**ANALYZING PUBLIC SENTIMENT TOWARDS A SPESIFIC POLITICAL PARTY: A SENTIMENT ANALYSYS PROJECT**

**BAB I**

**Introduction**

Seeing that the 2024 Indonesian Presidental Election is about to be held, it is important that votes of the citizens will decide the upcoming legislative members. We knew that the legislative candidates were going to represent different political parties. There are a lot of existing political parties in Indonesia, each of which also held their own records or history across the years. Whether it is the impact they have left or scandals that involved their representatives.  
 Because of it, there are vast kinds of opinions that are scattered around social media, especially on Twitter. Twitter itself consists of enormous amounts of posts, or we can call them tweets in forms of texts, filled with many debates about whether a specific politician should be in power or lower their image so they cannot be in power.

Therefore, it would be difficult for users to tell the responds of other users towards a certain political party. To classify and analyze the sentiments of citizens in forms of text, it would be easier with the help of a machine learning model, in which we propose a machine learning algorithm that is based on Logistic regression to analyze the tweets the users made on twitter and classify those tweets to be a positive, negative, or neutral opinion.

# **BAB 2**

# **Literature Review**

Sentiment Analysis is a way to analyze an opinion in the form of text towards a certain person, organization, or matter. The way a human thinks or processes data can be manipulated with various opinions. Whether it would be public statements or opinions towards certain candidates. Since sentiment analyses are usually held on social media, the datasets for it are taken from various social media such as Twitter, Facebook, Instagram, or others. Sentiment Analysis could be used to obtain feedback towards a certain organization, product, public figure, and other matters [1].

Twitter is known as one of the most popular social media platforms of the era, Because Twitter can provide microblogging up to 140 characters for a single tweet. Their users can post anything through their accounts if they obey the ToS (Term of Service), for example their daily routines, news, gossip, personal opinions, and others.

While logistic regression is a method capable of processing binary response data. When the response is binary, then it is made of ones and zeros, where one represents true and zero represents false. And actual values of ones and zeros can be wide apart based on the objectives of the study. (Hilbe, Joseph, Logistic Regression, 2011)

**BAB 3**

**METHOD**

The objective of this study is to build a machine learning model using the logistic regression method to classify opinions as positive, negative, or neutral based on the tweets from users, using the keywords “PDIP”, “PARTAI DEMOKRASI INDONESIA” or “PARTAI DEMOKRASI INDONESIA PERJUANGAN”. On the time ranging from 2014 until 2023. We managed to obtain and use 27.000 datasets for our project.

**Data Collection**

The data used in this study is using a python library called SNScrape. SNScrape is a technique used to collect necessary data automatically from websites such as articles and social media platforms. SNScrape uses keywords to find the specific texts that are going to be used in our model. The keywords in this paper are “PDIP”, “PARTAI DEMOKRASI INDONESIA,” and ”PARTAI DEMOKRASI INDONESIA PERJUANGAN”. We used the keywords on a social media platform called Twitter to collect the data needed for our model.

**Text Processing**

After that, data that has been collected will be labelled and do some Text Processing. The first is Case Folding. In Case Folding, we will process all the letters in the dataset to lowercase. Then, for the second we will do Tokenizing. Tokenizing is the process of deleting all characters such as symbols while keeping the letters a to z and separate all the sentences into Indvidual words. The third is Stopword Removal, Stopword Removal will be used to delete unnecessary words that have no meaning like “yg”, “pen”, “nya”, “nih”, etc. Lastly, the fourth process is normalization. In this process, the informal words will be changed to formal words.

**TF-IDF Weighting**

TF-IDF process will note the weight of all the words in the dataset. This process is used to count all words appearing in the dataset. For the weighting of sentences into negative or positive sentiments, a collection of words with negative and positive connotations is needed. Therefore, we collect data from GitHub which contains a corpus for words with negative connotations and words with positive connotations. The weighting system, we calculate the total value of the occurrences of negative and positive words in each data in the dataset. Then the total from the negative weighting calculation results and the total from the positive weighting calculation results are compared to each other. If the total weight of the negative is greater than the positive, the data is determined to have a negative sentiment. Likewise, if the positive weight is greater than the negative, then the data is determined to have a positive sentiment. However, if there is a balance in the total weight between positive and negative, it will be determined to be a neutral sentiment.

