Sentiment Analysis on Duterte Administration using Naïve Bayes Algorithm

Marjorie D. Agravante
Master in Information Systems
University of Southern Mindanao

Overview

Social networking sites became the outlet of internet users in expressing their sentiments. It is their main source of information, ideas and thoughts to give reaction to certain topics.

Twitter as a social media platform for expressing views and opinions on political issues is a relatively new occurrence. Nowadays, tweets are mostly composed of opinions on political and social issues. This can be seen as a good thing because people are being aware about what is going on around them. However, it also has disadvantages in the form of cultivating an even more aggressive online hate culture.

The administration of the former President Rodrigo R. Duterte (PRRD) has been the subject of discussion in this platform since the start of his term. Naïve Bayes Algorithm is a useful tool to determine the sentiment of the tweet.

Objectives

The study aimed to:

- determine the sentiments (positive and negative) of the tweets posted on the administration of the former President Rodrigo Roa Duterte (PRRD);
- create a word cloud – visual representation of words based on the tweets; and
- use the Naïve Bayes Algorithm to analyze the data.

Process and Methods

The study began by gathering tweets related to the PRRD Administration using the python library *snscrape*, a scraper for social networking services (SNS). The downloaded data (in CSV format) were divided into training and testing data. The percentage of training data sharing was 80% for the labeled data and 20% for the testing data. The training data was then labeled manually whether it is negative or positive by looking at the sentiment words contained while data testing was not labeled. The frequency of the sentiments must be balanced.

The data in testing set should not be in the training set so that it can be seen whether the classifier model is "right" in doing the classification.

The data were being processed using the python 3 scripts with Natural Language Processing (NLP) techniques to find the class polarity in words.

Stop words package is downloaded to be used later to clear the data from words that have no meaning or influence in the data.

The dataset was loaded into the system and plotted using *seaborn* library. Figure 1 shows the sentiment distribution of the train data. The figure can easily determine if the train dataset is balance.

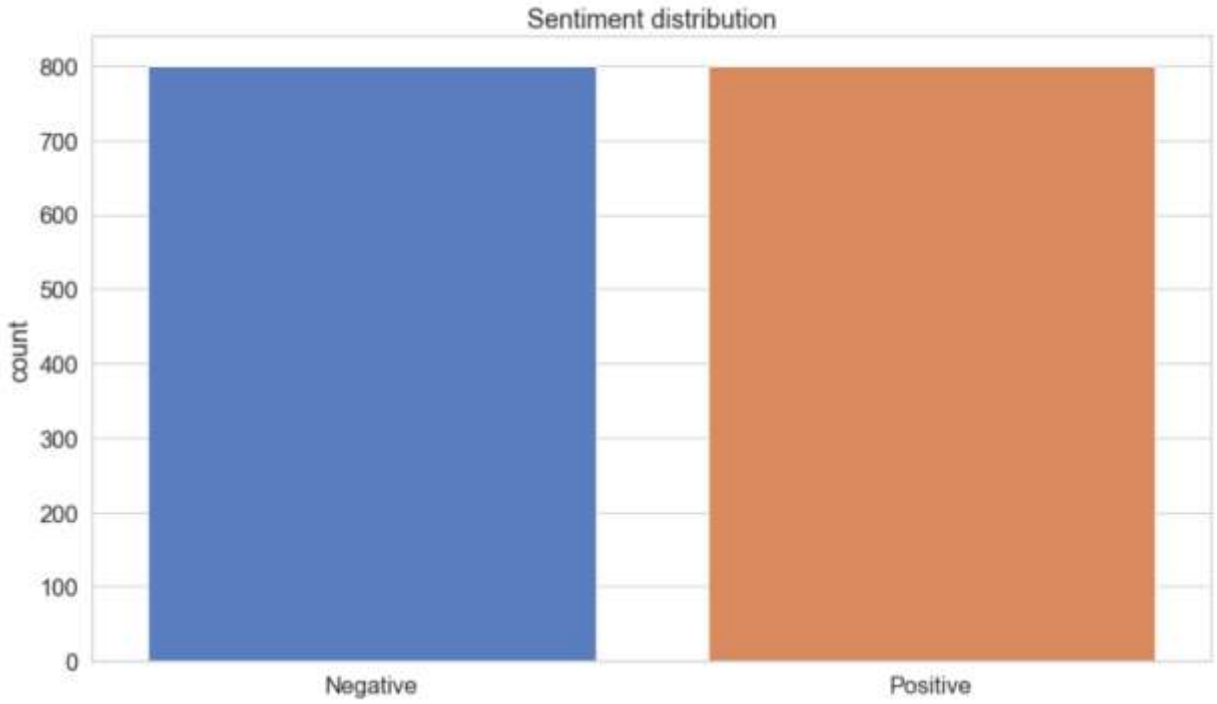

Figure 1 Sentiment Distribution.

After downloading the stop words, next is to show the word cloud which is based on the frequency of occurrence of the words used in the tweets. Figure 2 shows the visual representation of words.

Figure 2 Word Cloud of the Dataset.

