ABSTRACT

Sentiment analysis is the use of natural language processing, statistical or machine learning method to identify the attitude or emotion of the data whether it is positive or negative. As the internet are the main platform for people to seek and give opinions, the emergence of the data in user-based content website such as Twitter and Facebook could be used to analyse the emotion easily. This could change the way of getting the opinions of the people such as questionnaire or surveys which would involve longer and exhausting time to collect. It is important to gain the opinions of people towards movie as they could describe how they feel about the movie and what to expect in the future. The main objective of this research is to use machine learning Lingpipe Classifier which is an open source type of probabilistic classifier to analyse the sentiment of the data based on three (3) movies namely JRevolusi, RockBro and Boboiboy. Lingpipe works by analysing the data into two parts of analysis; first to train the data with the classifier and then gave a score for each data based on their attitude whether it is positive or negative. Results show the accuracy trained with the classifier on each movie above 80% consist of 1255 tweets extracted from Twitter. The classifier is trained with 950 datasets consist of English vocabulary and adjectives, with 50 of them are added with internet slangs and Malay language. Classification results showed that generally each movie received positive reviews from the users in Twitter. With the classifier favour the movie reviews, we could relate the review with the movie box office.

Keywords: Online social media, sentiment analysis, text mining, natural language processing, Lingpipe classifier
1.0 INTRODUCTION

Sentiment analysis or opinion mining have been widely known as the study consist of sentiments, attitudes, reaction and evaluation of the text content. It has brought a major change in the way opinion of the people are collected; method before include questionnaire and interview. The analysis can be applied in broad areas such as marketing, predicting trends in economy and services (Hussein, 2016, Khardel et al, 2016).

Such that we know that large number of opinion we could get through the famous microblogging site, Twitter. Because of the popularity of Twitter as an information sources, it led to the development and research in many areas. This could identify Twitter not only as a normal social media, but even connecting all the opinion to extract latest information about anything and current trending topics. Thus, huge and varied amount of knowledge can be extracted from the “tweets” (Injadat et al, 2015, Liu, 2012).

Large number of users expressing their voice about some movies they watch in Twitter. Such as in this project, the analysis will conclude on what people think about the movie that has been analysed. The opinion of the users will be then classified into positive or negative opinion by the machine learning algorithm. With this the producer may know the strength and weaknesses in the movie, thus can predict what people wanted in the movie next time.

Movie reviews are an important way to gauge the performance of a movie. While providing a numerical or stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews from the opinions of the people in social media is what gives us a deeper qualitative insight on different aspects of the movie. From that, a deeper analysis of the opinions can tell us if the movie in general meets the expectations of the people, thus giving the production company the ideas what generally people want in their movies.

The objectives of this research are to analyse the opinions and attitude of people towards the movie, to predict the trends and real-world events in the movie business and to study the relationship between the communication intensity in social media and the performance of the movie.
This project studied three (3) movies produced by a Malaysian company namely Primeworks Studio Sdn Bhd. Those three (3) movies which have been screened in the cinemas in Malaysia are: *BoBoiBoy: The Movie*, *J Revolusi* and *Rock Bro!*

### 2.0 METHODS

This section explained the method and design in Lingpipe Classifier to perform sentiment analysis. The design for data collection method has also been defined by their steps and flow. The system flow of this research project is constructed through five (5) steps: Collection of the review based on the topic from tweets are preprocessed and executed as part-of-speech tagging. Furthermore, the adjective from the tweets are located to execute the finding of their sentiment and count them. The counts are used as attributes in the process of classification. Before started the classification tasks, the tweets must be transformed into a format that is organized for the classification. Finally, the classification is performed by the machine learning algorithm which is Lingpipe Classifier (Pang et al, 2008).

#### 2.1 Method Briefing

Users on Twitter technically generate about 500 million tweets per day and around 200 billion tweets per year. It is one of the most social media used after Facebook and Instagram, which this will provide the project to retrieve the data from the wanted subject. This is all available in the Twitter App Management metadata that is available through the public APIs for free. The data would be extracted from the user tweets and the sentiment or emotion from the tweets will be analysed whether it is positive or negative. Twitters API are divided into two types: REST API and Streaming API (Kumar et al, 2016).