**Dataset Splitting**

The dataset will be split into two parts, that is a testing set and training set. For the testing set, it will use 10% from the total dataset and for the training set will use 90% from total dataset.

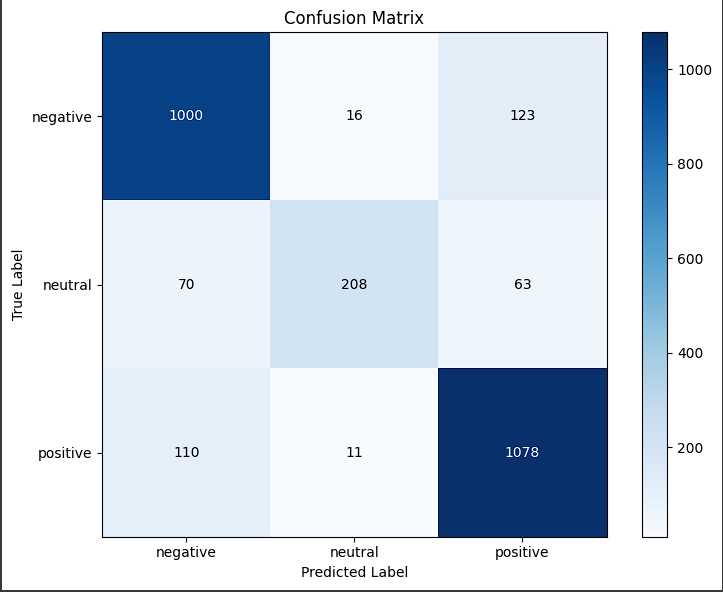
**Logistic Linear Regression**

The Logistic Regression Model that has been initiated will be fitted with training and testing data that has been splitted before. After that, model will predict testing data if that data is labeled with “Positive”, “Neutral”, or “Negative”

