**PERBANDINGAN KINERJA ALGORITMA KNN DAN LOGISTIC REGRESSION DALAM MENDETEKSI STRESS PADA TEKS**

**Hafizh Raihan Kurnia Putra1, Ja’far Shidqul Azzam2 , Muhammad Rusydi Hanan3**

1, 2, 3Universitas Brawijaya, Malang

Email: [1hachiman170@student.ub.ac.id](mailto:1hachiman170@student.ub.ac.id), 2jafarshidqul@student.ub.ac.id, 3rusdyhanan@student.ub.ac.id

\*Penulis Korespondensi

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

**Abstrak**

Dalam kehidupan seseorang tidak semua berjalan sesuai dengan apa yang mereka harapkan. Terjadinya peristiwa yang tak diinginkan menjadikan seseorang mengalami tekanan yang berlebihan sehingga menjadi stres. Stres adalah hubungan eksklusif antara orang dan lingkungan di sekitarnya yang jauh melampaui sumber daya kopingnya dan mengancam kesehatannya. stres bisa diobati secara mandiri dengan melakukan manajemen stress. Manajemen stress dilakukan dengan mencari penyebab hal-hal yang membuat stress kemudian mencari solusi untuk permasalahannya. Untuk mendeteksi keberadaan stres dan mengetahui penyebab stres tersebut pada seseorang, kita bisa menganalisis dari postingan orang tersebut pada sosial media. Contoh metode klasifikasi teks adalah *Logistic Regression* dan *K-Nearest Neighbor* (KNN). KNN adalah algoritma sederhana namun efektif yang digunakan untuk tugas klasifikasi namun memiliki kelemahan dalam klasifikasi data yang berukuran besar *Logistic Regression* adalah algoritma pembelajaran mesin yang efektif digunakan untuk melakukan klasifikasi pada data dengan dua kelas dan bisa menangani data yang berukuran besar. Penelitian ini juga dilakukan tuning untuk mencari *hyperparameter* terbaik untuk masing-masing model dimana didapatkan *Logistic Regression* dengan hyperparameter C bernilai 1, iterasi maksimal 100, dan tol 0.001 dan KNN dengan *hyperparameter* K bernilai 1, Leaf Size bernilai 15 dan P bernilai 2. Performa metode yang dibentuk masing-masing yaitu *Logistic Regression* dalam melakukan klasifikasi stres memiliki akurasi 0.91, dan KNN yaitu 0.54

**Kata kunci**: *deteksi stres, teks, KNN, logistic regression*

***A COMPARISON OF THE PERFORMANCE OF THE K-NEAREST NEIGHBORS (KNN) AND LOGISTIC REGRESSION ALGORITHM IN DETECTING STRESS IN TEXT***

***Abstract***

*In a person's life, not everything goes according to what they expect. The occurrence of unwanted events makes a person experience excessive pressure so that it becomes stressful. Stress is an exclusive relationship between people and the environment around them that far exceeds their coping resources and threatens their health. stress can be treated independently by doing stress management. Stress management is done by finding the causes of stressful things and then finding solutions to the problem. To detect the presence of stress and find out the cause of stress in a person, we can analyze the person's posts on social media. Examples of text classification methods are Logistic Regression and K-Nearest Neighbor (KNN). KNN is a simple but effective algorithm used for classification tasks but has weaknesses in the classification of large data. Logistic Regression is an effective machine learning algorithm used to classify data with two classes and can handle large data. This research also conducted tuning to find the best hyperparameters for each model where it was obtained Logistic Regression with hyperparameter C worth 1, maximum iteration 100, and toll 0.001 and KNN with hyperparameter K worth 1, Leaf Size worth 15 and P worth 2. The performance of the methods formed respectively, namely Logistic Regression in classifying stress has an accuracy of 0.91, and KNN is 0.54.*

**Keywords**: *stress detection, text, KNN, logistic regression*

# PENDAHULUAN

Dalam kehidupan seseorang tidak semua berjalan sesuai dengan apa yang mereka harapkan. Banyak sekali peristiwa-peristiwa yang tidak mereka inginkan terjadi dalam kegiatan sehari-hari mereka. Terjadinya peristiwa yang tak diinginkan yang terjadi membuat seseorang menjadi merasa gagal yang kemudian apabila terakumulasi bisa menjadikan seseorang mengalami tekanan yang berlebihan sehingga menjadi stres.

