# Using Logistic Regression Method to Classify Tweets into the Selected Topics

Indra, S.T[1], Liza Wikarsa, BCS, MComp[1], Rinaldo Turang, SKom, MKom[1]
Faculty of Engineering[1]
12013002@unikadelasalle.ac.id, lwikarsa@unikadelasalle.ac.id, tturang@unikadelasalle.ac.id

***Abstract*— Topics about health, music, sport, and technology are widely discussed in social network sites, especially in Twitter. Sharing information about those topics can enrich one's knowledge as well as increase the awareness of the current trends pertinent to the area of interests. Hence, this research aims to develop a web-based application that can classify tweets of netizens into these four categories of topics using one of machine learning methods called Logistic Regression. There are four main processes applied in this application that are fetching tweets, pre-processing, text feature extraction and machine learning. There are 1800 labeled tweets for each topic used as training data. Several processes were done in the pre-processing phase, including removal of URLs, punctuation, and stop words, tokenization, and stemming. Later, the application automatically converted the pre-processed tweets into set of features vector using Bag of Words. The set of features vector was applied to the Logistic Regression algorithm for the classification task. The trained classifier was then evaluated using 1800 tweets with 450 for each topic. Using Confusion Matrix, the results showed the accuracy of tweets classification into the selected topics is 92% which is considered very high.**

***Keywords—machine learning; topic analysis; health; music; sport; technology; text classification; logistic regression; Twitter;***

## I. INTRODUCTION

Twitter has more or less than 310 million monthly active users with total 500 million of tweets per day [1]. Tweet is a 140-characters message that can contain opinion or information about recent happenings or even user's emotion [2]. According to Wikarsa *et al.,* tweets are not properly structured because users do not care about spelling and grammatical construction when posting their tweets [3].

There are four most commonly discussed topics in Twitter such as health, music, sport, and technology. In support of this, Honigman pointed out that there are at least 40% of netizens worldwide, who dilligently access the heath information in the social media, are immensely influenced by the information that further affects the way they deal with their health care [4]. Music, on the othe hand, is the third of the top ten most frequently discussed topics in Twitter based on the research done by Franklin and his team [5]. Sport is widely discussed due to the subtantial number of sport lovers for different kinds of sports around the world [6]. Meanwhile, technology is greatly discussed in any social media like Facebook, Twitter and Instagram to obtain the latest information on technology developed at present. Technology plays a significant role on every aspect of our lives [7].

A hash tag is normally used to identifying the topic of the tweet. Having said that, many businesses can gather tweets based on the hashtags when they are in search of responses of netizens pertinent to their products and services. Governments and other interest groups can use the collected tweets to detect "human thinking patterns, group identification and recommendation, and also opinion about any specific topics of interests" [3, p. 1]. So, what if the tweet has no hash tag? They then have to read the whole tweet to know the content. It is surely time consuming and also requires a tremendous amount of efforts to understand the main content of the tweet.

In order to classify the contents of the collected tweets based on the selected topics, machine learning can be used to provide computational intelligences to the technology so it can learn and adapt from the given data independently. Application of machine learning methods, especially on social networking sites, can show varied results from auto recommendation, classification about specific interests, sentiments and so on [8]. There are several research done with regard to the use of machine learning to classiy text in the social network sites, including tweets classification for alcohol use-related or not [9] and real time tweets detection for small scale incidents [10]. Howeverit was found that the use of social network sites, especially Twitter, to classify tweets based on selected topics has not been done in a great extent. In addition, there is no application built that can classify tweets into those four topics automatically.

This research, therefore, aims to build a web based application which can fetch tweets directly from Twitter and classify tweets into four topics such as health, music, sport and technology automatically using one of machine learning methods called Logistic Regression. Logistic Regression method works by taking a set containing the feature of the input, in this case the sentence, and each of the features multiplied by the load or commonly referred to weight [8]. The previous research like tweets classification for alcohol use-related [9] and real time tweets detection for small scale incidents [10] revealed that Logistic Regression can generate better results in identifying and classifying text than any other methods like Naïve Bayes, Decision Tree, and others. In addition, this research will also use confusion matrix for the

 978-1-5090-4629-4/16/$31.00

classifier model to see the accuracy of the model using new data that is not used in previous training [11].

