Corpus Usage for Sentiment Analysis of a Hashtag Twitter

Conference Paper · October 2019

DOI: 10.1109/ICIC47613.2019.8985772

4 authors:

Herlawati Herlawati
Universitas Bhayangkara Jakarta Raya
81 PUBLICATIONS **353** CITATIONS
SEE PROFILE

Rahmadya T. Handayanto
Universitas Islam 45 Bekasi
37 PUBLICATIONS **318** CITATIONS
SEE PROFILE

Didik Setiyadi
STMIK Sinar Nusantara
41 PUBLICATIONS **186** CITATIONS
SEE PROFILE

Endang Retnoningsih
Institut Bisnis Muhammadiyah Bekasi
28 PUBLICATIONS **104** CITATIONS
SEE PROFILE

Corpus Usage for Sentiment Analysis of a Hashtag Twitter

Herlawati
Informatics Engineering
Universitas Bhayangkara Jakarta Raya
Bekasi, Indonesia 17121
mrs.herlawati@gmail.com

Rahmadya Trias Handayanto
Computer Engineering
Universitas Islam 45 Bekasi
Bekasi, Indonesia 17113
rahmadya.trias @gmail.com

Didik Setiyadi
Informatics Engineering
STMIK Bina Insani
Bekasi, Indonesia, 17114
didiksetiyadi@binainsani.ac.id

Endang Retnoningsih
Information System
STMIK Bina Insani
Bekasi, Indonesia, 17114
endang.retnoningsih@binainsani.ac.id

***Abstract*— Social media (Facebook, Instagram, Twitter, etc.) nowadays can be used for analyzing the objects, e.g. political views, products, services, etc. To understand the performance of an object, the sentiment analysis has been widely used to get the review from the consumers or users (positive or negative responses). Today, in big date era, which is a component of industry 4.0, many corpora are available and can be accessed freely. A corpus can be utilized to train the model through some methods. In this paper a Naïve-Bayes classifier was used to train a corpus from natural language toolkits (NLTK) corpora. As a case study, sentiment analysis for the sample movie "Avengers" was done from the twitter hashtag #avengersendgame. The paper also proposed the usage of a particular corpus to other different language implementations, e.g. for Indonesian language. Through the use of Tweepy and Pandas some Twitter tweets were retrieved and classified after pre-processing. The results showed the capability of the Naïve-Bayes classifier both for English and Indonesian language.**

Keywords—Naïve-Bayes Classifiier, Information Retrieval, Tweets, Tweepy, Pandas.

I. INTRODUCTION

Understanding the consumer needs is a step to create better product and service. Twitter and other social media can be used to get the information about the product and service. To process the information some algorithms have been widely implemented in sentiment analysis [1]–[3].

To check the twitter posts, the information retrieval method is needed. The sentiment analysis has been successfully implemented in various purposes [4]–[6]. With the text processing capability some regular expression can automatically process the posts before classification [7]–[10] [11], [12]. Similar to structured data, the unstructured data from twitter need a data mining method analysis, in particular classification [6]. Since the sentence whether it is positive or negative response difficult to be processed automatically through the system, the data need labelling process. Therefore the supervised classification is needed.

Corpora have been widely used and available in the internet. This text-based data can be used for another purpose. The language is the main problem, especially when implementation in non-English language. This study proposed the usage of corpus for sentiment analysis and the possibility to use for another language. In this case, the hashtag twitter was used for sentiment analysis.

Many classification methods, with their characteristics, have been implemented in a semantic analysis e.g. K-Nearest Neighbourhood (KNN) [7], Fuzzy C-Means (FCM) [8], Support Vector Machine (SVM) [9], [10], Naïve-Bayes Classification, etc [6]. The current study used Naïve-Bayes as the method to classify the tweets. Naïve-Bayes is not the only method for semantic analysis. Another method, e.g. SVM, KNN, etc., has also been widely used [8]. In this study, we focus on the automatic use of wrapping and analyzing the hashtag from twitter before classification using Naïve-Bayes method. This method uses probability as the basis for classification. The user not only sees the class but also the probability of the classification. This cannot be found in the hard-classification, e.g., KNN SVM [13]. The frameworks for corpus usage in different language were also proposed in this study. This research contributes to the utilization of a corpus from other open access data available in the internet for other purposes.

II. DATA AND METHODS

A. Data

This study used a corpus data for training the classifier. A corpus from previous research also been employed from Natural Language Toolkits (www.nltk.org), i.e. the Movie Review. This corpus contained more than 2000 reviews contained the near complete statements about a particular movie. It contains 1000 positive reviews and 1000 negative reviews. About 400 data was used for testing to check the accuracy. Each review has more than 200 words and has already been checked whether it is positive or negative response.