Stress adalah hubungan eksklusif antara orang dan lingkungan di sekitarnya yang jauh melampaui sumber daya kopingnya dan mengancam kesehatannya (Lazarus & Folkman, 1984). Seperti definisi yang sudah diterangkan, stress disebabkan oleh situasi sulit yang dialami seseorang, mulai dari kehilangan hal yang berharga, tertimpa sebuah musibah, atau mengalami permasalahan yang tidak bisa diatasi. Secara fisik, stres dapat menyebabkan kurangnya energi dari tubuh secara persisten, kurangnya nafsu makan, sakit kepala, dan lambung (Misra & Castillo, 2004).

Stress bisa diobati secara mandiri dengan melakukan manajemen stress. Manajemen stress dilakukan dengan mencari penyebab hal-hal yang membuat stress kemudian mencari solusi untuk permasalahannya. Oleh karena itu, mengetahui penyebab-penyebab stress kebanyakan orang perlu dilakukan untuk membantu orang-orang yang stress mengetahui penyebab stres mereka dan menyelesaikan permasalahan tersebut.

Untuk mendeteksi keberadaan stres dan mengetahui penyebab stres tersebut pada seseorang, kita bisa menganalisis dari postingan orang tersebut pada sosial media. Contoh metode klasifikasi/deteksi teks adalah *Logistic Regression* dan *K-Nearest Neighbor (KNN)*.

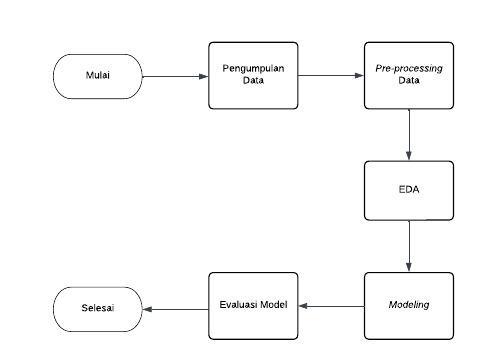
Pada penelitian yang dilakukan oleh Imron dkk pada tahun 2019, dilakukan deteksi stres berdasarkan teks postingan pada Twitter menggunakan algoritma *Improved KNN* untuk klasifikasi dan *Chi-square* untuk seleksi fitur. Pada penelitian ini, metode tersebut mendapatkan hasil rata-rata akurasi sebesar 83.3%.

Pada penelitian lainnya dalam melakukan klasifikasi level stress oleh Paseru dan Tapidingan (2020) dengan menggunakan algoritma KNN dan Naive Bayes, didapatkan akurasi masing-masing yaitu 86.61% untuk algoritma KNN dan 87.40% untuk algoritma Naive Bayes.

Berdasarkan penelitian terdahulu, nilai akurasi yang didapatkan oleh algoritma KNN sudah cukup baik. Namun nilai akurasi yang didapatkan oleh algoritma lain seperti Naive Bayes lebih tinggi dari algoritma KNN. Oleh karena dalam penelitian ini dilakukan klasifikasi stress menggunakan algoritma lain yaitu logistic regression dengan algoritma KNN sebagai pembandingnya dan penggunaan *hyperparameter* yang tepat untuk mendapatkan performa model terbaik.

# METODE PENELITIAN

Pada penelitian ini langkah-langkah yang akan dilakukan adalah pengumpulan data, EDA (*Exploratory Data Analysis)*, *pre-processing* data, *Modeling,* dan Evaluasi Model

Diagram 1. Flowchart Tahapan Penelitian

## 2.1. Pengumpulan data

Data yang digunakan pada penelitian ini didapatkan dari website Kaggle yang dikumpulkan oleh Reihaneh Namdari. Data tersebut memiliki jumlah baris sebanyak 27977 baris data dan 2 buah kolom. Kolom pertama berisi teks postingan beberapa orang dari sosial media Reddit tanpa disertai identitas penulis postingan. Sementara kolom kedua berisi angka yang hanya bernilai 0 atau 1, angka 0 menandakan bahwa tidak terdeteksi stres pada postingan tersebut dan angka 1 untuk sebaliknya.