This research is divided into four main processes that are fetching tweets, preprocessing, text feature extraction and machine learning. It is hoped that this application can provide information to everyone, both individual and interest groups, by showing trends or recent happenings in tweets based on the four topics selected.

## II. RESEARCH QUESTION AND OBJECTIVES

### A. Research Question

How to build a web based text classification application that can classify Twitter's users' tweets into four selected topics such as health, music, sport, technology using Logistic Regression method?

### B. Research Objectives

Objectives of this research are as follows:

1. To be able to fetch tweets from Twitter.
2. To be able to perform preprocessing process to fetched tweets.
3. To be able to convert preprocessed tweets into set of features vector.
4. To be able to classify tweets into four topics correctly with high accuracy rate.
5. To present the classified tweets according to the selected topics in a useful format, such as a graph or table by showing the accuracy of text classification.

## III. LITERATURE REVIEW

### A. Fetching Tweets

According to Wei, fetching tweets is a process of getting the tweets directly from Twitter server using Twitter API (Application Programming Interface) based on keywords used [12]. Figure 1 depicts how Twitter API works.

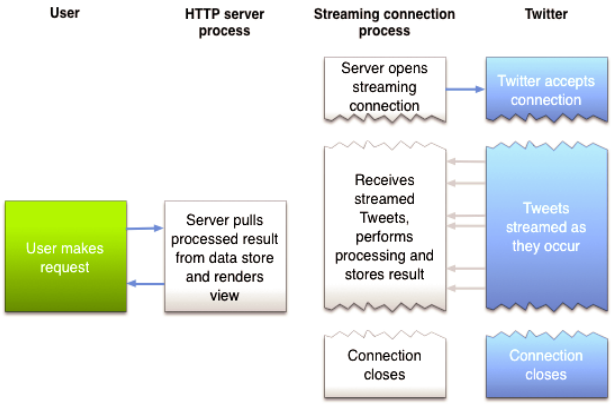



Figure 1. Fetching Tweets using Twitter API *[13]*

### B. Preprocessing

Preprocessing is a process of cleaning noise like links, punctuation or stop words that does not contain any useful information in the text [14]. There are few steps of preprocessing used in this research such as [15]:

- Remove URLs, to remove unwanted URL like http, https or something like them in the text.
- Remove punctuation, to remove punctuation marks like ",", ".", ":", ";" or something like them in the text.
- Tokenizing, to break text into each word.
- Remove stop words, to remove words like *a, most, and, is* and so on in the text because those words do not contain any useful information.
- Stemming, to change a word in the text into its base term or root term. Example, happiness to happy.

### C. Text Feature Extraction

Text Feature Extraction is a process of converting text into set of features in real number form side a vector that will be used as input for classification [16]. The process of extracting feature from text in this research used Bag of Words model that transforms all texts into a dictionary consist of all words appear in all texts. It later creates set of features in real number form inside a vector for each text where the value of each feature inside vector will be based on the frequency of each word counted in the text [17].

### D. Machine Learning

Machine Learning is one of computer science branches that focuses on providing the technology the ability to learn and adapt from given data so technology can learn and grow independently without being programmed explicitly by developers [8]. One of machine learning methods named Logistic Regression classifier is used for classification task in this research.

#### 1) Logistic Regression

Logistic Regression is one of many machine learning methods that works by taking input and multiplied the input value with weight value [8]. It is a classifier that learns what features from the input that are the most useful to discriminate between the different possible classes [18].

##### a) Computation of Logistic Regression Model

Logistic Regression is a discriminative model which means computing $P(y|x)$ by discriminating among the different possible values of the class $y$ based the given input $x$. The equation for this is as shown below:

$$P(c|x) = \sum_{i=1}^{N} w_i . f_i \quad \text{(i)}$$

The value of $P(y|x)$ cannot be actually calculated directly using the previous formula because it will result in value from $-\infty$ to $\infty$ which means it will not give output between value 0 and 1. To generate a value of an output that is in between value 0 and 1, the following *exp* function is used:

$$P(c|x) = \frac{1}{z} \exp \sum_i w_i . f_i \quad \text{(ii)}$$

 978-1-5090-4629-4/16/$31.00

To change the normalization factor $Z$ and specify the number of features as $N$ is as follows:

