**EVALUASI KINERJA ALGORITMA LOGISTIC REGRESSION DALAM KLASIFIKASI STRESS**

**Hafizh Raihan Kurnia Putra1, Jafar 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**

Stress adalah suatu keadaan ketika individu merasa tertekan, cemas, atau ketidakmampuan individu dalam menghadapi keadaan yang diberikan oleh lingkungan atau situasi tertentu. Orang yang kurang mampu mengatasi/menghadapi stress akan berakhir buruk seperti melukai dirinya sendiri, melukai orang lain, kemampuan berpikir terhambat, dan kasus terburuknya yaitu bunuh diri. Orang yang mengalami stress dapat terlihat dari postingan yang diunggah di social media. Hal tersebut dapat terlihat dari kata-kata yang digunakan dalam postingan yang diunggah. Proses deteksi stress yang kami lakukan menggunakan dataset yang berisi postingan pada platform Reddit. Metode klasifikasi yang akan digunakan dalam melakukan deteksi stress adalah *Logistic Regression* dengan metode pembanding untuk evaluasinya yaitu *KNN*. Pada penelitian ini juga akan melakukan vektorisasi terhadap teks menggunakan dua buah metode berbeda yaitu *Count Vectorizer* dan *TF-IDF Vectorizer*. Performa metode yang terbaik untuk klasifikasi stres dengan data yang berupa teks adalah metode *Logistic Regression* dengan menggunakan *TF-IDF Vectorizer* dan penggunaan fitur tambahan dengan hyperparameter C bernilai 1, iterasi maksimal 100, dan tol 0.001 yang memiliki akurasi senilai 0.74 atau 74 %, sedangkan metode *KNN* dengan hyperparameter K bernilai 1, Leaf Size bernilai 25 dan P bernilai 2 yaitu sekitar 0.61 atau 61 %.

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

***PERFORMANCE EVALUATION OF LOGISTIC REGRESSION ALGORITHM IN STRESS CLASSIFICATION***

***Abstract***

*Stress is a condition when individuals feel depressed, anxious, or the inability of individuals to deal with circumstances given by the environment or certain situations. People who are unable to cope with stress will end up hurting themselves, hurting others, inhibited thinking ability, and the worst case is suicide. People who experience stress can be seen from the posts uploaded on social media. This can be seen from the words used in the uploaded posts. Our stress detection process uses a dataset containing posts on the Reddit platform. The classification method that will be used in performing stress detection is Logistic Regression with a comparison method for evaluation, namely KNN. This research will also vectorize the text using two different methods namely Count Vectorizer and TF-IDF Vectorizer. The best method performance for stress classification with text data is the Logistic Regression method using TF-IDF Vectorizer and the use of additional features with hyperparameter C worth 1, maximum iteration 100, and toll 0.001 which has an accuracy of 0.74 or 74%, while the KNN method with hyperparameter K worth 1, Leaf Size worth 25 and P worth 2 is around 0.61 or 61%.*

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

# PENDAHULUAN

Stress adalah suatu keadaan ketika individu merasa tertekan, cemas, atau ketidakmampuan individu dalam menghadapi keadaan yang diberikan oleh lingkungan atau situasi tertentu. Stress dapat timbul dalam berbagai aspek kehidupan seseorang seperti tempat kerja, rumah, hubungan antara individu, situasi yang melibatkan perubahan besar, atau tekanan emosional. Orang yang kurang mampu mengatasi/menghadapi stress akan berakhir buruk seperti melukai dirinya sendiri, melukai orang lain, kemampuan berpikir terhambat, dan kasus terburuknya yaitu bunuh diri.

Orang yang mengalami stress dapat terlihat dari postingan yang diunggah di social media. Hal tersebut dapat terlihat dari kata-kata yang digunakan dalam postingan yang diunggah. Salah satu platform social media yang terkenal di luar Indonesia adalah Reddit. Dengan Reddit, semua orang dapat membuat postingan atau berkomentar sesuai dengan topik subreddit yang mereka inginkan.

Deteksi stress merupakan salah satu cara untuk mencegah dampak buruk dari stress yang terjadi pada orang lain. Proses deteksi stress yang kami lakukan menggunakan dataset yang berisi postingan pada platform Reddit.

Metode klasifikasi yang akan digunakan dalam melakukan deteksi stress adalah Logistic Regression dengan metode pembanding untuk evaluasinya yaitu KNN. Selain itu, pada penelitian ini juga akan melakukan vektorisasi terhadap teks menggunakan dua buah metode berbeda yaitu Count Vectorizer dan TF-IDF Vectorizer.