## 2.2. *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis adalah proses menganalisis data untuk lebih memahami data yang akan digunakan. Pada banyak kasus, EDA akan disajikan dalam bentuk grafis agar mampu mengeksplorasi data secara lebih terbuka dan mampu memperkuat analisis terhadap data (Romadhoni, 2020).

Jenis grafik yang akan digunakan tidak bisa dipakai secara asal. Kita perlu mengetahui jenis, karakteristik data, dan informasi apa yang ingin ditampilkan.

## 2.3. Preprocessing Data

*Pre-processing* data dilakukan dalam beberapa tahap. Tahap pertama adalah mengatasi *missing value* dengan cara menghapus data yang punya nilai yang hilang pada salah satu kolomnya. Setelah penanganan *missing value* dilakukan, maka tahap selanjutnya adalah menghapus data duplikat. Data duplikat akan dihapus kecuali data pertamanya sehingga salah satu data tetap akan tersimpan.

Tahap selanjutnya adalah melakukan cleaning pada teks. Hal ini dilakukan dengan menggunakan bantuan pustaka NLTK (*Natural Language Toolkit*). Pusatakan ini merupakan pustaka dalam bahasa Python yang populer untuk pemrosesan bahasa alami. NLTK menyediakan berbagai alat dan sumber data yang membantu dalam melakukan analisis, pemrosesan, dan pemodelan teks berbahasa manusia.

Cleaning pada teks dilakukan dengan melakukan beberapa hal yaitu penghilangan tag html dan karakter non-alphabet, mengubah semua huruf ke huruf kecil untuk memudahkan proses analisis dan pembentukan model. Proses tokenisasi yaitu memecah teks menjadi bagian yang terpisah dan penghapusan stopword juga dilakukan untuk meningkatkan performa model dan kemudahan analisis.

## 2.3.1. TF-IDF Vectorizer

TF-IDF (*Term Frequency-Inverse Document Frequency*) Vectorizeradalah alat atau metode yang digunakan dalam pemrosesan bahasa alami (*Natural Language Processing*/NLP) untuk mengubah teks menjadi representasi numerik yang dapat digunakan oleh algoritma pemodelan mesin.

TF-IDF Vectorizer menghitung nilai TF-IDF untuk setiap kata dalam teks, yang kemudian digunakan untuk menghasilkan vektor representasi numerik. Nilai TF-IDF mengukur seberapa penting sebuah kata dalam konteks koleksi dokumen atau korpus teks yang lebih besar.

## 2.4 Modeling

*Modeling* dilakukan dengan melakukan pembentukan model dengan algoritma *machine learning* untuk melakukan klasifikasi pada dataset. *Modeling* dilakukan dengan melatih metode klasifikasi dengan *hyperparameter* yang optimal dengan menggunakan data latih untuk kemudian diuji pada tahap evaluasi.

## 2.4.1. K-Nearest Neighbor

*K-nearest neighbors* (KNN) adalah algoritma sederhana namun efektif yang digunakan untuk tugas klasifikasi dan regresi dalam pembelajaran mesin. Ini adalah algoritma non-parametrik yang tidak membuat asumsi tentang distribusi data yang mendasarinya.

Dalam algoritma KNN, dengan diberikan sebuah data baru, K tetangga terdekat dari himpunan data latih diidentifikasi. Jumlah K ditentukan sebelumnya dan dapat disesuaikan sesuai dengan masalah yang sedang diselesaikan. Kemudian, label atau nilai target dari data baru diestimasi berdasarkan mayoritas label atau nilai target dari K tetangga terdekat tersebut.

Algoritma ini sangat populer karena mudah dimengerti dan dapat memberikan hasil yang cukup baik pada banyak tugas pembelajaran mesin (Alpaydin, 2020).

## 2.4.2. Logistic Regression

*Logistic Regression* adalah algoritma pembelajaran mesin yang digunakan untuk melakukan klasifikasi pada data dengan dua kelas atau lebih. Dalam *Logistic Regression*, variabel dependen atau target (kelas) adalah variabel biner (dua kelas) atau kategori (King & Zeng, 2001).