Before entering the process of breaking words or tokenizing, the tweets must be cleared with words that have no meaning or influence. These words were being grouped as stop words. Stop words usually contain pronouns and conjunctions. The tokenizing process was done by forming the unigram token. To do the tokenization process and calculate the frequency of occurrence of words, Python was being used to remove the stop word.

Naïve Bayes Algorithm was used to classify the dataset. This model is a representation of knowledge that was used to predict new data classes that have never existed. Next was the assignment of train and test data. Then, using the Multinomial Naïve Bayes method to conduct the learning process of the Naive Bayes classifier in the training data, it can also be known the sentence classification stage in the test data by using the results of the word probability of the training data. Then, predict the sentiment of the test data.

Compute the accuracy score of algorithm's prediction. Then, get the classification report to measure the quality of predictions from a classification algorithm (see Figure 3). Then, compute and plot the confusion matrix to assess the performance of a classification model.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	171
1	1.00	0.95	0.97	149
accuracy			0.97	320
macro avg	0.98	0.97	0.97	320
weighted avg	0.98	0.97	0.97	320

Figure 3 Classification Report of Naive Bayes Algorithm.

Results and Discussion

Sentiment classification using Naive Bayes algorithm was done through two stages: the learning process stage and the classification stage.

From the Figure 2, it clearly shows the top three words related to PRRD administration which are DuterтePalpak, SalamatPRRD and DuterteLegacy.

The accuracy score of the algorithm is **97.5%** which means 97.5% true positive accuracy rate and there is 2.5% false positive rate. It implies a very good accuracy rate (Allwright, 2022).

Figure 4 shows the Confusion Matrix that has True Negative of 171, False Negative of 0, True Positive of 141 and False Positive of 8.

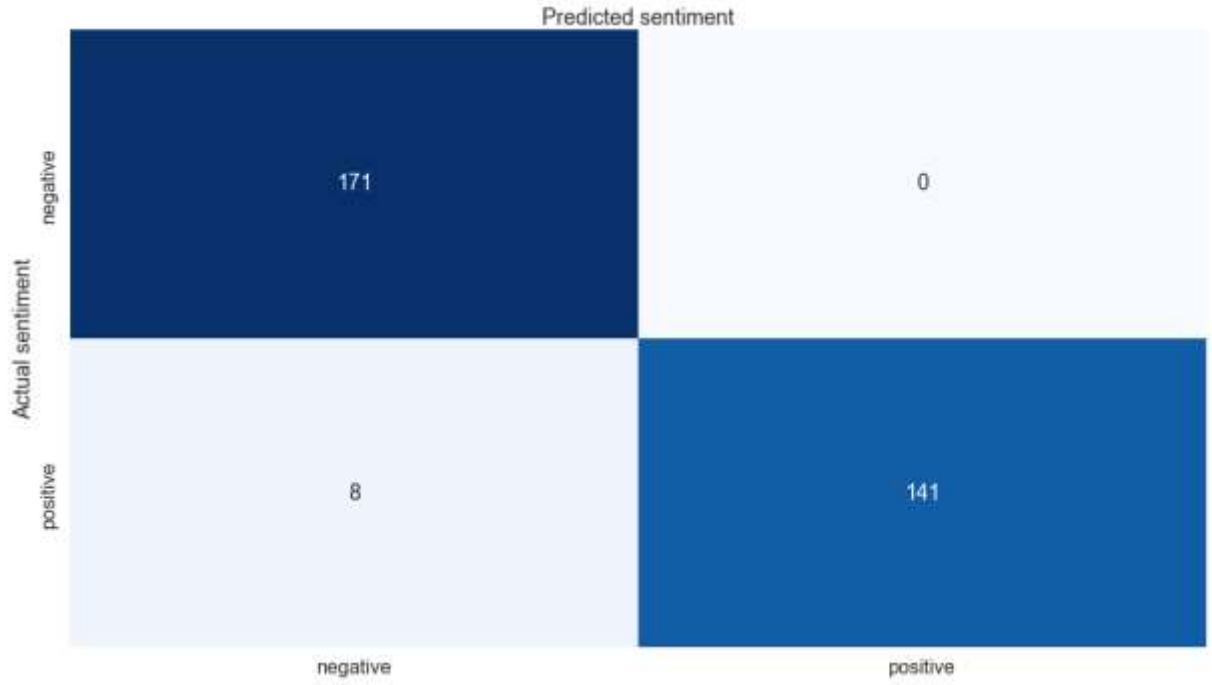

Figure 4 Confusion Matrix

Based on the data, it can be concluded that most tweets related to the PRRD administration result to negative sentiments.

References:

Allwright, S (2022, March 22). *What is a good accuracy score in machine learning?* Available at https://stephenallwright.com/good-accuracy-score/ [accessed August 30, 2022]

Hutchinson, A (2022, June 28). *New Study Shows Twitter is the Most Used Social Media Platform Among Journalists*. Available at https://www.socialmediatoday.com/news/new-study-shows-twitter-is-the-most-used-social-media-platform-among-journa/626245/#:~:text=According%20to%20a%20new%20Pew,being%20influenced%20by%20tweeted%20opinion [accessed August 30, 2022]

Martiti, CJ (2021). Implementation of Naive Bayes Algorithm on Sentiment Analysis Application. *Atlantis Press International B.V.*