no. 7, pp. 1263–1283, Nov. 2019. https://doi.org/10.1108/OIR-02-2018-0058

[6] H. Sun, E. Ch'ng, and S. See, "Influential spreaders in the political Twitter sphere of the 2013 Malaysian general election," *Ind. Manag. Data Syst.*, vol. 119, no. 1, pp. 54–68, Feb. 2019. https://doi.org/10.1108/IMDS-09-2017-0409

[7] N. Khamis, "Corpus-based Data for Determining Specialised Language Features," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 36–41, Feb. 2020. https://doi.org/10.30534/ijatcse/2020/07912020

[8] G. J. Wu, Z. "Jimmy" Xu, S. Tajdini, J. Zhang, and L. Song, "Unlocking value through an extended social media analytics framework," *Qual. Mark. Res. An Int. J.*, vol. 22, no. 2, pp. 161–179, Apr. 2019. https://doi.org/10.1108/QMR-01-2017-0044

[9] G. Casalino, C. Castiello, N. Del Buono, and C. Mencar, "A framework for intelligent Twitter data analysis with non-negative matrix factorization," *Int. J. Web Inf. Syst.*, vol. 14, no. 3, pp. 334–356, Aug. 2018. https://doi.org/10.1108/IJWIS-11-2017-0081

[10] M. Jayakrishnan, A. K. Mohamad, and A. Abdullah, "Journey of an Enterprise Architecture Development Approach in Malaysian Transportation Industry," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4, pp. 765–774, 2019.

[11] C. Udanor and C. C. Anyanwu, "Combating the challenges of social media hate speech in a polarized society," *Data Technol. Appl.*, vol. 53, no. 4, pp. 501–527, Sep. 2019.

[12] S. Tofighy and S. M. Fakhrahmad, "A proposed scheme for sentiment analysis," *Kybernetes*, vol. 47, no. 5, pp. 957–984, May 2018.

[13] M. Van M. Buladaco, "Sentiments Analysis On Public Land Transport Infrastructure in Davao Region using Machine Learning Algorithms," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 685–690, Feb. 2020. https://doi.org/10.30534/ijatcse/2020/97912020

[14] Z. Huang and Y. Liang, "Research of data mining and web technology in university discipline construction decision support system based on MVC model," *Libr. Hi Tech*, p. LHT-09-2018-0131, Jun. 2019.

[15] Z. Zhang and Y. Dai, "Combination classification method for customer relationship management," *Asia Pacific J. Mark. Logist.*, vol. ahead-of-p, no. ahead-of-print, Jul. 2019. https://doi.org/10.1108/APJML-03-2019-0125

[16] M. Jayakrishnan, A. K. Mohamad, and A. Abdullah, "Enterprise Architecture Embrace Digital Technology in Malaysian Transportation Industry," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4, pp. 852–859, 2019.

[17] M. Jayakrishnan, A. K. Mohamad, and A. Abdullah, "A Systematic Literature Review in Enterprise Architecture for Railway Supply Chain of Malaysia Transportation Industry," *Int. J. Eng. Res. Technol.*, vol. 12, no. 12, pp. 2473–2478, 2019.

[18] V. Diamantopoulou and H. Mouratidis, "Applying the physics of notation to the evaluation of a security and privacy requirements engineering methodology," *Inf. Comput. Secur.*, vol. 26, no. 4, pp. 382–400, Oct. 2018. https://doi.org/10.1108/ICS-12-2017-0087

[19] N. M. Geddes, "Adoption of renewable energy technologies (RETs) using a mixed-method approach," *J. Model. Manag.*, vol. ahead-of-p, no. ahead-of-print, Feb. 2020. https://doi.org/10.1108/JM2-03-2019-0082

[20] G. Coller, M. L. Frigotto, and E. Costa, "Management control system and strategy: the transforming role of implementation," *J. Appl. Account. Res.*, vol. 19, no. 1, pp. 141–160, Feb. 2018. https://doi.org/10.1108/JAAR-01-2016-0002