*Logistic Regression* menghitung probabilitas kelas target untuk setiap data input dengan menggunakan fungsi logit atau sigmoid. Ini menghasilkan nilai-nilai probabilitas yang dapat dikonversi menjadi prediksi kelas dengan mengatur ambang batas probabilitas.

Algoritma ini sangat populer karena sederhana, mudah diinterpretasikan, dan dapat digunakan untuk memodelkan banyak masalah klasifikasi yang berbeda.

## 2.4.3. Tuning Hyperparameter

*Hyperparameter* adalah parameter yang digunakan dalam proses pembentukan model. Parameter ini nantinya akan mempengaruhi performa model dalam melakukan klasifikasi nantinya, sehingga diperlukan penentuan *hyperparameter* yang sesuai untuk mendapatkan model dengan performa terbaik. *Hyperparameter* untuk masing-masing metode klasifikasi berbeda. Untuk metode KNN *hyperparameter-*nya adalah K, *Life Size,* dan P. Sedangkan untuk Logistic Regression, *hyperparameter-*nya adalah iterasi maksimal, C, dan tol.

*Tuning hyperparameter* adalah proses mencari kombinasi optimal dari nilai *hyperparameter* dalam sebuah model atau algoritma machine learning. *Tuning* *hyperparameter* dilakukan untuk mencari kombinasi nilai *hyperparameter* yang paling baik atau optimal untuk mencapai kinerja model yang optimal. Tujuan dari *tuning* *hyperparameter* adalah untuk meningkatkan kinerja model, mengurangi *overfitting* atau *underfitting*, dan mengoptimalkan performa prediksi pada data baru.

*Tuning hyperparameter* dilakukan dengan menggunakan pustaka gridsearchCV. Pustaka ini juga menyediakan cross validation sebagai evaluasi terkait *hyperparameter* yang terbaik.

## 2.5. Evaluasi Model

Evaluasi model dilakukan dengan menguji model yang sudah dibentuk dengan menggunakan data uji untuk mendapatkan performa model dalam melakukan klasifikasi data yang belum dikenal. Evaluasi akan dilakukan dengan membandingkan dan melihat nilai akurasi dari model yang telah dibentuk.

# HASIL DAN PEMBAHASAN

Pada penelitian pembentukan model akan dilakukan dengan melatih model dengan data latih dan mengujinya dengan data uji. Data latih dan data uji memiliki perbandingan 8:2.

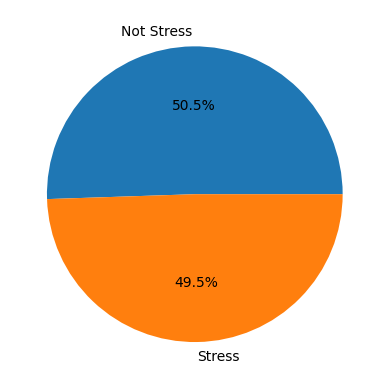
## 3.1. Tuning Hyperparameter Pada Tiap Metode

Hasil *tuning* *hyperparameter* pada algoritma Logistic Regression menunjukkan bahwa *hyperparameter* terbaik yang didapatkan adalah nilai *hyperparameter* C yaitu 1, nilai *hyperparameter* iterasi maksimal yaitu 100, dan nilai *hyperparameter* tol yaitu 0.001

Pada *tuning* *hyperparameter* algoritma KNN sendiri menunjukkan bahwa *hyperparameter* terbaik yang didapatkan adalah nilai *hyperparameter* K yaitu 1, nilai *hyperparameter* Leaf Size yaitu 15, dan nilai *hyperparameter* P yaitu 2.