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i.f_i)}{\sum_c \exp(\sum_{i=1}^{N} w_i.f_i)} \quad \text{(iii)}$$

It is common in language processing to use binary-valued features. The features are not just a property of the observation $x$ but also are a property for both the observation $x$ and the candidate output class $c$. Thus instead of $f_i$ or $f_i(x)$, $f_i(c,x)$ is used whereby feature $i$ from the class $c$ is assigned to be the given input of $x$. The final equation for computing the probability of $y$ being of class $c$ given $x$ is:

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i.f_i(c,x))}{\sum_{c' \epsilon C} \exp(\sum_{i=1}^{N} w_i.f_{i(c',x)})} \quad \text{(iv)}$$

*b) Conditional Maximum Likelihood*

Conditional Maximum Likelihood method is used by Logistic Regression as the estimator of weight or $w$ value. This method works by choosing $w$ value that maximizes the probable value of class $y$ given the observation $x$. Below is the equation of *Conditional Maximum Likelihood* [18]:

$$\hat{w} = argmax_w \sum_j \log P(y^j|x^j) \quad \text{(i)}$$

$$L(w) = \sum_j \log P\left(y^j \middle| x^j\right) = log \sum_j \exp(\sum_{i=1}^{N} w_i f_i\,(y^{(j)}, x^{(i=j)})) - log \sum_j \sum_{y' \epsilon Y} \exp(\sum_{i=1}^{N} w_i f_i\,(y^{(j)}, x^{(i=j)})) \quad \text{(ii)}$$

*2) Confusion Matrix*

Confusion Matrix is often used in classification task to measure the accuracy rate of the used classifier. Confusion Matrix uses new data that is not used in training process [11]. Below is the example of using Confusion Matrix in classification task:

| n = 165 | PREDICTED: NO | PREDICTED: YES | |
|---|---|---|---|
| ACTUAL: NO | TN = 50 | FP = 10 | 60 |
| ACTUAL: YES | FN = 5 | TP = 100 | 105 |
| | 55 | 110 | |

Figure 2. Confusion Matrix Example *[11]*

From the example above, the calculation to determine the accuracy rate of the used classifier can be formulated as follows:

$$Accuracy = \frac{(TP+TN)}{total} = \frac{(100+50)}{165} = 0.91 \quad \text{(i)}$$

*E. Related Works*

There are few related works used as references for the development of this application as shown in the following table.

Table 1. Related Works

| # | Researcher (Year) | Title | Description |
|---|---|---|---|
| 1 | Axel Schulz, Petar Ristoski and Heiko Paulheim (2013) [10] | I See a Car Crash: Real-time Detection of Small Scale Incidents in Microblogs | Real-time identification of posts related to small scale incidents using SVM, JRip and Naïve Bayes |
| 2 | Yin Aphinyanaphongs, Bisakha Ray, Alexander Statnikov dan Paul Krebs (2014) [9] | Text Classification for Automatic Detection of Alcohol Use-Related Tweets | Classifying users tweets to detect tweets related to alcohol using Multinomial Naïve Bayes, SVM, Bayesian Logistic Regression and Random Forests. |
| 3 | Danesh Irani , Steve Webb, Calton Pu (2010) [19] | Study of trend-stuffing on Twitter through text classification | To identify and classify tweets that may or may not relate to trend stuffing using *Naïve Bayes,* C4.5 *Decision Tree* and *Decision Stump* |

## IV. DESIGN AND IMPLEMENTATION

*A. Requirements*

The following will enlist the specification requirements for the web based application that can fetch tweets from Twitter and classify tweets into four selected topics such as health, music, sport and technology using Logistic Regression.

