Analisis Karakter dan Kata pada Tweet Hate Speech & Abusive dalam Bahasa Indonesia menggunakan Descriptive Analytics

Damar Pradiptojati
Binar Academy

Pendahuluan

Twitter merupakan salah satu media sosial yang populer di dunia dengan total pengguna sejumlah 396,5 pengguna. Indonesia pada tahun 2022 menduduki peringkat kelima sebagai negara dengan jumlah pengguna terbanyak, yaitu 18,5 juta pengguna menurut Statista.

Dalam dinamika penggunaan Twitter sebagai media komunikasi sosial, Indonesia tidak terlepas dari penyalahgunaan tweet, salah satunya adalah dengan melontarkan tweet yang bersifat menghina dan ujaran kebencian. Oleh karena itu, sebuah penelitian perlu dilakukan untuk mengamati pola dari berbagai jenis penyalahgunaan tweet tersebut.

Penelitian ini dilakukan untuk menganalisis panjang kalimat, jumlah kata, maupun pilihan kata dari tweet yang dilontarkan oleh pengguna Twitter dalam Bahasa Indonesia. Harapannya dari hasil analisis yang dilakukan menjadi bahan pertimbangan berbagai pihak kedepannya.

https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Metode Penelitian

Data pada penelitian ini bersumber dari website Kaggle dengan judul "*Indonesian Abusive and Hate Speech Twitter Text*". Data berikut adalah data yang memuat kumpulan komentar dalam bahasa indonesia diperoleh dari Twitter dan telah diklasifikasi dalam beberapa label antara lain *Hate Speech* dan *Abusive*, serta berbagai varian dari *Hate Speech*.

Metode analisis yang dipakai dalam penelitian ini menggunakan *Descriptive Analytics*. Jenis analisis tersebut dipergunakan karena memiliki fokus dalam mencari tahu kondisi data dan mendeskripsikan pola suatu data.

Analisis diproses berdasarkan kolom yang diproses yakni 1 variabel (*Univariate Analysis*) dan 2 variabel (*Bivariate Analysis*). Dalam setiap prosesnya diterapkan metode *Descriptive Statistic* dan Visualisasi. *Descriptive Statistic* digunakan untuk mencari tahu persebaran data secara angka sedangkan visualisasi digunakan untuk mencari tahu persebaran data secara visual.

Hasil dan Kesimpulan

Berdasarkan analisis yang sudah dilakukan terdapat temuan sebagai berikut:

- Berdasarkan Univariate Analysis:

  Melalui Descriptive Statistic:

  data yang diolah memiliki sejumlah outlier dengan jumlah yang tidak signifikan

  Melalui Visualisasi:

  Total karakter dan total kata memiliki panjang 1-293 karakter dan 1-85 kata.

  Tweet berlabel non *hate speech* lebih banyak daripada yang berlabel *hate speech*. Demikian pula tweet berlabel non *abusive* lebih banyak daripada yang berlabel *abusive*.

Hasil dan Kesimpulan

Tweet berlabel non *hate speech* rata-rata cenderung memiliki lebih banyak karakter dan kata (108 karakter & 17 kata) daripada yang berlabel *hate speech* (87 karakter & 14 kata).

Demikian pula tweet berlabel non *abusive* rata-rata cenderung memiliki lebih banyak karakter dan kata (112 karakter & 18 kata) daripada yang berlabel *abusive* (78 karakter & 13 kata).

Hasil dan Kesimpulan

Pada tweet *non hate speech* , kata yang sering muncul adalah:

- dan, di, yg, yang, itu, ini, ada, aku, orang, presiden

Sedangkan pada tweet *hate speech* , kata yang sering muncul adalah:

- yg, di dan, itu, jokowi, ini, cebong, yang, ya, aja

Pada tweet *non abusive* , kata yang sering muncul adalah:

- dan, di, yg, yang, itu, ini, ada, jokowi, presiden, orang

Sedangkan pada tweet *abusive* , kata yang sering muncul adalah:

- yg, di, itu, dan, cebong, ini, yang, lu, ya, aja

Hasil dan Kesimpulan

- Berdasarkan Bivariate Analysis:

  Melalui Descriptive Statistic:

  Total karakter dan total kata memiliki korelasi positif yang kuat.