After having a good accuracy, the classifier will be used for classify other particular objects, e.g. posts, comments, and other statements from social medias (Facebook, Instagram, Twitter, etc.). This study used a text-based corpus both for training and testing data of a hashtag-tweet.

Another data was used to implement the trained data. A hashtag twitter, #avengersendgame, was used. Twitter

(www.twitter.com) is a social media contained a post of short sentence for expressing the idea, feel, comment, etc. A script in Python, was created to grab some tweets related to a hashtag which is discussed in the following section.

B. Methods

Naïve-Bayes classification is named after Thomas Bayes, a probability and decision theory specialist. The posterior probability can be calculated by the following equation.

$$P(H|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})}. \quad (1)$$

P(H|X) is posterior probability, of *H* conditioned on *X*. Similarly, *P(X|H)* is the probability of *X* conditioned on *H*. This usually called likelihood. The *P(H)* is prior probability of *H* and *P(X)*, probability of *X*, used for normalized.

Naïve term in Naïve-Bayes classification appeared because this method neglects some facts, especially the relation between one class to other classes [13]. However, this method with the probability characteristics still widely used.

Corpus is the meaningful information shared from any institutions. The data sometimes have already been cleaned by the institutions who share that corpus for their own purposes, e.g. data mining, semantic analysis, etc. It can be used by other institution for another purpose different with the first owner. The corpus used in this paper was a movie review. Every review has been checked whether positive or negative sentiment. Fig 1 shows the framework of training and testing the model.

Two data were prepared, namely, training and testing data. Each data has to be stemmed to exclude the prefix and suffix. Python has already been prepared for the Porter stemmer. After training, the tested-hashtag was used to check the accuracy of the trained-data. The prediction from training (positive or negative sentiment) was compared to the answer from trained-data. If the accuracy was good enough, the model can be used to check the comments retrieved from twitter.

The framework used only for the text-based corpus. Python was used both for training and grabbing/wrapping hashtag twitter. Python 2.7 was used since the Python 3.0 has not been served for natural language toolkit (NLTK). The benefit of Python use is that this language is open source, free, and many toolkits (machine learning, information retrieval, etc.) are available.

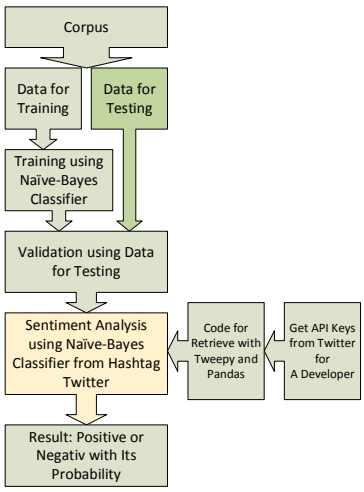

Fig. 1. Training and Testing Framework

After using the Tweepy module for grabbing/wrapping the tweets of a particular hashtag, the tweets were classified into positive or negative response. For case study, the hashtag "avengersendgame" was chosen. Following is a pseudocode to create the classifier.

```
[1] Importing the Naïve-Bayes Module and data trining from NLTK corpus.
[2] Define a function to extract feature.
[3] Preparing training data (movie_reviews from NLTK_DATA was chosen) with positive and negative review.
[4] Separating and labelling positive and negative reviews.
[5] Devide the data for training and testing
[6] Extract data for training
[7] Training the extract data using Naïve-Bayes classifier.
[8] Validate with test data.
[9] Find the most informative words from classifier object.
```

If the classifier is good enough, it can be used to classify other object. In this research a hashtag retrieved using the Tweepy module and was helped by other module: Pandas. Following is the procedure for retrieving some tweets.

1. Get Application Programme Interface (API) Keys (Consumer API Keys with API secret key) and Access Token (with Access Token Secret) from twitter (https://developer.twitter.com).
2. Install Tweepy and Pandas
3. Use code below to retrieved tweets from a hashtag. Pandas is useful for presenting the text in frame format and printing.

```
[1] import tweepy
[2] import pandas as pd
[3] import os
[4] #Twitter Access
[5] auth = tweepy.OAuthHandler( [Consumer API Keys],[Consumer API Secret])
[6] auth.set_access_token([Access Token Keys],[Access Token Secret])
[7] api = tweepy.API(auth,wait_on_rate_limit = True)
[8] df = pd.DataFrame(columns=['text'])
[9] msgs = []
[10] msg =[]
[11] for tweet in tweepy.Cursor(api.search, q='#avengersendgame', rpp=100).items(100):
[12] msg = [tweet.text]
[13] msg = tuple(msg)
[14] msgs.append(msg)
[15] print(msgs)
[16] df = pd.DataFrame(msgs)
[17] print(pd)
```

Line 11 shows the q as hashtag variable to be retrieved with numbers of tweets (e.g. 100 tweets). Some pre-processing can be used to gather the only text from retrieved tweets. Basic text processing is needed for cleaning the data

from hashtag twitter, especially to separate information from the xml code.

