Gold Level Challange

Tweet Analysis: Identifying Abusive and Hate Speech Language

Text Cleansing API Building and Data Analysis

Created by:

Hermawan

PRELIMINARY

Background

The current era of the internet allows the rapid and unfiltered spread of information, leading to the use of abusive language and hate speech to attack individuals or groups based on their race, religion, physical appearance, or sexual orientation. Such language is often found on social media and online games, especially MOBA games, where players may include school-age children or individuals with mental disorders.

As a junior data scientist in MOBA online games, I have decided to use Twitter data for descriptive analysis and pattern identification due to its vast amount of text data in the form of tweets. This analysis aims to provide evidence for the importance of text processing and cleaning abusive words in MOBA online games to create a healthy and non-toxic gaming environment. The results of the analysis will be used as a reference to develop an API that can process text, detect, and remove abusive words effectively.

Objectives

- Identifying tweet with abusive and hate speech words
- Analysis total words and total characters using univariate and bivariate EDA analysis
- Building an API Text Processing that can clean up rude words and replace slang words with normal words.

RESEARCH METHODS

DATA PREPARATION

```
#Cek kolom data tweet
df.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13169 entries, 0 to 13168
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Tweet           13169 non-null  object
 1   HS              13169 non-null  int64
 2   Abusive         13169 non-null  int64
 3   HS_Individual   13169 non-null  int64
 4   HS_Group        13169 non-null  int64
 5   HS_Religion     13169 non-null  int64
 6   HS_Race         13169 non-null  int64
 7   HS_Physical     13169 non-null  int64
 8   HS_Gender       13169 non-null  int64
 9   HS_Other        13169 non-null  int64
 10  HS_Weak         13169 non-null  int64
 11  HS_Moderate     13169 non-null  int64
 12  HS_Strong       13169 non-null  int64
dtypes: int64(12), object(1)
memory usage: 1.3+ MB
```

```
#Cek Duplicated Data Tweet
print('Masih ada {} duplicated data'.format(df.duplicated().sum()))
```

✓ 0.0s

```
Masih ada 125 duplicated data
```

```
#Membersihkan Duplicate Data Tweet
df = df.drop_duplicates()
print('Masih ada {} duplicated data'.format(df.duplicated().sum()))
print('Duplicated data sudah di hapus')
```

✓ 0.0s

```
Masih ada 0 duplicated data
Duplicated data sudah di hapus
```

```
#Cek Shape data hatespeech dari data uji
df.HS.value_counts()
```

✓ 0.0s

```
0    7526
1    5518
Name: HS, dtype: int64
```

```
HS_Other         3706
HS_Individual    3540
HS_Group         1978
HS_Religion       789
HS_Race           563
HS_Physical       322
HS_Gender         304
dtype: int64
```

```
#Cek Shape data abusive dari data uji
df.Abusive.value_counts()
```

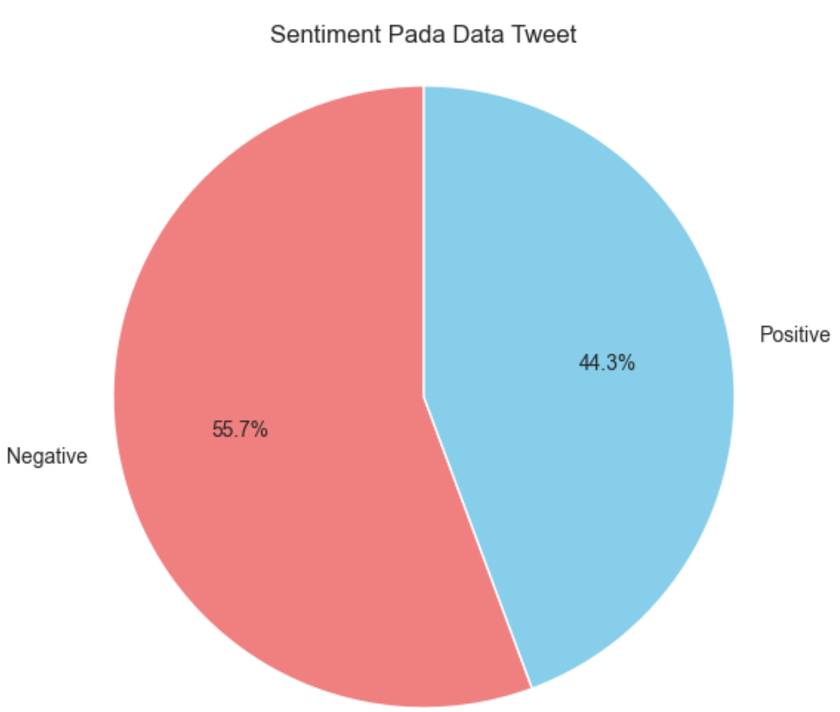
✓ 0.0s

```
0    8039
1    5005
Name: Abusive, dtype: int64
```