  Melalui visualisasi:

  Total karakter dan total kata menunjukkan suatu korelasi positif linier.

  Total karakter dan kata pada tweet *non hate speech* berjumlah lebih banyak daripada *hate speech*.

  Total karakter dan kata pada tweet *non abusive* berjumlah lebih banyak daripada *abusive.*

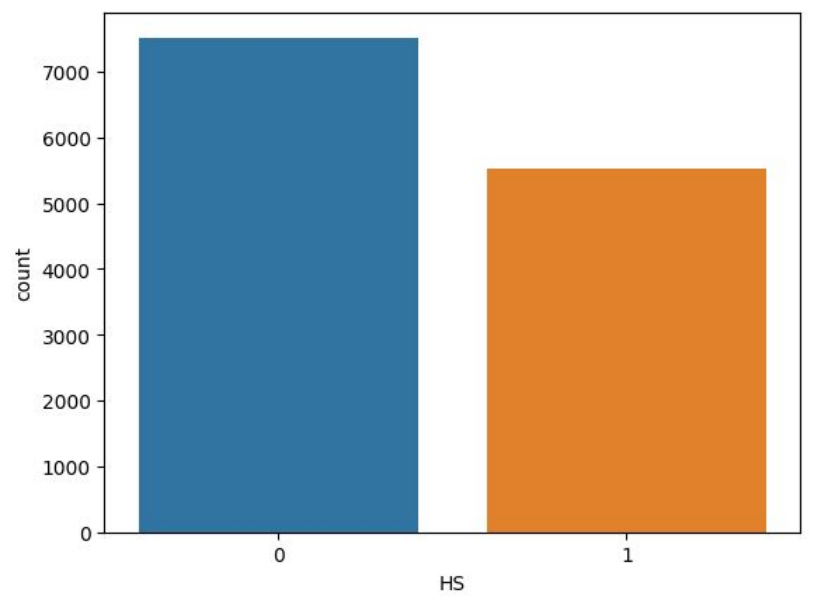
Hasil dan Kesimpulan

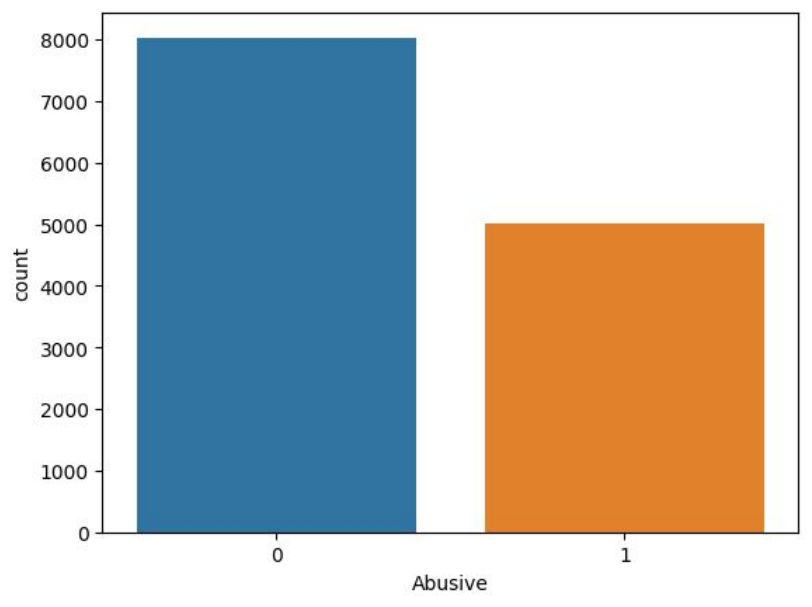
Setelah melalui proses penerjemahan dan penghapusan stopword, ditemukan bahwa

- 10 Kata Hate speech terbanyak:
    - kamu, indonesia, presiden, jokowi, kalau, orang, cebong, jadi, komunis
- 10 Kata Abusive terbanyak:
    - kamu, gue, kalau, cebong, orang, sama, sih, apa, jadi, mau