III. Result and Discussion

Python code (version 2.7) was implemented in Intel i5 processor with Windows 10 operating system. Data training from corpora were trained (1600 data points) with 400 data for testing. The accuracy result showed 73.5% with 10 most informative words: outstanding, insulting, vulnerable, ludicrous, uninvolving, astounding, avoids, fascinating, animators, and darker. Therefore, it was good enough to use the trained-model to classify the comment/review in Twitter.

Tweepy was used to retrieve the tweets in this study. Fig. 2 shows the result with 100 tweet samples of #avengerendgame hashtag. In the implementation, of course, we can use large samples to gather the benefit information of our products, brands, etc. As an information, the corpus is different with the hashtag in regard to the length of each review/comment. However, the Naïve-Bayes model successfully classify the tested-review whether positive or negative.

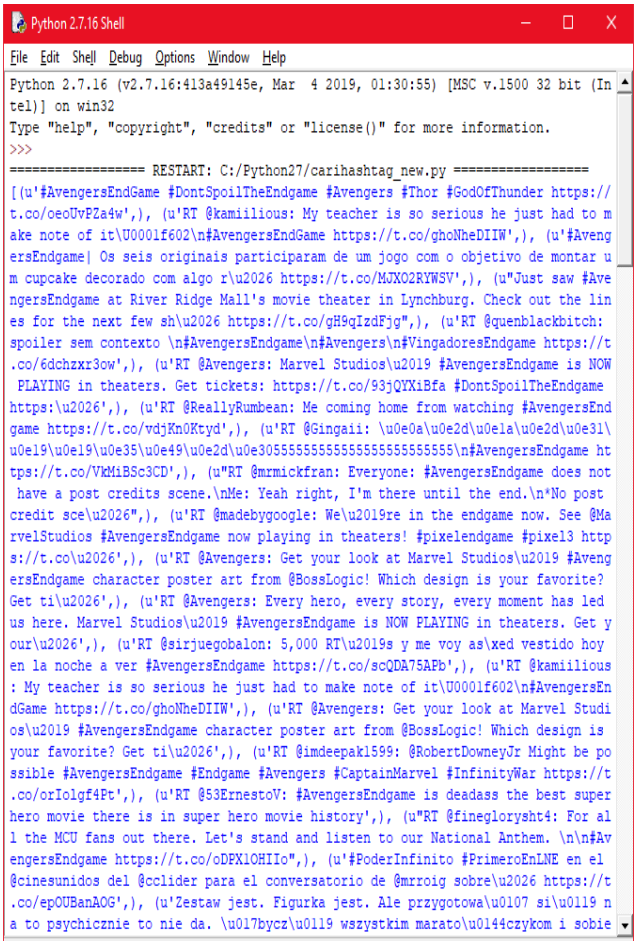

Fig. 2. Retrieving Tweets Result

Some processing methods used to eliminate the unnecessary alphabets (prefix, suffix, and infix) as well as the stop words which should be eliminated. The tweets inserted in Naïve-Bayes classification code to find the sentiment. Fig 3 shows the python result through the IDLE environment which shows the positive sentiment with probability of 0.51.

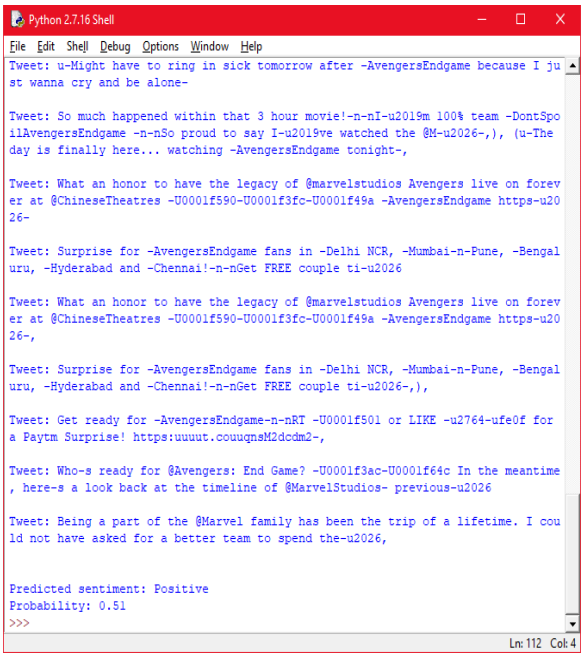

Fig. 3. Sentiment Analysis Result

The current result has a limitation. It only classify tweets in English. However, for other language, a translator is needed for a hashtag to be checked. Fig 4 shows the method to utilize an English corpus to another language, e.g. Indonesian.