To use the Streaming API of Twitter to extract the tweets, the registration and authentication should be applied at Twitter’s developer website (https://dev.twitter.com/index), which will be redirected to Twitter Application Management (https://apps.twitter.com/). Official Twitter account is needed to interact with the API and gain all access tokens. The tokens act as the session in the process of data mining from Twitter to Java Eclipse. Registered users are given the choice to change the access privilege and this function also limited to Streaming API. The most efficient settings are by enabling read, write and access direct message.
2.2 OAuth

For the users to use its API, Twitter Developer introduced OAuth as the authentication model used to provide authorized access. OAuth defined as Open Standard for Authentication which proves the user information intact and access to protected information. As vulnerable password is, OAuth approach using three-way handshake that would underline the concept of OAuth to provide security on data. This would eliminate any possibilities the user’s data being shared with other 3rd party application without user’s permission.

In this research, streaming API will be used for the data and opinion collection. Streaming API provide developers to access entire Twitter stream on a small subset. The API must run on a separate thread than the code used to maintain the stream to provide real-time stream of Twitter data. Initially, they would deliver tweets based on search terms or specific words in real-time.

To use streaming API require a login into some Twitter account and will provided with OAuth authentication. GET and POST methods deliver common streaming of API at the endpoint.
Table 1: Streaming API endpoints

<table>
<thead>
<tr>
<th>Method</th>
<th>Resource</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET</td>
<td>user</td>
<td>Stream tweets of a user</td>
</tr>
<tr>
<td>GET</td>
<td>status/sample</td>
<td>Streams sample of tweets</td>
</tr>
<tr>
<td>POST</td>
<td>statuses/filter</td>
<td>Streams tweets matching predicate</td>
</tr>
</tbody>
</table>

2.3 Twitter4j

Twitter4j ([http://twitter4j.org/en/](http://twitter4j.org/en/)) is an open source Java library which provide API for the convenience to access Twitter API. In the API, the interaction available includes posting a tweet, getting the timeline of user, interaction in direct messages and search for tweets etc. The library eases the mining operation which could also ensure the security of a user configured with OAuth credentials.

The Streaming API is very reliable in real time as it handles the thread creation and listens to events. These will create a listener which listens to tweet updates from users (Sun et al, 2016).

2.4 Classification of data

In Lingpipe Classifier, the classification of data will go through two part of analysis which are basic polarity analysis and basic subjectivity analysis respectively. The first part of the analysis would focus on identifying the adjectives in the tweet, followed by producing the sentiment score for each tweet by the user on the subject.

2.4.1 Basic Polarity Analysis

The analysis would begin by training the collected dataset of tweets from Twitter and testing it with basic classifier of subjectivity data. The research includes additional dataset of modern internet language slang and also from Bahasa Malaysia slang that is familiar with defining the good or bad of the movie, as seen in the example of Table 2.
The constructor will set up the member variables using the command-line argument. Firstly, the
directory is just set to be the directory named txt_sentiment relative to the toplevel polarity data
directory given as the first command-line argument. The category array is initialized using the
directory names under txt_sentiment, which in this case are "positive" and "negative".

This method would run through the categories. It will then create a directory using the
polarity data directory and the name of the category. The text would then be read from the file
using the LingPipe utility method Files.readFromFile, and then used to train the classifier for the
specified category.

As each tweet is processed, the number of tests is incremented. Then the classifier is used
to produce a classification for a review string in a single line. Next, the classification of the text
subject from Lingpipe is extracted as the result category of classification. If the result category
matches the test category, the number of correct classifications is incremented.

### 2.4.2 Basic Subjectivity Analysis

The second form of sentiment analysis would determine the sentence whether it is positive or
negative based on its sentiment score. The negative score ranging from -0.001 to -1 while vice
versa for positive score. The analysis follows the same class pattern as the last analysis but with a
slightly different data format and addition of the classifier evaluation framework.

The remaining code in the train () method will compile the model to a file and transforms
the format, resulting the model being much faster at runtime. The first line would create an
evaluator from the classifier and the array of categories. When the evaluation loop is done,
thetoString() method have to be called on the evaluator to print out the results. The classifier will be them used to perform classification on the input sentence with reference to the reference category. The resulting scored classification is then added as an evaluation case with the specified reference category for computing results (Zhou, 2016).