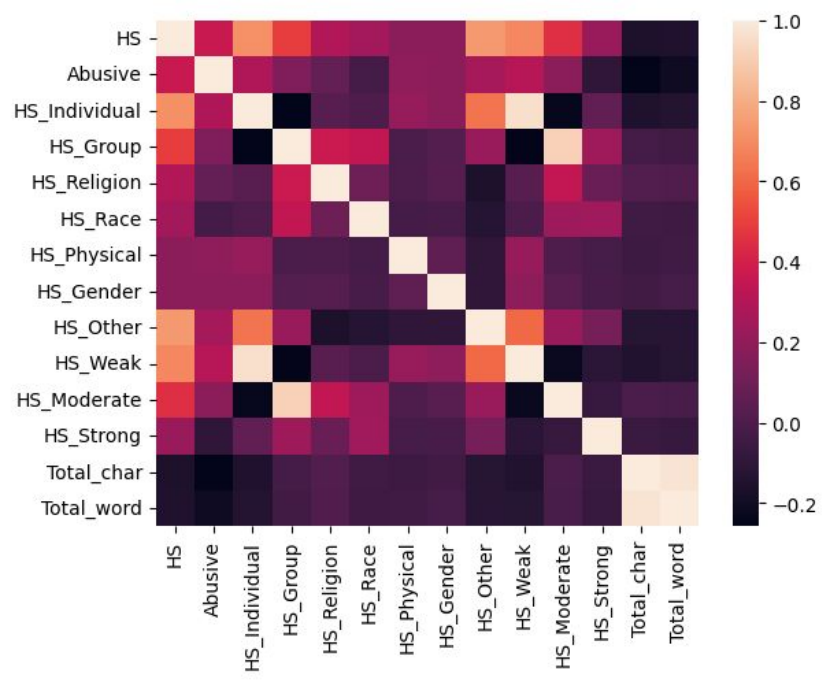
Berdasarkan data di atas, dapat disimpulkan bahwa ujaran kebencian cenderung tertuju pada situasi politik di Indonesia, didukung dengan banyaknya isu buzzer yang menggunakan media sosial untuk melakukan provokasi. Sedangkan ujaran yang bersifat penghinaan memiliki kecenderungan ke arah personal.

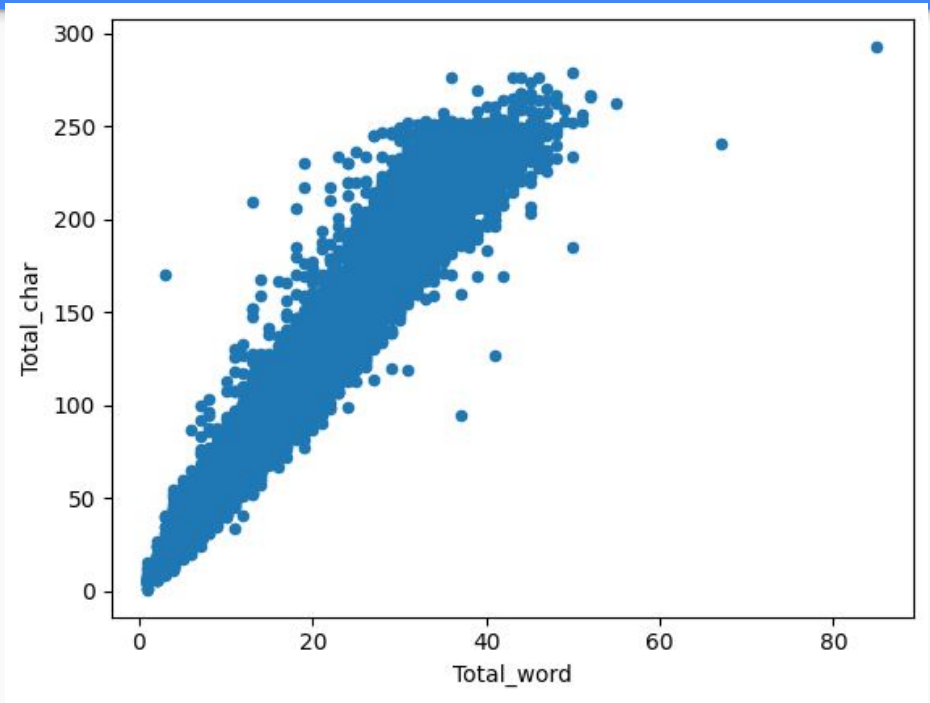
Appendix
Perbandingan Jumlah Positive & Negative Hate Speech & Abusive