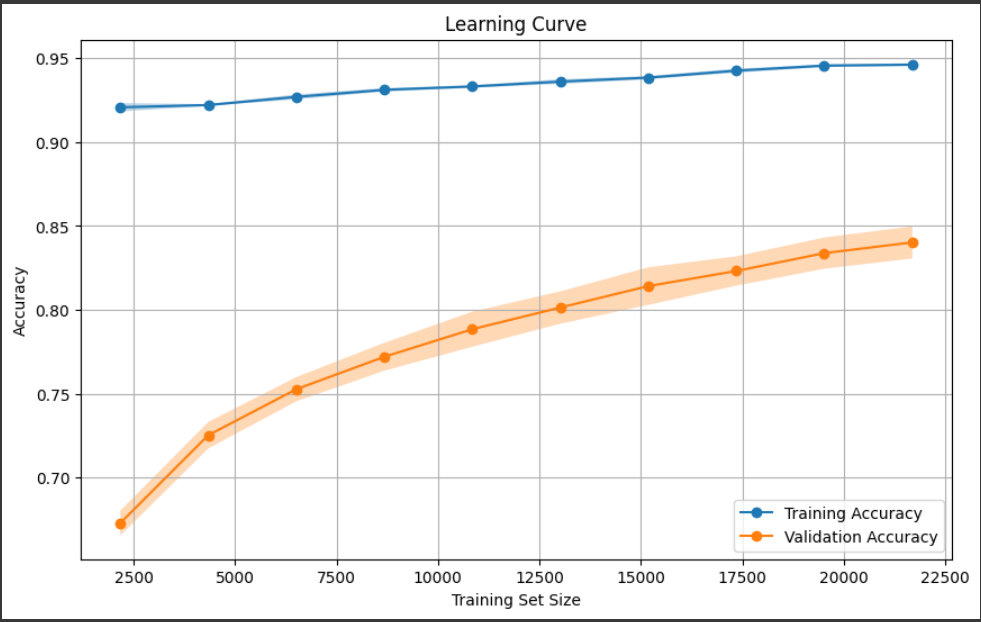
**BAB IV**

**Result**

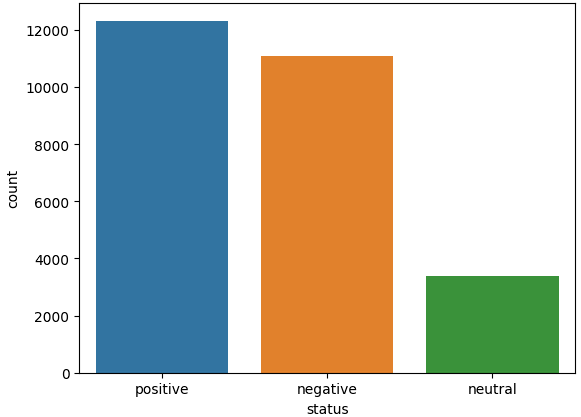
We have proposed the results of our Logistic Regression Model. With our model, we can achieve accuracy of 85%. The figure below provides our model’s performance. The confusion matrix will display our model’s prediction compared to real labels value and for the column will show predicted labels. And for the value within the cell show how many models predict that data.



Other than the confusion matrix, our model also determined the validation accuracy in which we can conclude that the size of the training dataset corresponds to the validation accuracy score. Which means, the smaller the training dataset’s size the lower the validation accuracy score, while the larger the training dataset’s size the higher the validation accuracy score becomes. The figure below will show you the graph that helped us determine this conclusion.



And in the end, we also got the objective of the model which is to classify each of the data that we’ve used to its corresponding classes, which are positive, negative, or neutral. And the classification is based on what we’ve discussed on the TF-IDF weighting stage. From the figure below we can conclude that it shows the cumulative result from the datasets that we’ve taken. There are roughly 12.000 datasets labelled “positive”, about 11.000 datasets labelled “negative”, and there are about 4.000 datasets labelled “neutral”.



**BAB V**

**Conclusion**

With our research, we found that the sentiment analysis can provides a value insight to some public opinion about president election. Through analyzing a lot of tweets from twitter we can have a deeper understanding of expression by users for our President Candidate especially from PDIP party.

The result from our model shows that, using Logistic Linear Regression algorithm we can achieve good accuracy in Sentiment Analysis problem. The Logistic Linear Regression algorithm can classify which data is a “Positive” statement, “Neutral” statement, or “Negative” statement for our dataset with good accuracy.

In conclusion, our research is to contribute to the growth of the Indonesian Politics and research of Sentiment Analysis. By using the Twitter data and creating logistic linear regression, we provide some insights to expression of Indonesian peoples during a presidential election. We also give better understanding of Indonesian opinion for political purpose. We hope this research will contribute to the development of Sentiment Analysis problems specifically for 2024 Indonesian Presidental Election to help Indonesian people to choose their President.

REFERENCE

[1] Mohd Ridzwan Yaakub Mohd, Muhammad Iqbal Abu Latiffi, Liyana & Safra Zaabar. “A Review on Sentiment Analysis Techniques and Applications.” IOP Conference Series: Materials Science and Engineering, vol. 551, no. 1, 1 Aug. 2019, p. 012070, DOI: 10.1088/1757-899x/551/1/012070.

LINK FILE PROJECT :

<https://drive.google.com/drive/folders/1fcl-epanOYrAzlhaC5oANSrdMa3vhByN?usp=sharing>

Untuk kodingannya, terdapat pada file ‘Experiment\_NLP.ipynb’

<https://drive.google.com/file/d/1jDGHeBcTwlpyWmfgn8ZaJVom9y9qMhlj/view?usp=sharing>

By :

* ANDREW NUGRAH – 2502031785
* DAVID AGUSTIAN WIJAYA RASA – 2540118353
* EDWARD EFFENLIE – 2540119671
* MATTHEW DIPAPUTRA WIDJAJA - 2502030006