3.0 RESULTS

As the streaming API is being executed with Twitter4j library, the selected query will mine the tweets. The extracted tweets are displayed along with the random id executed from OAuth, date and the time of tweets as seen in Figure 2.

Figure 2: Some of the extracted tweet from ‘JRevolusi’ query

After that, the first analysis is being done respectively according to each dataset of collected tweets from Twitter. The analysis involved with the classifier’s polarity training cases which include words and subjectivity of the text to start train with. Figure 3 shows the evaluation of polarity analysis:
<table>
<thead>
<tr>
<th>Movie title</th>
<th>Results of polarity analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRevolusi</td>
<td><img src="" alt="BASIC POLARITY DEMO" /></td>
</tr>
<tr>
<td></td>
<td>Training:</td>
</tr>
<tr>
<td></td>
<td># Training Cases=5000</td>
</tr>
<tr>
<td></td>
<td># Training Char=816958</td>
</tr>
<tr>
<td></td>
<td>Evaluating:</td>
</tr>
<tr>
<td></td>
<td># Test Cases=438</td>
</tr>
<tr>
<td></td>
<td># Correct=368</td>
</tr>
<tr>
<td></td>
<td># Correct=0.878</td>
</tr>
<tr>
<td>Rockbro!</td>
<td><img src="" alt="BASIC POLARITY DEMO" /></td>
</tr>
<tr>
<td></td>
<td>Training:</td>
</tr>
<tr>
<td></td>
<td># Training Cases=5000</td>
</tr>
<tr>
<td></td>
<td># Training Char=816958</td>
</tr>
<tr>
<td></td>
<td>Evaluating:</td>
</tr>
<tr>
<td></td>
<td># Test Cases=369</td>
</tr>
<tr>
<td></td>
<td># Correct=312</td>
</tr>
<tr>
<td></td>
<td># Correct=0.845</td>
</tr>
<tr>
<td>Boboiboy:</td>
<td><img src="" alt="BASIC POLARITY DEMO" /></td>
</tr>
<tr>
<td>Movie</td>
<td>Training:</td>
</tr>
<tr>
<td></td>
<td># Training Cases=5000</td>
</tr>
<tr>
<td></td>
<td># Training Char=816958</td>
</tr>
<tr>
<td></td>
<td>Evaluating:</td>
</tr>
<tr>
<td></td>
<td># Test Cases=450</td>
</tr>
<tr>
<td></td>
<td># Correct=362</td>
</tr>
<tr>
<td></td>
<td># Correct=0.804</td>
</tr>
</tbody>
</table>

Figure 3: Evaluation of polarity analysis

The second analysis of subjectivity are being done to detect the emotion of the text hence evaluate each tweet for the sentiment score. Note that the accuracy is based on the score given in both analysis and the original score for the data subjectivity, in the range of -0.10 to +0.10. Figure 4 shows the evaluation of subjectivity analysis:
<table>
<thead>
<tr>
<th>Movie title</th>
<th>Results of polarity analysis</th>
</tr>
</thead>
</table>
| **JRevolusi** | BASIC SUBJECUTIVITY DEMO  
Data Directory=C:\Users\YFP\Lingpipe\Data\tcv_jrevolusi  
Training,  
# Sentence plot=5000  
# Sentence quote=5000  
Compiling,  
Model file=subjectivity.model  
# Training cases=960  
# Training chars=1179085  
Evaluating,  
CLASSIFIER EVALUATION  
Total Count=489  
Total Correct=985  
Total pos=470  
Total neg=189  
Total accuracy=0.857 |
| **Rockbro!** | BASIC SUBJECUTIVITY DEMO  
Data Directory=C:\Users\YFP\Lingpipe\Data\tcv_rockbro  
Training,  
# Sentence plot=5000  
# Sentence quote=6000  
Compiling,  
Model file=subjectivity.model  
# Training cases=960  
# Training chars=1179085  
Evaluating,  
CLASSIFIER EVALUATION  
Total Count=969  
Total Correct=912  
Total pos=948  
Total neg=120  
Total accuracy=0.921 |
| **Boboiboy: The Movie** | BASIC SUBJECUTIVITY DEMO  
Data Directory=C:\Users\YFP\Lingpipe\Data\tcv_boboiboy  
Training,  
# Sentence plot=5000  
# Sentence quote=5000  
Compiling,  
Model file=subjectivity.model  
# Training cases=960  
# Training chars=1179085  
Evaluating,  
CLASSIFIER EVALUATION  
Total Count=450  
Total Correct=962  
Total pos=546  
Total neg=105  
Total accuracy=0.817 |

Figure 4: Evaluation of subjectivity analysis
The data would be evaluated and given the score based on how the machine learn from the training cases and dataset as seen in Figure 5.