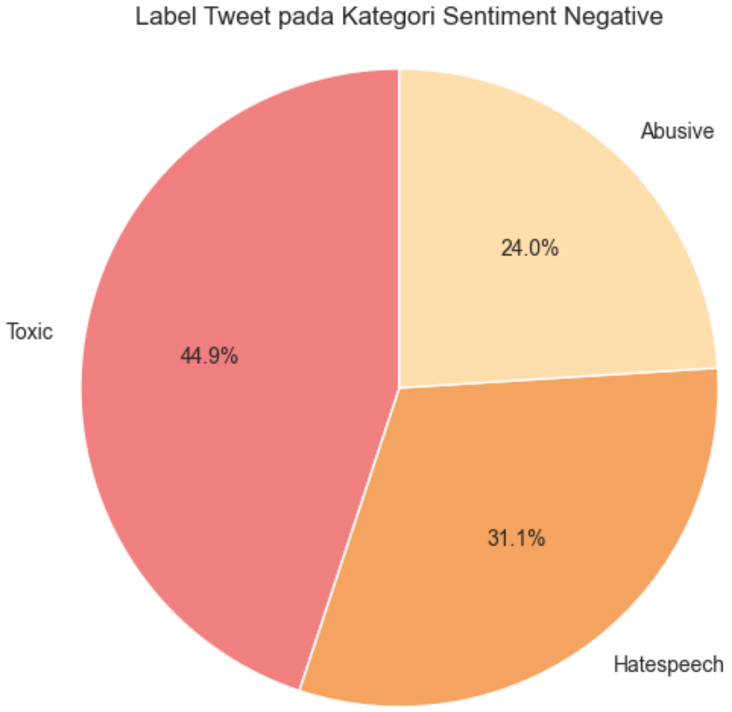
- The data used is secondary data obtained from Kaggle
- The data has 13 columns and 13169 rows, with no missing values.
- However, there were 125 duplicated data, so the number of data after removing duplicates was reduced to 13044 rows.
- In the data, there are 5518 tweets categorized as hates peech, with the most common types being HS_Other, HS_Individual, and HS_Group. Additionally, there are also 5005 data containing the word abusive.

RESEARCH METHODS

EXPLORATORY DATA ANALYSIS




```
Negative    7261
Positive    5783
Name: Sentiment, dtype: int64
```




```
Toxic          3262
Hatespeech     2256
Abusive        1743
Name: Label, dtype: int64
```

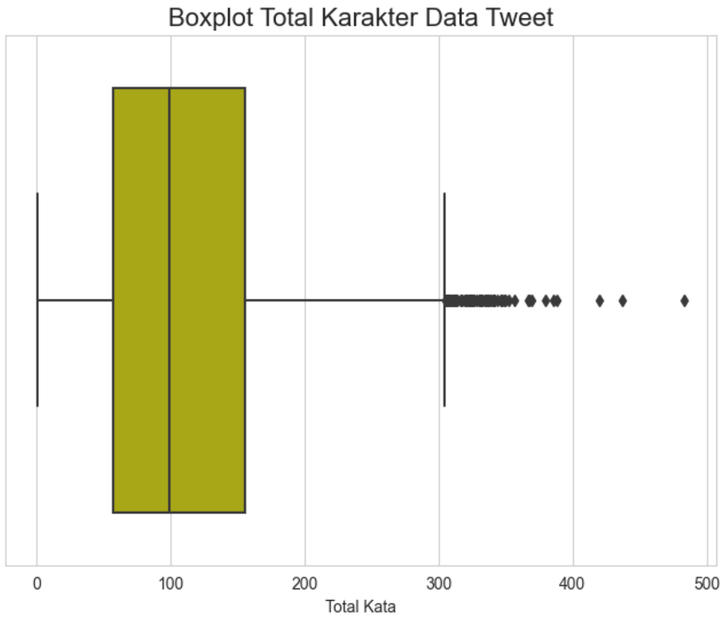
- By looking at the HS and Abusive collumn in the dataframe, a new column is created to determine the sentiment of the tweets. If HS and Abusive do not have values, the sentiment label is Positive. If either HS or Abusive (or both) have values, the sentiment label is Negative.