Google translate (www.translate.google.com) or other translating tools are available for converting the corpus to another language. Another language, e.g. Indonesian, have their own stemming and lemmatization. Sastrawi is the famous toolkit for stemming in Indonesian language. However, another alternative can be used, i.e. translating the tested-tweets into English as shown in Fig 4.

Fig 4a shows the corpus utilization through the language translation. In this study, Indonesian language was used. Specific stemming is available in Indonesian, i.e. the Sastrawi. After training the classifier, Indonesian hashtag can be classified by the trained-classifier. In the other hand, Fig 4b shows another corpus utilization method. Instead of translating the corpus, this method translating the Indonesian hashtag into English. The English trained-classifier would be able to classified the converted-hashtag. For instant usage, the method shown in Fig 4b is more appropriate since translating the tested-hashtag need less processing than translating the corpus data. But if the classifier will be used frequently, the translating the corpus, stemming, and training the classifier should be done.

Of course, it is better if we create our own corpus. But some pre-processing tasks should be conducted to avoid the inaccurate result. Therefore, it is safe to use the established corpus. In this study the trained-data hardly separated into positive or negative and avoiding neutral because we have to decide whether a sentiment is positive or negative.

The accuracy of the classification in this study depends on the translator. Therefore, another grammar tools, e.g. Grammarly, smallseotools, etc. should be used, instead of just a word-to-word translator. For the implementation the user-friendly interface would make the management decide the next marketing strategy.

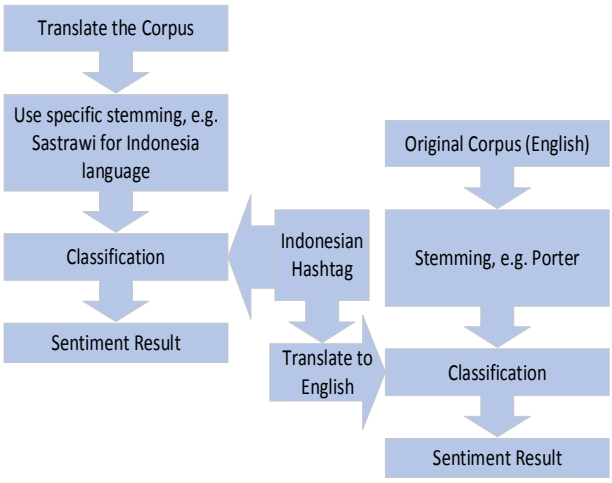

Fig. 4. Corpus usage for other languages Classification

For testing the proposed corpus utilization, a hashtag #aadc2 was retrieved. Fig 4b method was used by translating the Indonesian hashtag into English before classification. Fig 5 shows the grab result of #aadc2 hashtag as tested data. After pre-processing (stemming, deleting the standard xml code, and following the classifier standard script), the tested-hashtag were translated using a translation tool, for example www.translate.google.com (Fig 6).

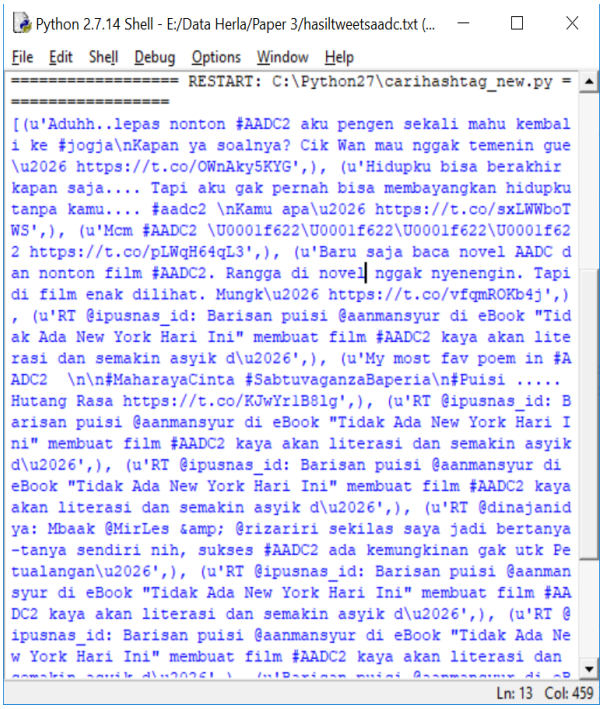

Fig. 5. Corpus usage for other languages Classification

After translation a sentiment analysis engine with the classifier inside it translated the tested-hashtag. This step is similar to analyzing #avengerendgame hashtag in section 2.