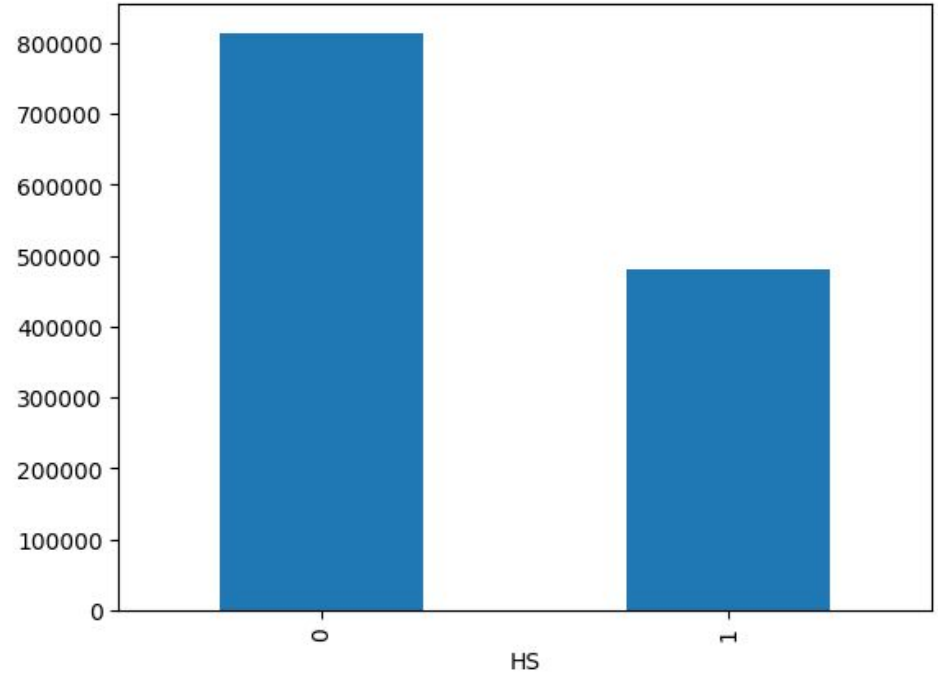


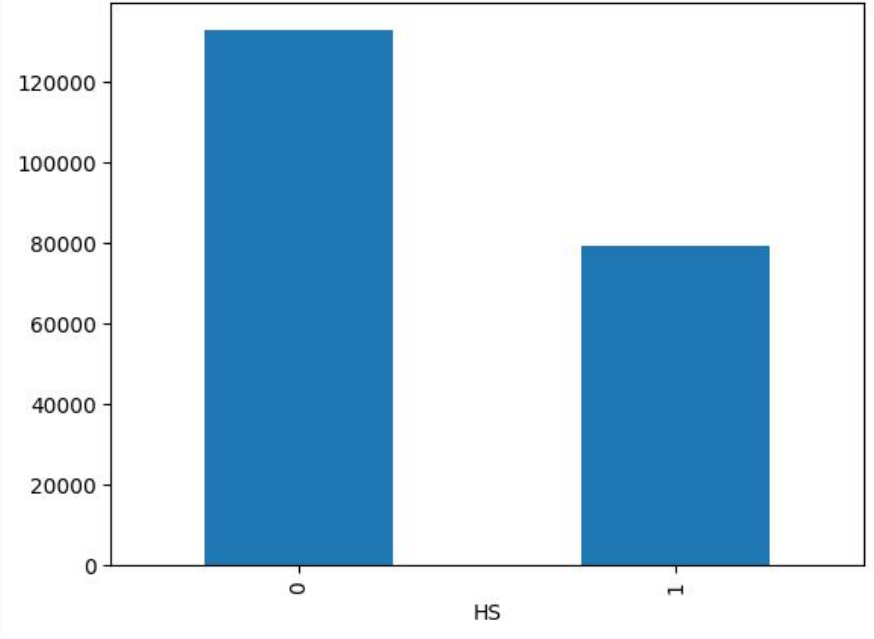
Appendix
Heat map korelasi antar variabel, serta hubungan korelasi positif linier antara total char dan total word





Appendix
Perbandingan jumlah total char dan total word pada variabel hate speech





Appendix
Perbandingan jumlah total char dan total word pada variabel abusive