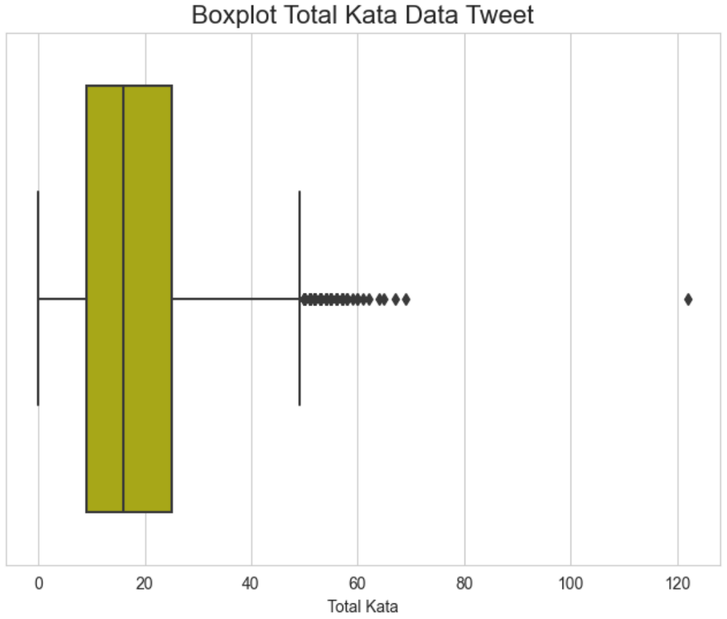
- From the dataset, it is found that there are more tweets with negative sentiment 7261(55,7%) than those with positive sentiment 5783(44,3%), with a difference of 11.4% of the total 13,044 data.

- From 7261 negative sentiment tweets, it can be concluded that tweets containing both abusive and hate speech words (Toxic) are 3262 (44.9%), while tweets containing only hate speech are 2256 (31.1%), and tweets containing only abusive words are 1743 (24.0%).

RESEARCH METHODS

EXPLORATORY DATA ANALYSIS






```
Quartile dan Interquartile Range Total Karakter

Batas Bawah 'total_char': -91.5
Nilai minimun: 1
Tidak ada outlier dari sisi batas bawah

Batas Atas 'total_char': 304.5
Nilai maksimum: 483
Ada outlier dari sisi batas atas
```