Pada penelitian yang dilakukan oleh Safri, Yofi & Arifudin, Riza & Muslim (2018) diketahui bahwa menggunakan metode K-Nearest Neighbor sebagai salah satu metode untuk melakukan klasifikasi pada data. Kasus pada penelitian ini yaitu klasifikasi pada dataset mengenai kartu indonesia sehat. yang diberikan pada orang miskin. Dengan menggunakan metode K-Nearest Neighbor didapatkan nilai akurasi berjumlah 0.82 atau 82%.

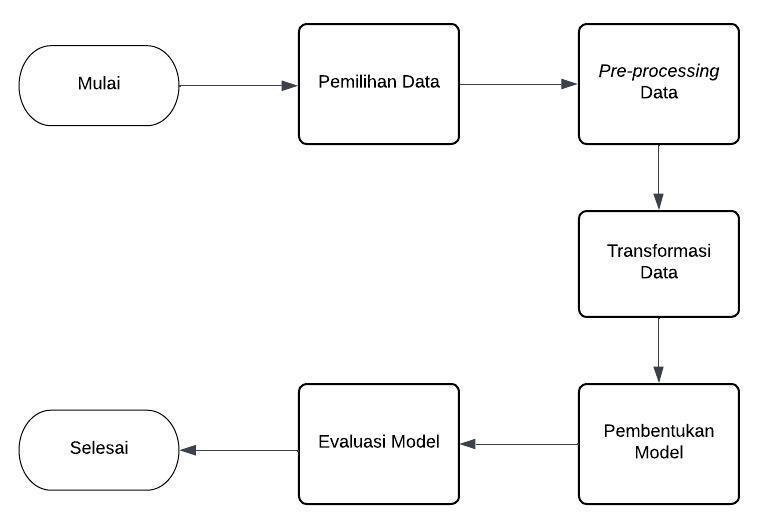
Selain itu, terdapat juga penelitian lain yang menggunakan metode K-Nearest Neighbor sebagai salah satu metodenya. Penelitian ini dilakukan oleh Hidayati, Nur dan Hermawan, Arief (2021) pada kasus klasifikasi kelulusan siswa, dengan dataset yang terdiri dari 10 fitur. Dengan menggunakan metode K-Nearest Neighbor dengan nilai K=7 dan menggunakan perhitungan jarak euclidean dan manhattan, menghasilkan akurasi sebesar 85.28%.

Terdapat penelitian yang dilakukan oleh Sah, Ramesh Kumar dan Ghasemzadeh, Hassan (2021) yang melakukan klasifikasi pada dataset stress dengan menggunakan metode Convolutional Neural Network (CNN)

Berdasarkan penelitian-penelitian yang telah dilakukan, deteksi stres dengan menggunakan algoritma KNN sudah banyak dilakukan dengan akurasi yang cukup baik namun masih bisa ditingkatkan lagi. Oleh karena itu dalam penelitian ini dilakukan klasifikasi stress menggunakan algoritma lain yaitu Logistic Regression dengan algoritma KNN sebagai pembandingnya dan juga kami menggunakan TFIDF Vectorizer dan CountVectorizer sebagai algoritma vektorisasi teks.

# METODE PENELITIAN

Pada penelitian ini akan dilakukan langkah-langkah yaitu pemilihan data, *pre-processing* data,, transformasi data, pembentukan model*,* dan evaluasi model

Diagram 1. Flowchart Tahapan Penelitian

## 2.1. Pemilihan Data

Data yang digunakan pada penelitian ini didapatkan dari website Kaggle yang dikumpulkan oleh Kreesh Rajani. Data yang digunakan merupakan dataset yang berisi komentar-komentar yang ditulis orang-orang dalam sosial media Reddit Data tersebut memiliki jumlah baris sebanyak 2820 baris data dan 7 buah kolom. Kolom pertama berisi nama subreddit tempat postingan, kolom kedua adalah angka unik yang menunjukkan identitas postingan, kolom ketiga berisi rentang kalimat dalam postingan, kolom keempat berisi teks dari postingan. Sementara kolom kelima berisi angka yang hanya bernilai 0 atau 1, angka 0 menandakan bahwa tidak terdeteksi stres pada postingan tersebut dan angka 1 untuk sebaliknya. Kolom keenam berisi tingkat kepercayaan diri dari pembuat postingan dan kolom terakhir berisi informasi mengenai waktu postingan sesuai dengan format yang relatif (e.g. “Baru saja”, “2 Jam lalu”, “Minggu lalu”, dll.)

