**Development of Mini Application of Indonesian Abusive Words Filter on Twitter with Descriptive Methods.**

**Written by: Febrian Nur Alam**

**DSC WAVE-7**

The increase in social media users has caused an increase in communication activities between netizens in social media. For example, on twitter, citizens can communicate through tweets. Tweets on twitter can have a negative characteristic and this characteristic needs to have special attention because it is highly likely to contain hate speech. Mini Aps in its implementation is able to process data in the form of text or comma-separated values files.The Tweet data that has been taken then analyzed, how many tweets are toxic, what kind of distribution of hate speech is in it, and what category the hate speech belongs to. This article presents a review of several tweets that utilize machine learning to help detect text containing hate speech. It will then be processed by replacing "Alay" words and removing all abusive words. This results in standardized words.

1. Introduction

The development of technology that is increasingly advanced today makes it easier for humans to carry out various daily activities. Currently, the fastest growing technology is information and communication technology. One of them is the growth of online social media such as Twitter which is used to share facts or opinions of its users. According to the Big Indonesian Dictionary (KBBI), Facts are things that are true, while opinions are expressions that describe the feelings or emotions of the writer or speaker.

There are many daily active users of Twitter in Indonesia. According to Adriansah, Country Industry Director of Twitter Indonesia, the number of users in Indonesia increased by 17% to 145 million users in the third quarter of 2019.

With many Twitter users generating a variety of Tweets, these Tweets can be positive and some can be negative. There will be negative comments that are a problem because they usually contain elements of hate speech and can lead to legal sanctions for the person doing the threatening.

The Cybercrime Unit of the National Police Criminal Investigation Unit (Bareskrim Polri) has reprimanded 125 social media accounts in connection with content reported to contain hate speech. Of the accounts that received the warning, twitter accounts had the largest number, namely 79 accounts. This figure was recorded in the period February 23 to March 11, 2021.

The government in preventing and overcoming problems related to hate speech has issued legislation in the form of the ITE Law. Article 28 paragraph 2 of the ITE Law states that citizens are prohibited from disseminating information to cause hatred.

Part of machine learning, text analysis, has algorithms that can identify or classify text objects. This text analysis can be used to fight hate speech on social media as it can detect cyberbullying, offensive language, and cyber hate.

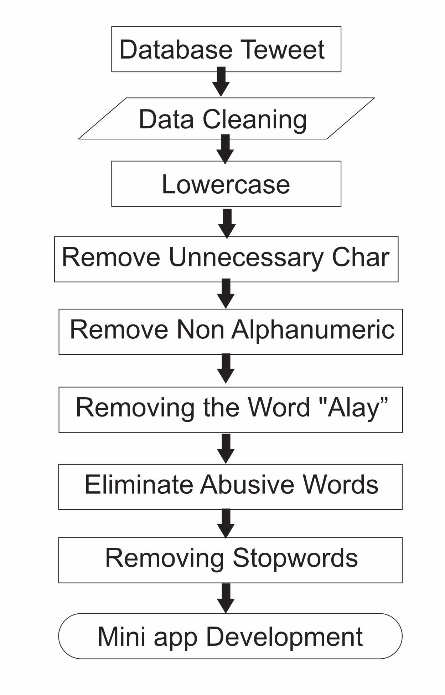
This research will create a system that can be used to classify hate speech, "Alay" words and identify non-standard language in a Tweet, then convert it into a standard language.

1. Dataset and Feature

This dataset was obtained from Kaggle with the title "Indonesian Abusive and Hate Speech Twitter Text". In this data set, there is a collection of Tweets that have been grouped based on several categories including. Hate speech data abbreviated as HS, Abusive, HS\_Individual, HS\_Group, HS\_Religion, HS\_Race, HS\_Physical, HS\_Gender, HS\_Other, HS\_Weak, HS\_Moderate and HS\_Strong.

* 1. Preprocessing Stages

The first step before a text data is processed further is to preprocess the text, the goal is to explore, process and organize the data. In text classification, text is grouped into several groups. The preprocessing stage is carried out to prepare the data so that the text data can be more structured by removing the noise in the data so that it is easier to process the data.

Lower case is used to convert the entire text in the document into a uniform form, the function of case folding is to make all letters in the document lowercase. Examples of the use of case folding can be seen in the table below.

Remove non-aplhabeth numeric is used to remove all characters that are not alphabetic characters.

Normalize "Alay" at this stage all the words "Alay" will be replaced with standard words.

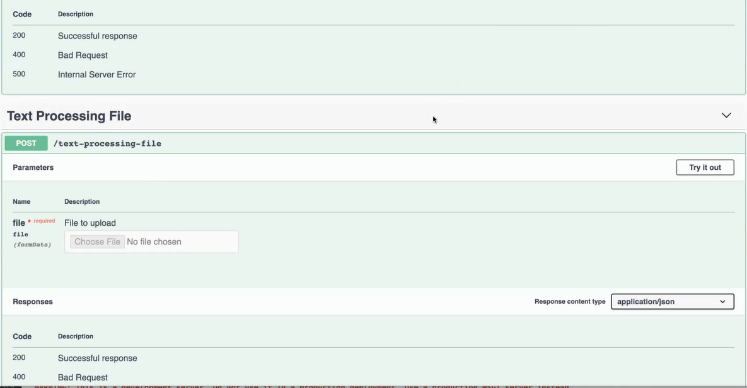
Ramove Abusive, in this process Tweets that contain rich abusive words will be removed so as to produce standardized words that do not contain abusive words.

Remove stopwrod, serves to reduce characters that have nothing to do with hate speech.