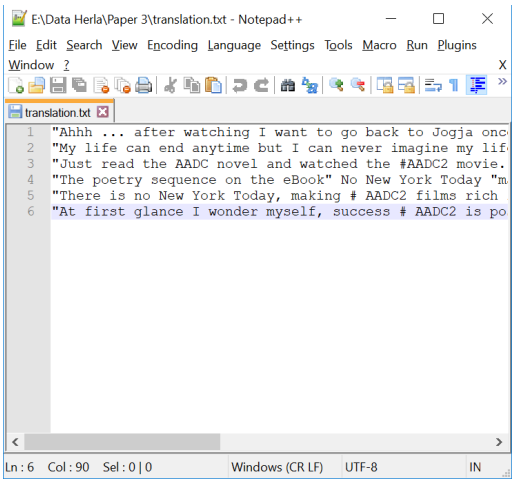

Fig. 6. Tested-hashtag after translation.

Fig 7 shows that the #aadc2 has positive sentiment with the probability of 0.66. Therefore, this Indonesian movie shows the good acceptance in the market.

Fig. 7. Sentiment Analysis Result.

This method can be used for other object, e.g. products, services, ideas, etc. In addition, this system can be used to detect some bad comments, bullying statements, terrorist discussion, etc. In addition, the combination method between Naïve-Bayes with other methods might be interesting to get better accuracy.

IV. Conclusions

Both the corpus and information from social medias is free and can be access easily today. With the Twitter API, the information from tweets can be mined and classify for sentiment analysis automatically. This study used Naïve-Bayes classifier to classify the post whether positive or negative which is trained from a corpus. The result showed that this method was able to classify some tweets retrieved from a particular hashtag into positive and negative response. The study also showed the ability of an English-based corpus doing the sentiment analysis for a non-English tested-hashtag, after classifying their translated-hashtags (with some different stemming and lemmatization methods according the tested-hashtag language). The study can be implemented for other purposes such as for analyzing the policies, business products, elections, and so on. Future research on image-based corpus

will be done from the social media e.g. Facebook, Instragram, etc., as well as with a hybrid method and a combination of other machine learning tools, e.g. Keras, TensorFlow, etc.

ACKNOWLEDGEMENTS

The authors thank to Universitas Bhayangkara Jakarta Raya, Universitas Islam 45, STMIK Bina Insani and others who participate in data analysis.

REFERENCES

[1] A. Srivastava, V. Singh, and G. S. Drall, "Sentiment Analysis of Twitter Data," *Int. J. Healthc. Inf. Syst. Informatics*, vol. 14, no. 2, pp. 1–16, 2019.

[2] K. Saranya and S. Jayanthy, "Onto-based sentiment classification using machine learning techniques," *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*, vol. 2018-Janua, pp. 1–5, 2018.

[3] B. Pang and L. Lee, "A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," 2002.

[4] M. Arora and V. Kansal, "Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis," *Soc. Netw. Anal. Min.*, vol. 9, no. 1, 2019.

[5] B. Dahal, S. A. P. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Soc. Netw. Anal. Min.*, vol. 9, no. 1, 2019.

[6] J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," *Inf. Fusion*, vol. 52, pp. 344–356, 2019.

[7] I. M. A. Agastya, T. B. Adji, and N. A. Setiawan, "Comparison of distributed K-means and distributed fuzzy C-means algorithms for text clustering," *Commun. Sci. Technol.*, vol. 2, no. 1, pp. 11–17, 2017.

[8] S. Ghosh and S. Kumar, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 4, pp. 35–39, 2013.

[9] D. Novitasari, I. Cholissodin, and W. F. Mahmudy, "Hybridizing PSO With SA for Optimizing SVR Applied to Software Effort Estimation," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 1, p. 245, 2016.

[10] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, p. 1502, 2016.

[11] R. E. Banchs, *Text Mining with MATLAB*. Barcelona: Springer, 2013.

[12] C. Manning, *An Introduction to Information Retrieval*, no. c. Cambridge: Cambridge University Press, 2009.

[13] M. Han, Jiawei; Kamber, *Data Mining, Concept and Techniques*. San Fransisco: Elsevier Inc., 2006.