*1) Application can fetch tweets by selected topics using Twitter API.*

*2) Application can store fetched tweets into database.*

*3) Application can perform preprocessing that consists of 5 steps described before.*

*4) Application can perform text feature extraction to tweet by converting tweet into set of features in real number form inside a vector which will be used as input.*

*5) Application can be used to give label (topic) to each tweet used as training data.*

*6) Application can learn through training process conducted using training data.*

*7) Application can perform evaluation process to measure accuracy rate of the trained classifier.*

*8) Application can perform testing process using the trained classifier to classify tweets to four topics such as health, music, sport and technology automatically.*

*B. Design Modeling*

This particular section will briefly explain the modeling of the tweets classification application by showing all the processes and datasets used in the application. Once the tweets

 978-1-5090-4629-4/16/$31.00

have been successfully fetched from Twitter using Twitter API, these data will go to the preprocessing process. When the preprocessing process is done, it later moves on to the training process where the tweets for training dataset will be used to training the classifier that uses Logistic Regression. After that, the trained classifier is evaluated using Confusion Matrix to show the accuracy rate of the trained classifier. Finally, the trained classifier can be used to automatically classify new tweets to the four selected topics as well as showing the probability value of each topic in a tweet.

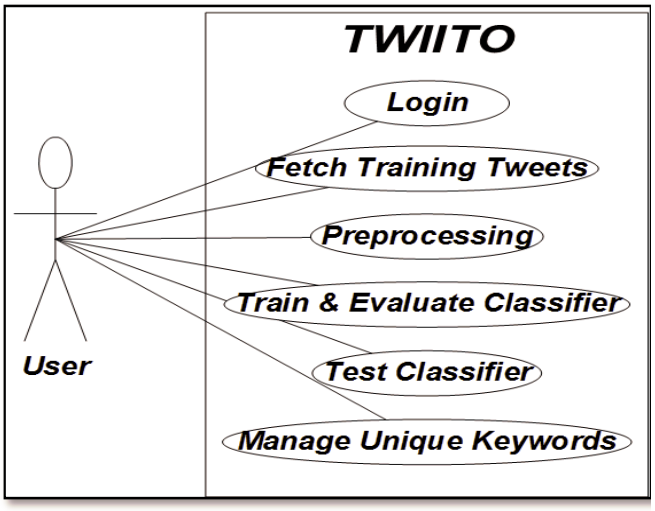



Figure 3. Use Case Diagram

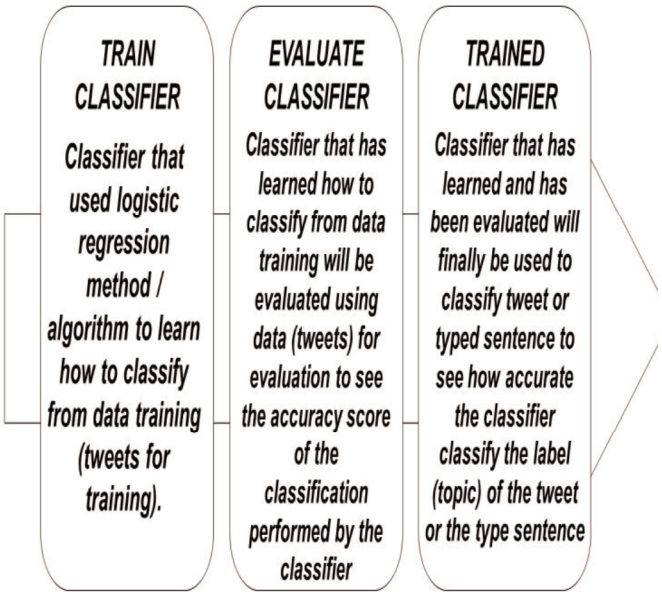



Figure 4. Datasets for the Tweet Classification Application

In the design process, the data model is divided to three datasets that are training dataset, evaluation dataset and testing dataset. Training dataset is used to training the classifier, evaluation dataset is 20% of data taken from training dataset used to evaluate the trained classifier and testing dataset is used to letting the trained classifier classify the tweets into one of the four selected topics automatically.

### C. Development Phases

Figure 5 below shows the processes inside the tweets classification application.

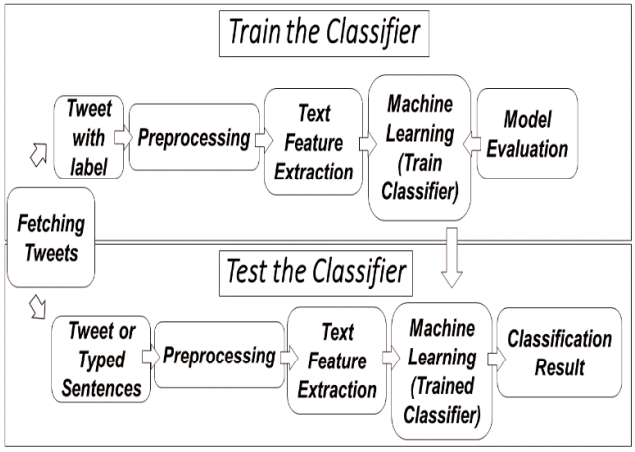



Figure 5. Processes Inside the Tweets Classification Application

*1) Fetching Tweets*

It is done using the Twitter API with additional filters based on keyword used for each topic.