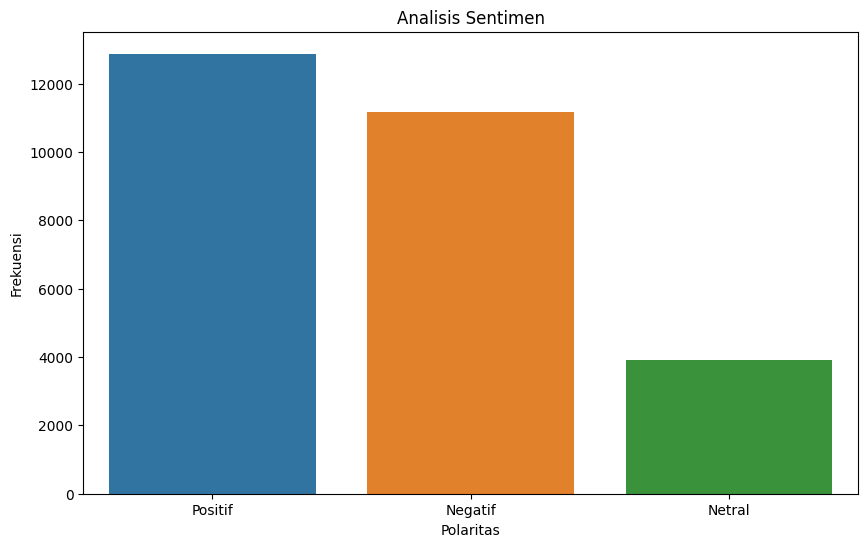
## 3.2. Hasil *Exploratory Data Analysis (EDA)*

Demi memahami data teks yang kami gunakan, kami telah membuat beberapa grafik yang dapat membantu kami dalam mengeksplorasi data.



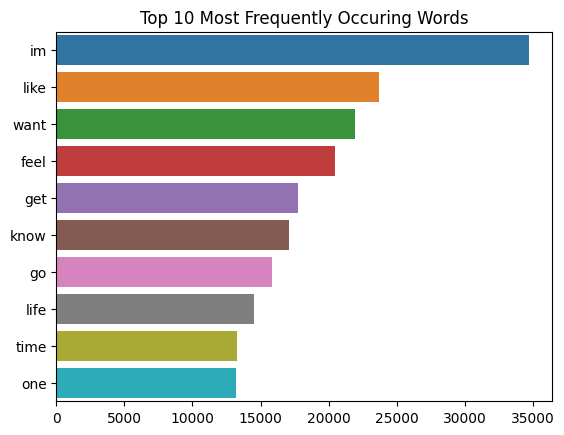
Gambar 1. Distribusi kategori

Data yang akan kami gunakan sudah seimbang dari segi pemerataan kategori/kelas. Hal ini sangat penting untuk mencegah adanya *underfitting* dan *overfitting* pada salah satu label.



Gambar 2. Analisis sentimen data

Sesuai hasil analisis sentimen, diketahui bahwa dari sekitar 27 ribu data teks postingan, diantaranya terdapat sekitar 40% konten negatif. Sementara itu, sisanya berupa konten positif dan konten yang netral.

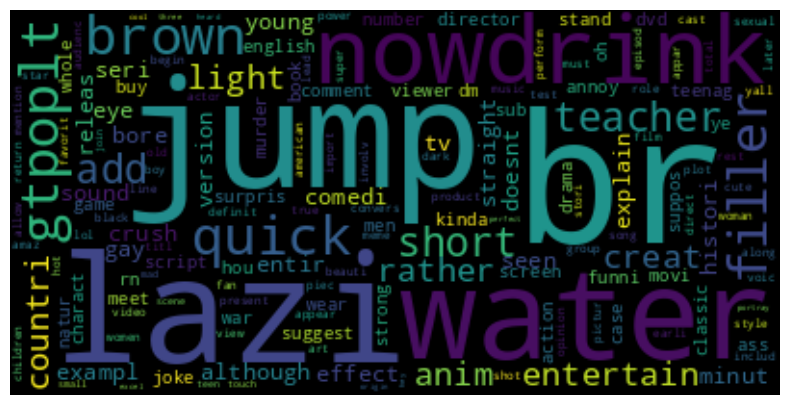


Gambar 3. Grafik 10 kata yang sering muncul

Berdasarkan gambar 3 di atas, kita dapat mengetahui jenis-jenis kata apa saja yang paling sering digunakan oleh penulis postingan. Jenis kata-kata yang sering digunakan adalah kata yang berkaitan dengan keinginan, perasaan, dan kehidupan.



