Analisis Text dalam Tweet menggunakan Descriptive Analysis

Brain Dior (Binar Academy)

Pendahuluan

Analisis data teks telah menjadi salah satu bidang yang berkembang pesat dalam ilmu data dan analitik. Dalam konteks analisis media sosial, seperti Twitter, data teks yang dihasilkan oleh pengguna menjadi sumber informasi yang berharga. Descriptive analysis atau analisis deskriptif merupakan salah satu metode yang digunakan dalam menggali pemahaman dari data teks tersebut.

Dengan memanfaatkan analisis deskriptif pada data teks dari tweet, kita dapat menggali pemahaman yang lebih dalam tentang pola, tren, dan makna di balik teks tersebut. Analisis ini dapat membantu dalam pengambilan keputusan, pemetaan opini publik, atau identifikasi isu-isu penting yang berkembang di platform media sosial seperti Twitter.

Analisis deskriptif penelitian ini bertujuan untuk memberikan gambaran umum tentang karakteristik, pola, dan makna yang terkandung dalam teks. Dalam konteks analisis data teks dari tweet, analisis deskriptif dapat digunakan untuk mengidentifikasi kata kunci yang sering muncul, tema yang dominan, sentimen yang terkandung, serta hubungan antara kata-kata atau topik tertentu.

Metode Penelitian

Data pada penelitian ini bersumber dari kaggle oleh okkyibrohim yang sudah dipublikasikan dalam paper berikut untuk undergraduate project nya. Lebih spesifik lagi, ini adalah dataset untuk deteksi ujaran kebencian dan bahasa kasar dengan label ganda dalam Twitter Indonesia.

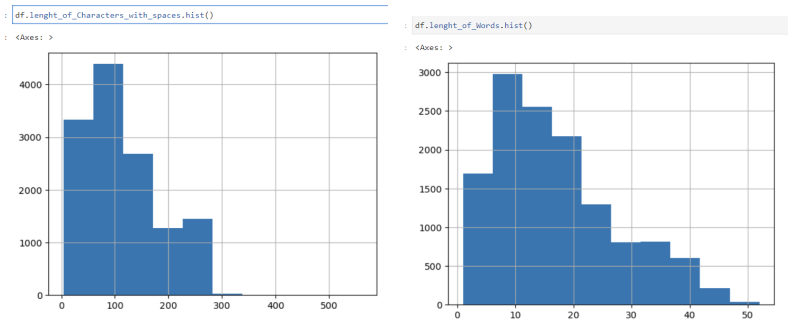
Metode analisis yang dipakai dalam penelitian ini menggunakan Descriptive Analytics. Karena bertujuan mendeskripsikan pola dari data. Jenis analisis tersebut dirasa cocok karena fokus pada mencari tahu kondisi data dan mempelajari pola suatu data.

Analisisnya diproses dengan berdasarkan kolom yang diproses yakni variabel (Univariate Analysis) dan variabel (Bivariate Analysis). Dalam setiap prosesnya menerapkan metode Descriptive Statistic dan Visualisasi. Descriptive Statistic digunakan untuk mencari tahu persebaran data secara angka sedangkan visualisasi untuk mencari tahu persebaran data secara visual.

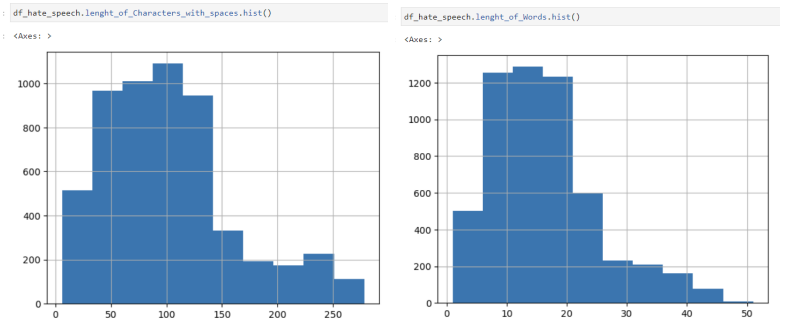
Hasil dan Kesimpulan

Berdasarkan analisis yang sudah kita lakukan dapat hasilnya dapat dijabarkan sebagai berikut :

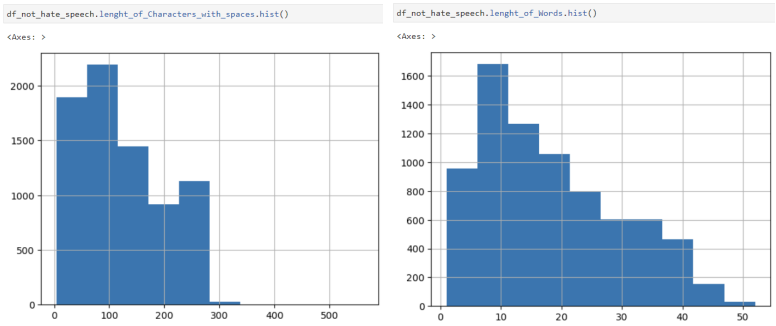
- Berdasarkan Univariate Analysis :
    - Dalam Descriptive Statistic menunjukkan data yang kita olah memiliki outlier namun tidak terlalu signifikan
    - Dalam visualisasi menunjukkan:
        - **Length of tweet dan total words** memiliki panjang 100-200 karakter dan 10-20 kata.