## 2.2. 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.

**2.2.1 Pembentukan Fitur Baru**

Pembentukan fitur baru dilakukan untuk meningkatkan performa model yang telah dibuat. Pembentukan fitur baru sendiri bisa dilakukan dengan beberapa cara. Pada penelitian ini pembentukan fitur baru dilakukan dengan cara menghitung jumlah kata yang bersentimen positif dan kata yang bersentimen negatif sehingga dapat mempermudah model dalam melakukan klasifikasi terkait postingan tersebut masuk tergolong dalam kategori stres (negatif) ataupun tidak stres (positif).

Penghitungan kata positif dan negatif dilakukan dengan menggunakan pustaka NRCLex yang digunakan untuk mengukur emosi perasaan dari sebuah badan dari teks. Pustaka ini memiliki *dictionary* untuk perasaan yang memiliki sekitar 27.000 kata yang didasarkan pada National Research Council Canada (NRC) affect lexicon dan sinonim set dari pustaka NLTK.

## 2.3 Transformasi Data

Transformasi data dilakukan dengan melakukan vektorisasi terhadap data teks. Hal ini dilakukan agar data dapat diproses pada tahap selanjutnya yaitu tahap pembentukan model. Pada penelitian ini terdapat dua Jenis vektorisasi yang akan digunakan untuk mendapatkan model yang baik yaitu Count Vectorizer dan 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.

Hal ini berbeda dengan Count Vectorizer yang hanya akan menghitung kemunculan kata dalam suatu teks dan akan mengembalikan vektor yang berisi vektor jumlah kemunculan katanya saja tanpa ada perhitungan lebih lanjut.

## 2.4. Pembentukan Model

Pembentukan model dilakukan dengan membentuk model dari algoritma *machine learning* untuk melakukan klasifikasi pada dataset. Hal ini dilakukan dengan melatih model klasifikasi dengan *hyperparameter* yang optimal dengan menggunakan data latih untuk kemudian diuji pada tahap evaluasi dengan menggunakan data uji.

## 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 Logistic Regression *hyperparameter-*nya adalah iterasi maksimal, C, dan tol. Sedangkan untuk metode KNN *hyperparameter-*nya adalah K, *Life Size,* dan P.

*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 untuk algoritma Logistic Regression dan RandomizedSearchCV untuk algoritma KNN. Kedua pustaka ini 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.

Evaluasi dilakukan pada masing-masing vectorizer yang digunakan yaitu Count Vectorizer dan TF-IDF Vectorizer untuk membandingkan dan membentuk model dengan performa terbaik. Evaluasi dengan penambahan fitur baru juga dilakukan untuk mengukur pengaruh penggunaan fitur baru dalam melakukan klasifikasi.

Analisis terhadap hasil klasifikasi juga akan dilakukan untuk mengetahui penyebab dari *misclassification*  yang terjadi sehingga bisa digunakan sebagai pertimbangan dalam meningkatkan performa model dalam melakukan klasifikasi.

# 3. 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.

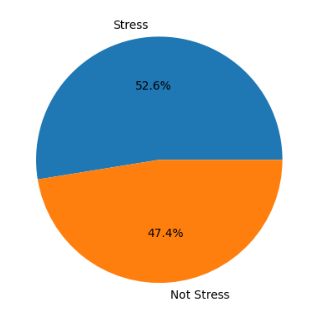
## 3.1. Tuning Hyperparameter Pada Tiap Metode

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 25, dan nilai *hyperparameter* P yaitu 2.

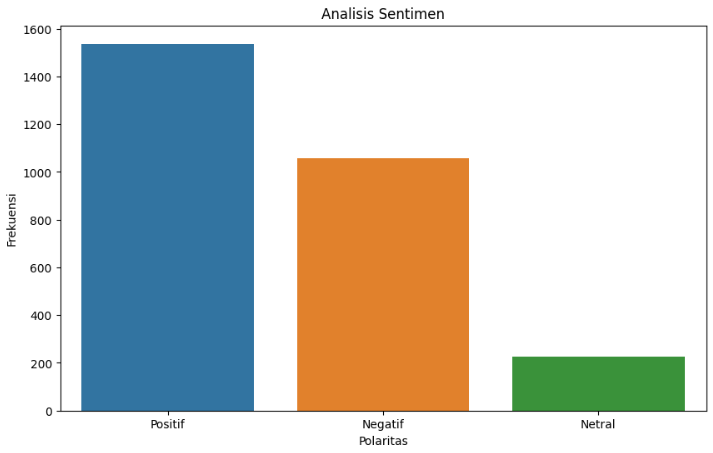
## 3.2. Exploratory Data Analysis (EDA)

Kami melakukan visualisasi EDA untuk melakukan eksplorasi terhadap data yang akan kami gunakan.