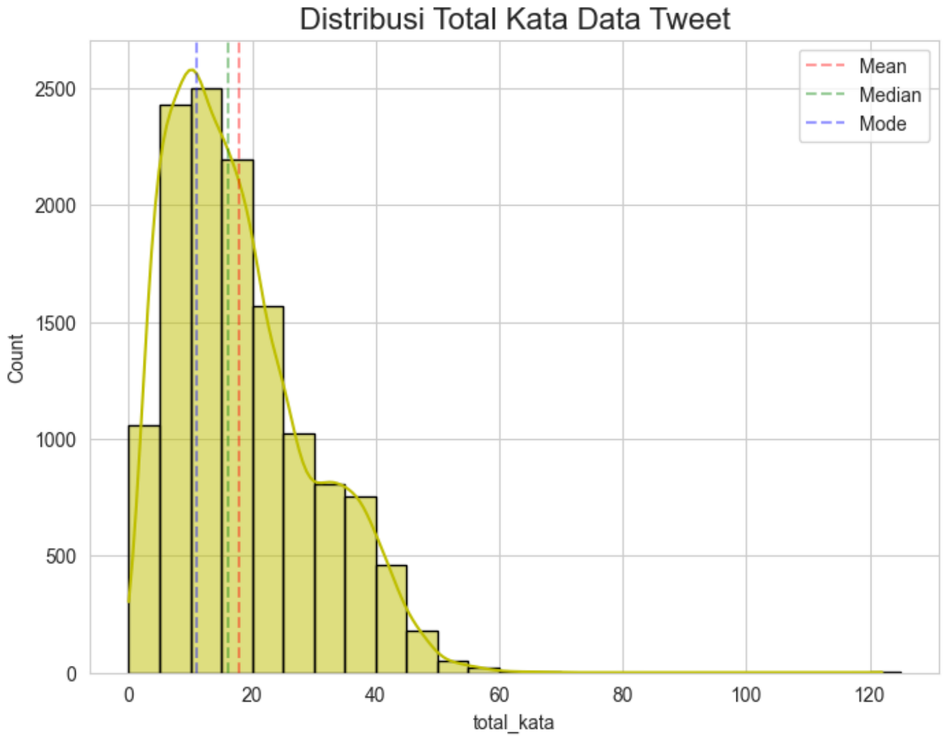
```
Quartile dan Interquartile Range Total Kata

Batas Bawah 'total_kata': -15.0
Nilai minimum: 0
Terdapat outlier dari sisi batas bawah

Batas Atas 'total_kata': 49.0
Nilai maksimum: 122
Tidak terdapat outlier dari sisi batas atas
```

- The total number of characters has a quartile and interquartile range. The lower boundary of 'total_char' is -91.5, with a minimum value of 1, and no outliers on the lower side. On the other hand, the upper boundary of 'total_char' is 304.5, with a maximum value of 483, and outliers exist on the upper side.

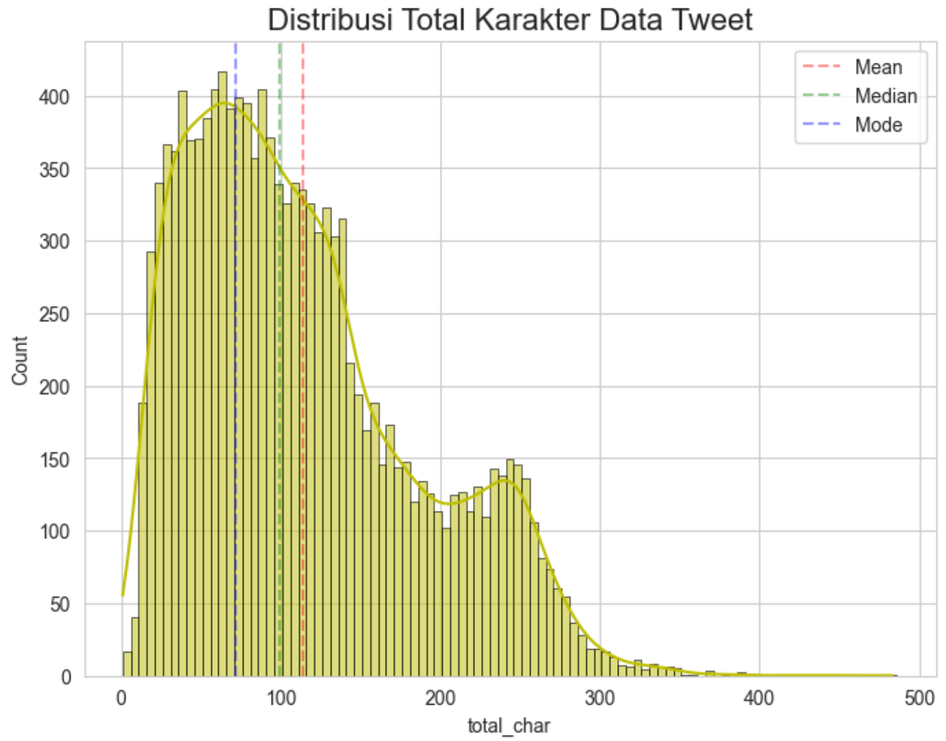
- The Interquartile Range for total words is 49, with a minimum value of 0 and a maximum value of 122. However, there are outliers on the lower end with a lower boundary of -15, while no outliers were found on the upper end