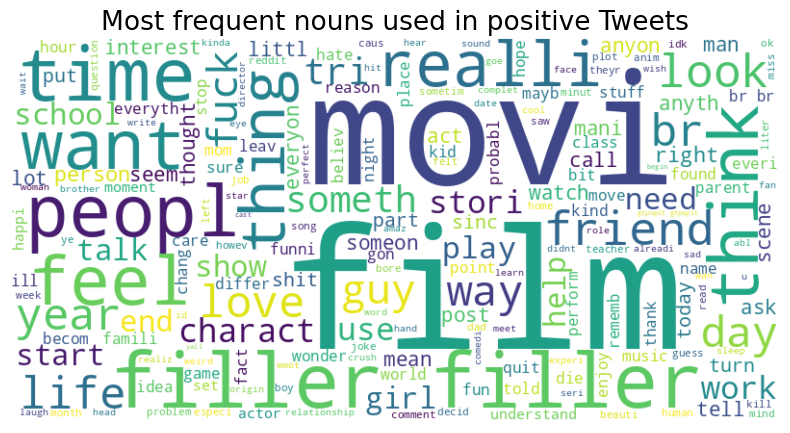
Gambar 4. Frekuensi kata pada post orang yang stress ditampilkan dalam WordCloud



Gambar 5. Frekuensi kata pada post orang yang tidak stress ditampilkan dalam WordCloud

Seperti yang terlihat pada WordCloud yang telah disajikan, kata-kata yang banyak muncul pada postingan orang yang stres adalah kata-kata yang sangat berkaitan dengan perasaan mendalam seperti permintaan maaf, makian, dan batin. Sementara itu, kata-kata yang banyak muncul pada post orang tidak stres bersifat acak dan bebas sesuai hal-hal menarik yang ada disekitar mereka.

Untuk mendapat pengetahuan lebih lanjut maka akan dilakukan pencarian kata benda yang digunakan saja pada data yang stres. Dapat diketahui kata yang sering muncul berkaitan dengan film, bioskop, cerita, dan kesharian mereka.



Gambar 6. Frekuensi kata benda pada post orang yang tidak stress

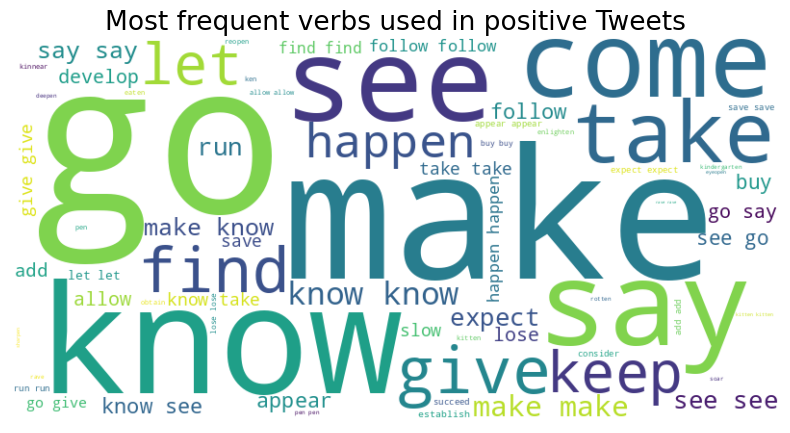
Pada gambar 6 ditampilkan kata benda yang sering ditulis seseorang pada postingan reddit seperti hobi, film dan sejenisnya. Namun pada postingan orang stress, kata benda yang sering muncul dapat dilihat pada gambar berikut.



Gambar 7. Frekuensi kata benda pada post orang yang tidak stress

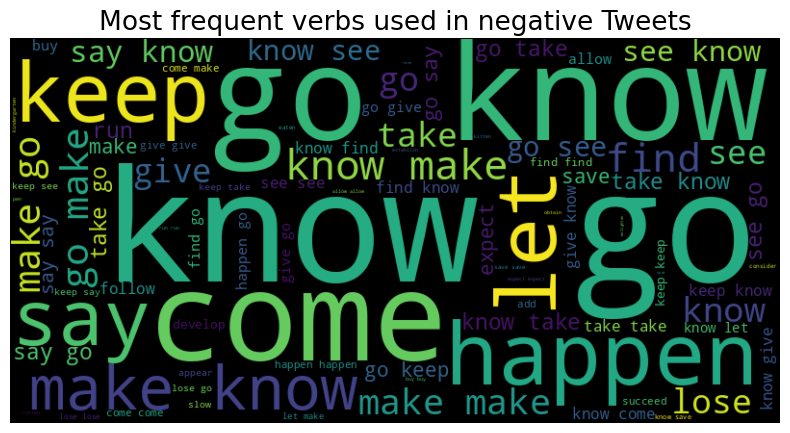
Pada postingan orang stress, kata benda yang sering muncul seperti perasaan, waktu, kehidupan, teman, dan orang lain.