Gambar 1. Distribusi Kelas

Dari visualisasi yang ditampilkan, terlihat bahwa terdapat perbedaan jumlah data pada kedua kelas sebanyak 5.2%. Perbedaan tersebut masih tergolong tidak jauh sehingga dapat dikatakan jumlah data kedua kelas hampir seimbang.



Gambar 2. Analisis Sentimen

Analisis sentimen dilakukan untuk mengetahui tingkat kepositifan dalam sebuah dataset teks. Dalam kasus kami, kami menemukan bahwa terdapat sekitar 1600 data bernilai positif, sekitar 1000 data bernilai negatif, dan sekitar 200 data sisanya bernilai netral.



Gambar 3. WordCloud Label Stress



Gambar 4. WordCloud Label Tidak Stress

Berdasarkan visualisasi WordCloud, kita dapat melihat bahwa kata-kata yang sering muncul pada postingan dengan label stress adalah kata-kata yang terkait dengan perasaan negatif dan kekerasan. Sementara pada postingan dengan label tidak stress berisi kata-kata terkait keinginan seseorang untuk melakukan sesuatu dan kata benda di sekitar mereka.

## 3.3. Pengujian Algoritma

Pengujian pertama dilakukan dengan menggunakan Count Vectorizer pada transformasi data yang berupa teks pada dataset dan menggunakan hyperparameter berupa hyperparameter terbaik yang didapatkan dari proses tuning yang telah dilakukan sebelumnya yaitu algoritma Logistic Regression dengan hyperparameter C bernilai 1, iterasi maksimal 100, dan tol 0.001 dan algoritma KNN sendiri dengan hyperparameter K bernilai 1, Leaf Size bernilai 25 dan P bernilai 2.

Pengujian pada algoritma Logistic Regression menggunakan Count Vectorizer menunjukkan akurasi yang didapatkan yaitu 69% atau 0.69. Untuk algoritma KNN sendiri yang menggunakan Count Vectorizer menunjukkan akurasi yang didapatkan yaitu 59% atau 0.59.

Pengujian selanjutnya dilakukan dengan menggunakan TF-IDF Vectorizer untuk transformasi datanya dan hyperparameter yang telah digunakan sebelumnya.

Pengujian dengan menggunakan TF-IDF Vectorizer kali ini pada algoritma Logistic Regression menunjukkan akurasi yang lebih baik yaitu 73% atau 0.73. Sedangkan pada algoritma KNN menunjukkan juga menunjukkan peningkatan performa dengan akurasi sebanyak 60% atau 0.60.

Penggunaan TF-IDF menunjukkan sedikit peningkatan performa kedua model dalam klasifikasi yaitu 4% untuk model Logistic Regression dan 1% untuk model KNN.

Untuk meningkatkan performa maka model juga dilatih dengan menggunakan fitur tambahan yang berupa kata negatif dan kata positif. Oleh karena itu pengujian ini dilakukan dengan menggunakan TF-IDF Vectorizer untuk transformasi data dan menggunakan fitur tambahan serta hyperparameter yang sudah digunakan pada pengujian sebelumnya.

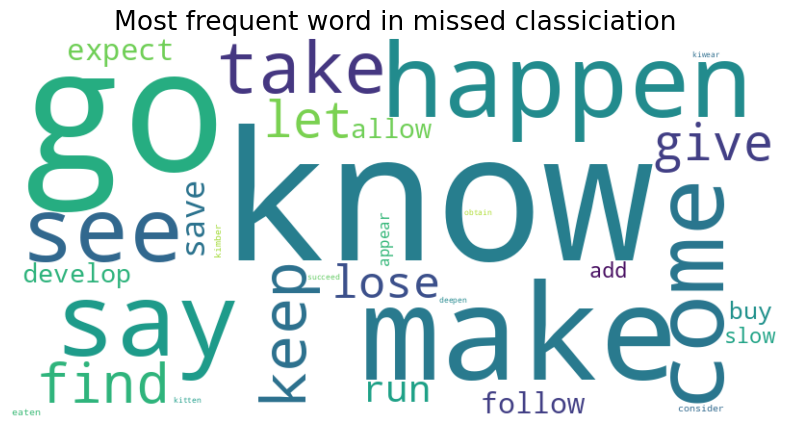
Pada pengujian dengan menggunakan TF-IDF Vectorizer dan penambahan fitur, model Logistic Regression mendapatkan akurasi 74% atau 0.74, sedangkan model KNN mendapatkan akurasi sebanyak 61% atau 0.61.