RESEARCH METHODS

EXPLORATORY DATA ANALYSIS



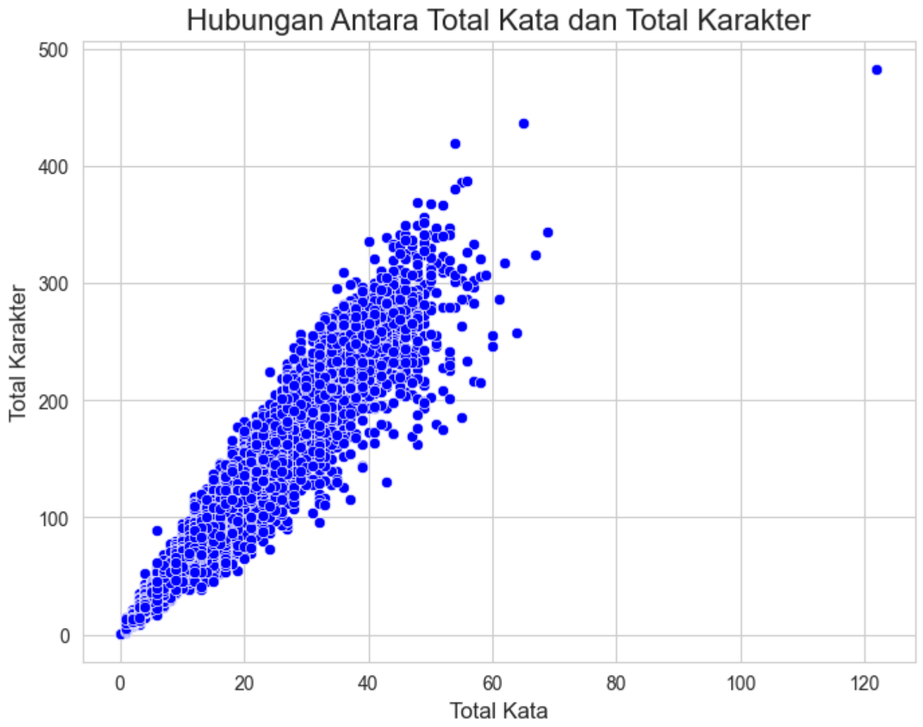

```
EDA Total Kata Pada Tweet
Mean: 17.95
Median: 16.00
Mode: 11.00
Variance: 129.70
Standard Deviation: 11.39
Skewness: 0.85
Kurtosis: 0.51
```

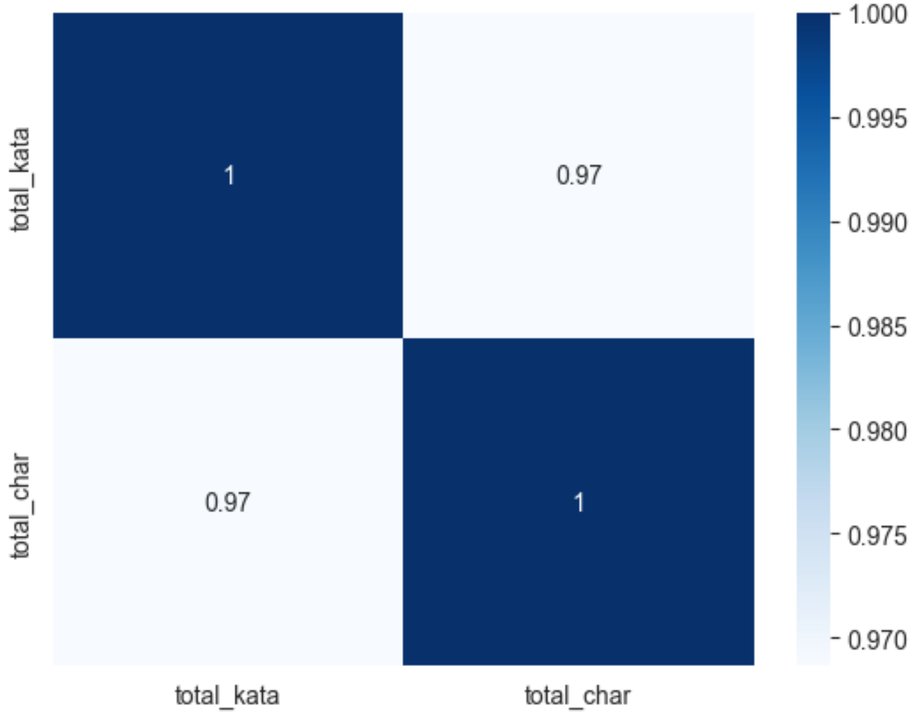



```
EDA Total Karakter Pada Tweet
Mean: 113.86
Median: 99.00
Mode: 72.00
Variance: 5174.96
Standard Deviation: 71.94
Skewness: 0.75
Kurtosis: -0.15
```