*2) Preprocessing*

Once the tweets are successfully fetched, then the tweets need to be cleaned in preprocessing process. After the tweets are cleaned, the next step to do is to pass the tweets to text feature extraction process.

*3) Text Feature Extraction*

Using Bag of Words model to convert tweets into set of features vector in real number form. The set of features vector then passed to Logistic Regression classifier.

*4) Machine Learning (Train Classifier)*

*a) To make classifier learn from the tweets for training so it can classify the tweets correctly, is first using equation (ii) in Logistic Regresssion, for example given one tweet with class or label health:*

- *P*(*health | study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
  = *exp* (-3.4*1+-0.4*1+2.2*2+-0.56*1+-1.12*1+1.2*1+-2.2*1+-0.12*1+1.1*1+-0.6*1+0.4*1+1)
  = *exp* (-0.3) = 0.74.
- *P*(*music | study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
  = *exp* (-2.4*1+-0.4*1+2.2*2+-0.56*1+-1.12*1+1.2*1+-1.78*1+-0.12*1+1.1*1+-2.6*1+0.4*1+1)
  = *exp* (-0.88) = 0.41.
- *P*(*sport | study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
  = *exp* (-3.4*1+-0.4*1+2.2*2+-0.56*1+-1.76*1+1.2*1+-2.2*1+-0.82*1+1.1*1+-0.6*1+0.4*1+1)
  = *exp* (-1.64) = 0.19.
- *P*(*technology | study suggest gmo wheat silence permanently damage human gene result serious health issue health*)

 978-1-5090-4629-4/16/$31.00

= *exp* (-3.4*1+-0.4*1+2.2*2+-1.56*1+-1.12*1+1.2*1+-2.2*1+-0.12*1+1.1*1+-0.6*1+0.4*1+1)
= *exp* (-1.3) = 0.27.

*b) After calculating the probability value of each label to the given tweet, the next step is to using equation (iv):*

- *P*(*health* | *study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
$= \frac{0.74}{(0.74+0.41+0.19+0.27)} = 0.46$
- *P*(*music* | *study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
$= \frac{0.41}{(0.74+0.41+0.19+0.27)} = 0.25$
- *P*(*sport* | *study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
$= \frac{0.19}{(0.74+0.41+0.19+0.27)} = 0.11$
- *P*(*technology* | *study suggest gmo wheat silence permanently damage human gene result serious health issue health*)
$= \frac{0.27}{(0.74+0.41+0.19+0.27)} = 0.16$

*c) Training results f*rom calculation can be seen that the probability value of the tweet being in class or label health is higher than the probability of the tweet being in other classes. The value will gradually increase by the time classifier has successfully trained using all tweets for training.

*5) Machine Learning (Evaluate Classifier)*

To evaluate classifier, evaluation dataset which is 20% of the collected data from training dataset is used to measuring the accuracy rate of the trained classifier.

## *D. Implementation of Application Interfaces*

The next three figures will show the interfaces and the functionalities provided in this application.


Figure 6. Fetching Training Tweets


Figure 6. Train and Evaluate Classifier

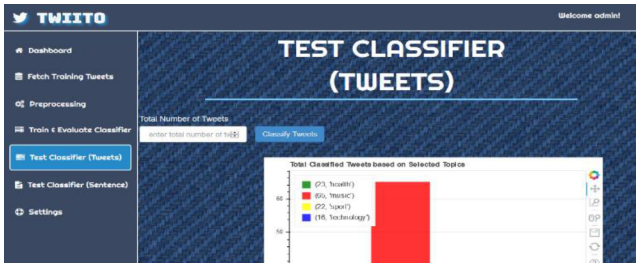
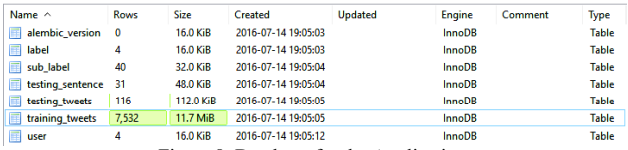