        - **Length of tweet dan total words** yang **termasuk** hate speech (HS) yakni 100-150 karakter dan 15-25 kata.



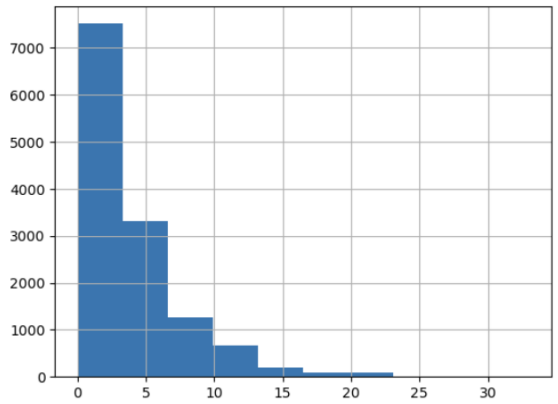
        - **Length of tweet dan total words** yang **tidak** termasuk hate speech (**not HS**) yakni 100-200 karakter dan 10-20 kata.



- Informal words mentioned memliki rentang 2-7 kali di mention

```
df.alay_words_mentioned.hist()
```

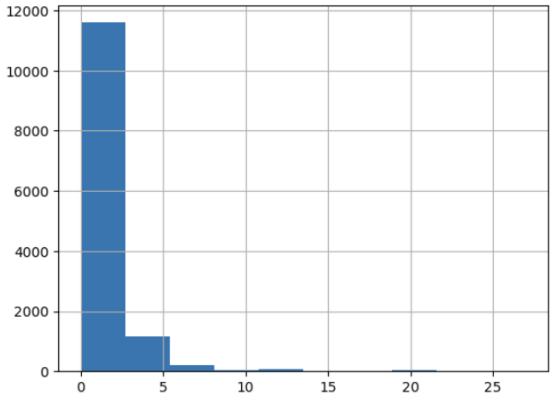
```
<Axes: >
```



- User mentioned memliki rentang 2-7 kali di mention

```
df.user_mentioned.hist()
```

```
<Axes: >
```



- Pada sentimen **positif**, kata yang sering muncul adalah "USER", "di", "dan", "yg", "yang", dan "itu".

```
text = ' '.join(df_not_hate_speech['Tweet'])
wordcloud = WordCloud().generate(text)

# Generate Plot
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

- Pada sentimen **negatif**, kata yang sering muncul adalah "USER", "URL", "dan", "yg", "yang", dan "Indonesia".

```
text = ' '.join(df_hate_speech['Tweet'])
wordcloud = WordCloud().generate(text)

# Generate Plot
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

- Pada saat **user mentioned lebih dari 5 kali dalam satu tweet**, kata yang sering muncul adalah "USER", "cebong", "dan", "yg", "yang", dan "Presiden".

```
text = ' '.join(df_user_mentioned['Tweet'])
wordcloud = WordCloud().generate(text)

# Generate Plot
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

- Pada saat **informal word mentioned lebih dari 5 kali dalam satu tweet**,, kata yang sering muncul adalah "USER", "gue", "di", "yg", "ada", dan "dgn".

```
text = ' '.join(df_alay_mentioned['Tweet'])
wordcloud = WordCloud().generate(text)

# Generate Plot
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

- Berdasarkan Bivariate Analysis:
    - Dalam Descriptive Statistic menunjukkan variabel **User Mentioned** dan **Informal Word Mentioned** memiliki korelasi positif.

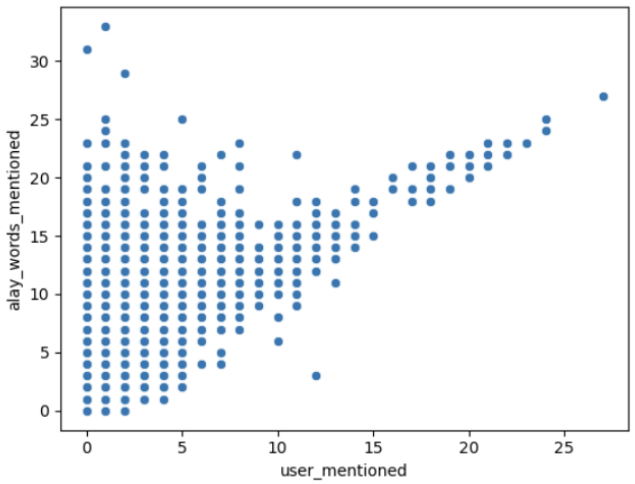
```
]: df[['user_mentioned','alay_words_mentioned']].corr()
```

]:

	user_mentioned	alay_words_mentioned
user_mentioned	1.000000	0.584776
alay_words_mentioned	0.584776	1.000000

    - Dalam visualisasi menunjukkan:
        - Variabel **User Mentioned** dan **Informal Word Mentioned** terkonfirmasi memiliki korelasi positif.

```
: sns.scatterplot(x=df['user_mentioned'], y=df['alay_words_mentioned'])
: <Axes: xlabel='user_mentioned', ylabel='alay_words_mentioned'>
```



Dari hasil di atas dapat disimpulkan user mentioned dan informal word mentioned memiliki korelasi positif. Pada saat user mentioned lebih dari 5 kali dalam satu tweet, kata yang sering muncul adalah "USER", "cebong", "dan", "yg", "yang", dan "Presiden". Dari kata yang sering muncul tersebut dapat diinterpretasikan ada unsur hate speech (HS) dari tweet tersebut.