Dari akurasi dari kedua model pada pengujian ini, penggunaan fitur tambahan meningkatkan akurasi walaupun tidak terlalu signifikan yaitu sebanyak 0.1 atau 1% saja untuk kedua model.

## 3.4. Analisis Hasil Evaluasi

Hasil dari pengujian-pengujian yang telah dilakukan sebelumnya menunjukkan akurasi yang tidak buruk untuk model Logistic Regression dengan akurasi tertinggi yang didapatkan adalah 74%, sedangkan untuk model KNN memiliki akurasi yang cukup rendah untuk tertingginya yaitu 61% saja. Akurasi tersebut tentunya belum bisa dibilang baik sehingga perlu ditingkatkan lagi.

Sebagai dasar pertimbangan untuk meningkatkan performa model, analisis terhadap misklasifikasi data dilakukan untuk mengetahui penyebab mengapa data tersebut salah diklasifikasikan. Pada penelitian ini analisis dilakukan dengan mengecek fitur nilai fitur tambahan dan kata-kata yang sering muncul dalam data yang salah diklasifikasikan.



Gambar 5. Kata pada data yang salah diklasifikasikan

Dari data yang salah diklasifikasikan maka kata yang sering muncul adalah know, go, happen, make, say, come, dan beberapa kata lainnya. Kata tersebut merupakan kata yang bisa dibilang cukup netral, dimana tidak menunjukkan sentimen seperti positif atau negatif. Untuk memastikan maka kata tersebut juga akan nilai sentimennya menggunakan pustaka NRCLex dan dihasilkan bahwa nilai dari sentimen penghitungan positif dan negatif dari kata-kata tersebut tidak berbeda jauh yaitu 1.76 untuk sentimen positif dan 1.62 untuk sentimen negatif. Hal ini menunjukkan bahwa data yang memiliki kata-kata yang netral atau tidak menunjukkan sentimen, cenderung salah diklasifikasikan.

| **Emosi** | **Scores** |
| --- | --- |
| Positif | 0.17618857371154614 |
| Negatif | 0.16220535357570914 |

Tabel 1. Nilai Sentimen kata pada data yang salah diklasifikasikan

Analisis selanjutnya dilakukan terhadap fitur tambahan yang digunakan. Disini data-data yang salah diklasifikasikan akan dilihat nilai dari fitur tambahan tersebut yaitu jumlah kata positif dan kata negatif. Dari hasil pengamatan, dapat diketahui bahwa perbedaan nilai dari jumlah kata positif dan negatif kebanyakan sangat kecil. Hal ini menunjukkan bahwa dalam data tersebut tidak terlalu condong ke arah positif atau negatif sentimennya sehingga model tidak bisa mengklasifikasikan dengan sempurna.

# 4. KESIMPULAN

Orang stres kebanyakan menggunakan kata-kata yang menggambarkan perasaan mereka atau kata-kata yang bersentimen seperti meminta, stress, sedih dan lainnya. Hal ini berbeda dengan orang yang tidak stres yang cenderung menggunakan kata-kata random di keseharian mereka atau kata-kata yang tidak memiliki sentimen atau nilai perasan seperti email, teman, mengubah, dan lainnya.

Performa metode yang terbaik untuk klasifikasi stres dengan data yang berupa teks adalah metode Logistic Regression dengan menggunakan TF-IDF Vectorizer dan penggunaan fitur tambahan yang memiliki akurasi yang tidak buruk yaitu senilai 0.74 atau 74 %, sedangkan metode KNN sendiri cukup rendah yaitu sekitar 0.61 atau 61 % walaupun sudah menggunakan fitur tambahan dan TF-IDF Vectorizer.

*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 25 dan P bernilai 2.

Data yang salah diklasifikasikan kebanyakan memiliki kata dengan nilai sentimen yang tidak condong ke arah positif atau negatif. Data yang salah diklasifikasikan juga memiliki nilai yang tidak jauh beda pada fitur tambahan yaitu jumlah kata negatif dan jumlah kata positif pada data tersebut sehingga sedikit sulit untuk diklasifikasikan.

# 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.