Figure 7. Test Classifier

## *E. Implementation of Database*



| Name | Rows | Size | Created | Updated | Engine | Comment | Type |
|---|---|---|---|---|---|---|---|
| alembic_version | 0 | 16.0 KiB | 2016-07-14 19:05:03 | | InnoDB | | Table |
| label | 4 | 16.0 KiB | 2016-07-14 19:05:03 | | InnoDB | | Table |
| sub_label | 40 | 32.0 KiB | 2016-07-14 19:05:04 | | InnoDB | | Table |
| testing_sentence | 31 | 48.0 KiB | 2016-07-14 19:05:04 | | InnoDB | | Table |
| testing_tweets | 116 | 112.0 KiB | 2016-07-14 19:05:05 | | InnoDB | | Table |
| training_tweets | 7,532 | 11.7 MiB | 2016-07-14 19:05:05 | | InnoDB | | Table |
| user | 4 | 16.0 KiB | 2016-07-14 19:05:12 | | InnoDB | | Table |

Figure 8. Database for the Application

# V. TESTING

The following table will show the accurate rate when testing is performed on the evaluation data.

Table 2. Accuracy Rate Based on Evaluation Data Using Same Amount of Unique Keywords per Topic

| | Total of Training Data Tweets | Total of Evaluation Data Tweets | Total Unique Keywords per Topic | Accuracy Rate (%) |
|---|---|---|---|---|
| 1. | 6000 | 1500 | 10 | 92.93% |
| 2. | 8000 | 2000 | 10 | 93.15% |
| 3. | 9000 | 1800 | 10 | 93.38% |

Table 3. Accuracy Rate Based on Evaluation Data Using Different Amount of Unique Keywords per Topic

| | Total of Training Data Tweets | Total of Evaluation Data Tweets | Total Unique Keywords per Topic | Accuracy Rate (%) |
|---|---|---|---|---|
| 1. | 6000 | 1500 | 15 | 85.6% |
| 2. | 8000 | 2000 | 15 | 87.1% |
| 3. | 9000 | 1800 | 15 | 88.9% |

The results of testing performed using Confusion Matrix on evaluation data show that using the same amount of unique keywords per topic for training and evaluation data can give a higher accuracy rate than using different amount of unique keywords per topic. In summary, the total of unique words and

 978-1-5090-4629-4/16/$31.00© 

total training data have significant impacts on the rate of accuracy achieved as it may increase or decrease the accuracy to certain levels.

## VI. CONCLUSION AND RECOMMENDATION

### A. Conclusion

1. Application can fetch tweets from Twitter using Twitter API.
2. Application can perform preprocessing which is divided into 5 parts such as remove URLs, remove punctuation, tokenizing, remove stop words and stemming to the tweets.
3. Application converts tweets into set of features vector in real number form.
4. Application can learn from tweets that have been labeled and used as training data.
5. Application can classify unlabeled tweets into one of four selected topics correctly.
6. The testing result on evaluation data shows that the total of unique words and training data have significant impacts on the accuracy rate achieved.
7. Growing volumes and varieties of available data in Twitter can be further classified into four selected topics through iterative learning from the datasets without being explicitly programmed where to look. In other words, it requires little human intervention due to its ability to independently adapt.
8. The implementation of Logistic Regression in the application showed that this method is able to automatically build analytical model that can be applied to new datasets resulting in a higher rate of accuracy in classifying tweets into selected topics.

### B. Recommendation

1. Tweets classification application is later expected to be able to classify the tweets based on selected topics as well as selected emotions contained in the tweet.
2. Application expected to eventually be able to classify tweets into selected topics in more specific way (e.g. music about pop, music about rock, etc).
3. Application can automatically delete tweet that does not contain selected topics in it
4. Application expected to be used in other social media like Facebook, Instagram or Path, making it easier for users to be able to know what is being said by other users automatically in their statuses or posts shared.