RESEARCH METHODS

EXPLORATORY DATA ANALYSIS





there was a strong positive linear correlation between total_kata and total_char, with a correlation coefficient of 0.97

RESEARCH METHODS

TEXT NORMALIZATION

Original Data

Data.csv

Text Processing

Text cleansing using RegEx library

Remove Abusive words with help of abusive.csv

Replace non standard words with the help of kamusalay.csv

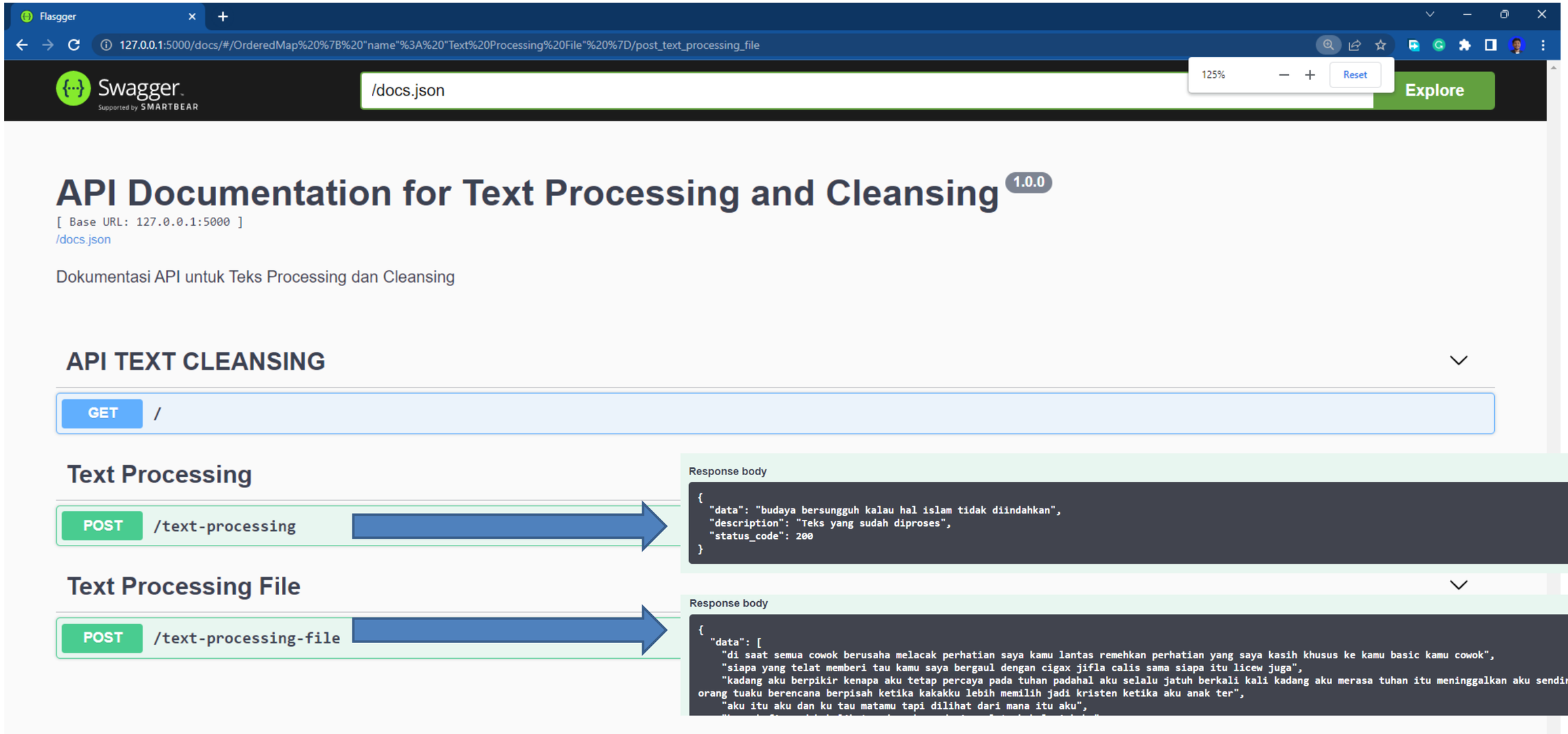
Clean Data

TWEET COMPARISON

```
Tweet lama: USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT TAPI DILIAT DARI MANA ITU AKU'

Tweet baru: aku itu aku dan ku tau matamu tapi dilihat dari mana itu aku
```

RESEARCH RESULT

API PROCESSING TEXT CLEANSING



RESEARCH RESULT

RESULT

- The descriptive analysis reveals that 56.7% of the tweets in the dataset have negative sentiment, which is higher than the 44.3% of tweets with positive sentiment. This implies a significant difference of 11.4% in the total of 13,044 data. Further analysis on the negative sentiment tweets shows that 44.9% of them contain both abusive and hate speech words (Toxic), while 31.1% contain only hate speech and 24.0% contain only abusive words.

- In conducting exploratory data analysis (EDA) using univariate and bivariate analysis, the dataset was analyzed for two variables: total_char (total characters) and total_kata (total words). The mean value for total_char was 113.86 with a median of 99 and a mode of 72. The range for total_char was 482, and there were no outliers on the lower end of the quartile and interquartile range. However, there were outliers on the upper end, with a maximum value of 483. The variance for total_char was 5175.35, which was greater than the mean, while the standard deviation was 71.94, which was smaller than the mean. The skewness for total_char was 0.75, which indicated a positive skew, and the kurtosis was -0.15, indicating a platykurtic distribution.