Figure 5: Some of the classification result - JRevoluti

The following Table 3 shows some examples of the given sentiment score of classified tweets in different movies. Table 4 shows the results of the classification of tweets.

Table 3: Example of classified tweets

<table>
<thead>
<tr>
<th>Tweets</th>
<th>Sentiment Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malam ni nk layan #RockBro</td>
<td>0.649</td>
</tr>
<tr>
<td>Not too fond of rockbro new movie</td>
<td>-0.815</td>
</tr>
<tr>
<td>Sebelum ni kami ramai-ramai tengok #rockbro Seronok layan filem melayu</td>
<td>0.580</td>
</tr>
<tr>
<td>Yg ada power power macam bobaiboy the movie?</td>
<td>0.805</td>
</tr>
<tr>
<td>Well..Bobaiboy the Movie was a very good attempt.</td>
<td>0.753</td>
</tr>
<tr>
<td>Cerita orang Malaysia buat yang paling best setakat ni, citer Boboiboy The Movie. Haha. Tak sabar tunggu next movie</td>
<td>0.794</td>
</tr>
<tr>
<td>Tak best aaaa bobaiboy movie!</td>
<td>-0.227</td>
</tr>
</tbody>
</table>
Table 4: Classification of tweets

<table>
<thead>
<tr>
<th>Movie title</th>
<th>Positive</th>
<th>Negative</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRevolusi</td>
<td>270</td>
<td>168</td>
<td>0.857</td>
</tr>
<tr>
<td>Rockbro!</td>
<td>249</td>
<td>120</td>
<td>0.821</td>
</tr>
<tr>
<td>Boboiboy: The Movie</td>
<td>345</td>
<td>105</td>
<td>0.817</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>864</strong></td>
<td><strong>391</strong></td>
<td></td>
</tr>
</tbody>
</table>

4.0 DISCUSSION

The discussions of the result presented in this section include the visualization of the data produced. Generally, from the classification tasks, each movie receives positive reviews from the users in Twitter. The score basically favour each movie with more positive than negative emotion based on the classifier.

With the classifier favour the movie reviews, we could relate the review with the movie box office. JRevolusi and RockBro each gross an honourable amount while Boboiboy: The Movie topped the list of box office. This could relate with more positive review in their movie (345) compared to JRevolusi (270) and RockBro! (249).

![Figure 6: Sum of box office for the movies](image-url)
The scatter plot in Figure 7 shows the accuracy of the learning classifier in evaluating each of the text subjectivity in the range of -0.10 to +0.10. For example, the subjectivity of the word ‘power’ is decided around +0.70 based on the positive emotion word. However, several tweets with word ‘power’ classified with score out of the range, defining the accuracy for the word in both analysis.

![Figure 7: Scatter plot for subjectivity sentiment accuracy](image)

### 5.0 CONCLUSION

This research was defined as the sentiment analysis project on movies in Twitter using Lingpipe classifier. The project was aimed to collect and extract the data from Twitter about a subject and evaluate them into score that would define the data whether it contain positive or negative emotion

The objective for this research was to analyse the opinions and attitude of people towards the movie. Furthermore, it was also to study the relationship between the communication intensity in social media and the performance of the movie so a trend could be seen based on the opinions. Based on the result, the objective for the research have been summarized and achieved.

Using Java language in Eclipse platform, the method was implemented using Lingpipe classifier. Extraction method of the data from Twitter were fully using the Streaming API. In
summary, the objective of the project was achieved and sentiment analysis were successfully executed in this project.

In future, the research project could be expanded to a much broader subject, not only in entertainment industry but also politics and sports. The scope of the project would be easily defined given the availability and reliability of the data from Twitter. This would create an endless possibility in exploring more subjects in the field of data science.
REFERENCES