* 1. Data Classification

The data is grouped into several categories, HS and Abusive are grouped into Toxic data, then based on the HS category it is grouped into HS\_Individual, HS\_Group, HS\_Religion, HS\_Race, HS\_Physical, HS\_Gender and HS\_Other and finally Hate Speech data is grouped based on the level of existing HateSpeech. The data is grouped into HS\_Weak, HS\_Moderate and HS\_Strong.

* 1. The min-app development process

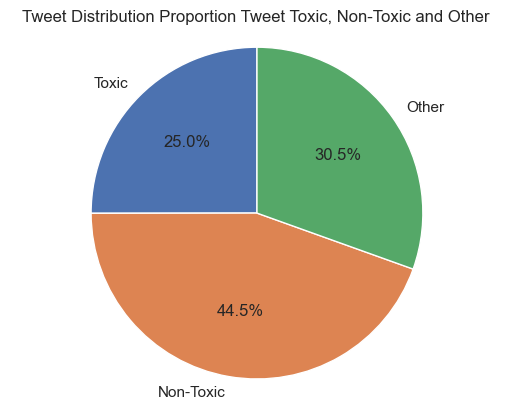


The Mini-Application was developed to make the tweet cleaning process easier. The following is a view of the mini-application developed using swagger. The incoming text is processed by the system that has been built. It will remove unnecessary characters, make all tweets lowercase, remove abusive words and replace "alay" words with standard words.

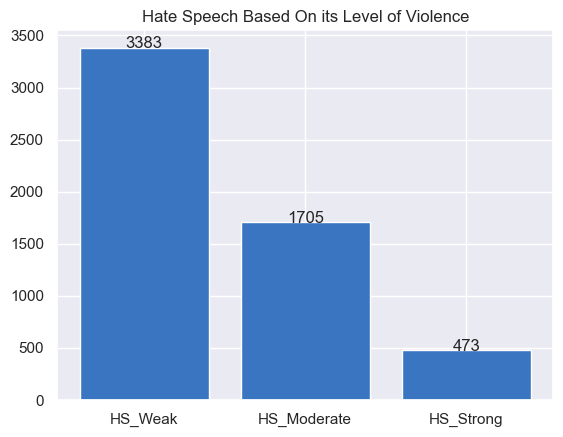
The data processing process begins with and must be done according to the analysis procedure. The extracted database is processed through the Preprocessing phase using a Python application to undergo a series of actions that include lowercase, removing unnecessary char, remove non-alphanumeric, removing the word “alay”, stopword, and mini apps development.

1. Results on 13.169 data that have been analyzed

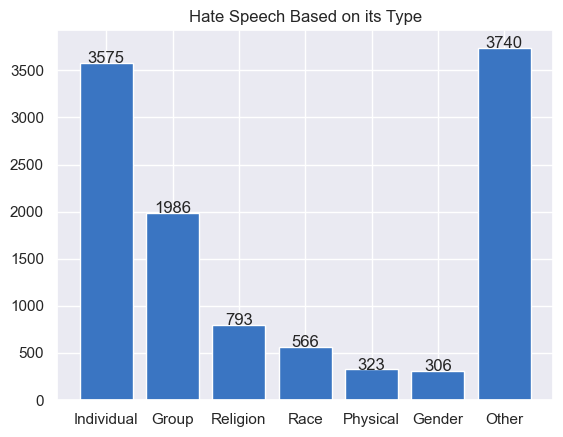
Based on the results of tweet sentiment analysis, shows that Non-Toxic tweets are the most tweets which are around 5,860 (44.5%) tweets, followed by other tweets which are tweets containing hate speech or abusive as many as 4,014 (30.5%) tweets and the least is toxic tweets 3,295 (25%) tweets from a total of 13,169 tweets.



Based on the level of violence, hate speech is divided into 3 types, namely weak hate speech (3383), Moderate hate speech (1705), and strong hate speech (473).



Based on its type, hate speech is divided into 7 types, namely individual hate speech (3575), group hate speech (1986), religious hate speech (793), racial hate speech (566), physical hate speech (323), gender hate speech (306), and other hate speech (3740).



With the amount of hate speech in this data, it is necessary to develop a mini-application that can be used to process harsh words into standardized words.

After the development of the mini-application, it is found that all the words entered into the application This application can be processed into standardized words and this mini application has 2 types of input, namely input in the form of text and input in the form of CSV-type files.



Conditions in the initial tweet data created using wordcloud, it can be seen that there are still many abusive words, there are still tweets that use capital letters. Tweets still contain many unnecessary characters such as numbers and symbols.



After the data was inputted into the mini-app, it became cleaner, the abusive words disappeared, all letters were changed to lowercase, and the word "alay" was changed to a standard word.

1. CONCLUSION

Based on the results of tweet sentiment analysis with descriptive methods, Non-toxic tweets are more dominant, namely as many as 5,860 (44.5%) tweets, followed by Toxic tweets as many as 3,295 (25%) other tweets as 4,014 (30.5%) from a total of 13,169 tweets. And the mini-app has succeeded in eliminating abusive words and replacing "Alay" words with standard words.

1. BIBLIOGRAPHY

A. P. J. Dwitama. Deteksi Ujaran Kebencian Pada Twitter Bahasa Indonesia. Menggunakan Machine Learning: Reviu Literatur. Jurnal SNATi. Volume 1. Nomor 1. 2021.

F. F. Mailo., L. Lazuardi. Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas Indonesia. Journal Of Information System For Public Health, Vol. 4,No. 1, April 2019.

O. H. Rahman., G. Abdillah., A. Komarudin. Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. Jurnal Rekayasa Sistem dan Teknologi Informasi. Vol. 5 No. 1 (2021) 17-23.

I. F. PUTRA., Indonesian Abusive and Hate Speech Twitter Text (2020) <https://www.kaggle .com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text/code>