Dilakukan EDA lebih lanjut untuk menganalisis perbedaan kata kerja pada postingan orang stres dan tidak stres.



Gambar 8. Frekuensi kata benda pada post orang yang tidak stress

Dapat dilihat pada Gambar 8. Dilakukan EDA lebih lanjut untuk menganalisis perbedaan kata kerja pada postingan orang stres dan tidak stres.



Gambar 9. Frekuensi kata benda pada post orang yang tidak stress

Dapat diketahui hasilnya bahwa kata kerja yang digunakan pada tweet negatif terlihat lebih beragam dan lebih banyak walaupun kebanyakan kata kerja yang digunakan sama.

## 3.3. Pengujian Algoritma

Pengujian algoritma Logistic Regression dengan hyperparameter C bernilai 1, iterasi maksimal 100, dan tol 0.001 menunjukkan akurasi yang didapatkan yaitu 91% atau 0.91.

Untuk algoritma KNN sendiri dengan hyperparameter K bernilai 1, Leaf Size bernilai 15 dan P bernilai 2 menunjukkan akurasi yang didapatkan yaitu 54% atau 0.54.

Berdasarkan Hasil diatas, dapat dilihat bahwa performa metode Logistic Regression dalam melakukan klasifikasi stres berdasarkan stres sangat baik dengan akurasi 0.91, berbeda dengan KNN yang akurasinya jauh lebih rendah yaitu 0.54.

# KESIMPULAN

Penyebab stres kebanyakan orang didominasi oleh perasaan dan keinginan seseorang dalam hidup mereka. Penyebab ini sangat umum dan kurang terlihat spesifik terkait cara mengatasi permasalahan tersebut. Penyebab lainnya adalah terkait pekerjaan, kesehatan, dan mental. Penyebab ini lebih spesifik dari penyebab sebelumnya sehingga solusinya sendiri bisa mudah untuk diusahakan.

Performa metode yang terbaik untuk klasifikasi stres dengan data yang berupa teks adalah metode Logistic Regression yang memiliki akurasi cukup tinggi yaitu senilai 0.91 atau 91 %, sedangkan metode KNN sendiri cukup rendah yaitu sekitar 0.54 atau 54 %.

*Hyperparameter* yang digunakan untuk mendapatkan performa terbaik masing-masing model adalah Logistic Regression dengan hyperparameter C bernilai 1, iterasi maksimal 100, dan tol 0.001 dan KNN dengan hyperparameter K bernilai 1, Leaf Size bernilai 15 dan P bernilai 2

.

# DAFTAR PUSTAKA

MAULANA, M.I., INDRIATI, SOEBROTO, A.A., 2019. Klasifikasi Tingkat Stress Berdasarkan Tweet pada Akun Twitter Menggunakan Improved K-Nearest Neighbor dan Chi-square. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (JPTIIK), 3(7), p.6662-6669.

TAPIDINGAN, Y. C., & PASERU, D. 2020. Comparative Analysis of Classification Methods of KNN and Naïve Bayes to Determine Stress Level of Junior High School Students. Indonesian Journal of Information Systems, 2(2),p.80–89.

LAZARUS, R.S., FOLKMAN, S.C., 1984. The Handbook of Behavioral Medicine. New York: Guilford.

MISRA, R., CASTILLO, L.G., 2004. Academic Stress Among College Students: Comparison of American and International Students. International Journal Of Stress Management, 11(2), p.132-148.

ALPAYDIN, E., 2010. Introduction to Machine Learning (2nd ed.). Cambridge, MA: MIT Press.

KING, G., & ZENG, L., 2001. Logistic Regression in Rare Events Data. Political Analysis, 9(2), P. 137–163.