## REFERENCES


[1] The Verge, "The Verge," More People Reportedly use Snapchat than Twitter Everyday, 2 June 2016. [Online]. Available: http://www.theverge.com/2016/6/2/11839394/snapchat-passes-twitter-in-daily-active-users. [Diakses 15 June 2016].

[2] Social Media Today, "Social Media Today," Twitter 101: What is Twitter Really About?, 12 April 2013. [Online]. Available: http://www.socialmediatoday.com/content/twitter-101-what-twitter-really-about. [Diakses 15 April 2016].

[3] L. Wikarsa, S. N. Thahir dan R. Mandala, "A Text Mining Application of Emotion Classifications of Twitter's Users Using Naïve Bayes Method," IEEE, Indonesia, 2016.

[4] Referral MD, "Referral MD," 24 Oustanding Statistic & Figures on How Social Media has Impacted the Health Care Industry, August 2013. [Online]. Available: https://getreferralmd.com/2013/09/healthcare-social-media-statistics/. [Diakses 15 April 2016].

[5] Brandwatch, "Brandwatch," Social Media is Revolutionising the Music Industry, 29 August 2013. [Online]. Available: https://www.brandwatch.com/2013/08/social-media-the-music-industry/. [Diakses 16 April 2016].

[6] UK Essays, "UK Essays," Role of Sport in Modern Society Cultural Studies Essay, 23 March 2015. [Online]. Available: https://www.ukessays.com/essays/cultural-studies/role-of-sport-in-modern-society-cultural-studies-essay.php. [Diakses 16 April 2016].

[7] Pew Research Center, "Pew Research Center," How Social Media is Reshaping News, 24 September 2014. [Online]. Available: http://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/. [Diakses 16 April 2016].

[8] S. Raschka, "Python Machine Learning," dalam *Python Machine Learning*, Birmingham, Packt Publishing Ltd, 2015.

[9] Y. Aphinyanaphongs, B. Ray, A. Statnikov dan P. Krebs, "Text Classification for Automatic Detection of Alcohol Use-Related Tweets," dalam *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference*, Redwood City, 2014.

[10] A. Schulz, P. Ristoski dan H. Paulheim, "I See a Car Crash: Real-Time Detection of Small sclae Incidents in Microblogs," dalam *10th Extended Semantic Web Conference*, Montpellier, 2013.

[11] K. Markham, "Data School," Simple Guide to Confusion Matrix Terminology, 26 March 2014. [Online]. Available: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. [Diakses 10 April 2016].

[12] X. Wei, "Social Media & Text Analytics," Twitter API tutorial, 1 July 2015. [Online]. Available: http://socialmedia-class.org/twittertutorial.html. [Diakses 16 April 2016].

[13] Envato Tuts+, "Envato Tuts+," Building with the Twitter API Using Real-Time Streams, 17 November 2014. [Online]. Available: http://code.tutsplus.com/tutorials/building-with-the-twitter-api-using-real-time-streams--cms-22194. [Diakses 16 April 2016].

[14] Analytics Vidhya, "Analytics Vidhya," Steps for Effective Text Data Cleaning (with Case Study Using Python), 16 November 2014. [Online]. Available: http://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/. [Diakses 10 April 2016].

[15] S. Bird, E. Klein dan E. Loper, "Natural Language with Python," dalam *Natural Language with Python*, Sebastopol, O'Reilly Media Inc, 2009.

[16] J. Perkins, "Python 3 Text Processing with NLTK 3 Cookbook," dalam *Python 3 Text Processing with NLTK 3 Cookbook*, Birmingham, Packt Publishing Ltd, 2014.

[17] Udacity, "Udacity," Intro to Machine Learning, [Online]. Available: https://www.udacity.com/course/viewer#!/c-ud120/l-2892378590/e-3068768537/m-3008458614. [Diakses 10 April 2016].

[18] D. Jurafsky dan J. Martin, "Speech and Language Processing (3rd ed. draft)," 24 August 2015. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/7.pdf. [Diakses 16 April 2016].

[19] D. Irani, S. Webb dan P. Carlton, "Study of trend-stuffing on Twitter through text classification," CiteSeerx7M, Pennsylvania, USA, 2010.


 978-1-5090-4629-4/16/$31.00