- For total_kata, the mean was 17.95, the median was 16, and the mode was 11. The range for total_kata was 122, and there were outliers on the lower end of the quartile and interquartile range. However, there were no outliers on the upper end. The variance for total_kata was 129.71, which was greater than the mean, while the standard deviation was 11.39, which was smaller than the mean. The skewness for total_kata was 0.85, which indicated a positive skew, and the kurtosis was 0.51, indicating a platykurtic distribution.

- Furthermore, there was a strong positive linear correlation between total_kata and total_char, with a correlation coefficient of 0.97.

- The API created has 2 endpoints for each model (to process text and file data) and can clean up punctuation, rude words and replace slang words with normal words.

RESEARCH RESULT

CONCLUTION AND RECOMMENDATION

- The analysis of Twitter data with analisys descriptive has revealed a high prevalence of negative sentiment and abusive/hate speech in online communication. This highlights the need for effective text processing and filtering tools to create a healthy and non-toxic gaming environment.

- The EDA analysis with bivariate and univariate showed that the length of tweets varies widely, with most tweets containing around 100 characters and 16 words. However, there are also outliers with very long tweets. Additionally, there was a strong positive correlation between total_char and total_kata, indicating that longer tweets also tend to have more words.

- Finally, the developed API provides a useful tool for processing text data, detecting and removing abusive language, and replacing slang words with standard words, which can be implemented in MOBA online games to improve the quality of communication and reduce toxicity.

THANK YOU

CONTACTS :

[Hermawan]

[GitHub Repository]