Klasifikasi Data Tweet dengan Menggunakan Metode Klasifikasi Multi-Class Support Vector Machine (SVM) (Studi Kasus : PT.KAI)

Dhina Nur Fitriana[1], Yuliant Sibaroni, S.T, M.T[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]dhnnur@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id

Abstrak

Informasi dalam bentuk dokumen berbasis teks yang tidak terstruktur semakin banyak dan menjadi hal biasa keberadaannya di internet. Informasi tersebut sering ditemukan secara mudah dan dimanfaatkan oleh pelaku bisnis atau perusahaan melalui media sosial. Salah satu media sosial yang dibahas pada penelitian ini adalah Twitter. Twitter menempati peringkat ke-6 sebagai media sosial yang banyak diakses saat ini yaitu sebanyak 52 % pengguna di Indonesia. Pemakaian Twitter memiliki kelemahan yaitu data teks tidak terstruktur dan berjumlah banyak yaitu mencapai 2400 *tweet* per harinya. Hal ini mempersulit pelaku bisnis atau perusahaan mengetahui sentimen publik terhadap suatu layanan dengan sumber daya terbatas. Sentimen publik pada Twitter perlu diklasifikasikan ke dalam netral tidak hanya positif dan negatif agar dapat mempermudah perusahaan mengetahui sentimen publik untuk pelayanan yang lebih baik di masa yang akan datang. Metode *Support Vector Machine* (SVM) merupakan metode klasifikasi yang optimal dibandingkan metode Naïve Bayes. Kekurangan dari metode Support Vector Machine (SVM) yaitu menggunakan fungsi pemisah yang memisahkan data ke dalam dua kelas, jika kelas yang ingin dipisahkan lebih dari dua maka dibutuhkan modifikasi dan mempengaruhi waktu pelatihan dan ukuran memory yang dibutuhkan. Untuk menangani kasus klasifikasi *non-biner* pada penelitian ini diperlukan pendekatan *multi-class Support Vector Machine* (SVM) yang menangani klasifikasi tiga kelas. Penelitian ini menggunakan pendekatan *One Againts All* sebagai model untuk menentukan kelas yang tepat. Pendekatan *One Againts All* memiliki akurasi yang lebih baik dibandingkan *One Againts One*. Penelitian ini berisi hasil implementasi metode *multi-class Support Vector Machine* (SVM) OAA dengan lima fitur yang berbeda yaitu unigram, bigram, trigram, unigram+bigram, dan wordcloud saat mengklasifikasikan data *tweet* dalam jumlah yang banyak. Nilai akurasi tertinggi berasal dari pengujian model TF-IDF unigram yang dikombinasikan dengan metode klasifikasi *multi-class Support Vector Machine* (SVM) dengan nilai parameter *gamma 0.7* yaitu 80.59. *Multiclass Support Vector Machine* (SVM) dapat mengklasifikasikan kelas netral dengan baik karena banyaknya opini yang bersifat netral yaitu sebanyak 365 kalimat dari 402 kalimat netral namun, jika menggunakan metode *Support Vector Machine binary class* opini netral sulit diklasifikasikan.

Kata kunci : Klasifikasi Teks, *Multi-class Support Vector Machine, Term Frequency-Inverse Document Frequency*

Abstract

Information in the form of unstructured texts is increasing and becoming commonplace for its existence on the internet. This information is easily found and utilized by business people or companies through social media. One of them is Twitter. The use of Twitter has the disadvantage of an unstructured and large amount of text data, which reaches 2400 tweets per day. Consequently,, it is difficult for business people or companies to know public opinion towards service with limited resources. Public opinion on Twitter need to be classified into positive, negative, and neutral sentiments in order to know the response of customers for better service in the future. The Support Vector Machine (SVM) method is more optimal than the Naïve Bayes method. The weakness of the Support Vector Machine (SVM) method is that it uses a separator function that separates data into two classes. If the class wants to be separated more than two, modification is needed and affects the training time and memory size required. There are two approaches to implementing the multiclass Support Vector Machine method by combining several binary SVMs, namely One Against All (OAA) and One Against One (OAO). In this paper, this research contains the results of classifying multi-class Support Vector Machine (SVM) methods with five different weighting features for classifying tweet data and finding the best accuracy value when processed with large amounts of data. The results show that the TF-IDF feature extraction approach with unigram feature outperforms other methods allowing the classifier to achieve highest accuracy when work with larger datasets. The unigram TF-IDF combined with multi-class SVM has the highest average accuracy value of 80.59 compared to the other four models namely 52.53 bigrams, 53.54 trigrams, Unigrams + bigrams 76.13, and word cloud 70.33. The highest f-measure value gets from SVM multi-class method with the unigram feature and gamma parameter value of 0.7 which is 80.59. Multiclass SVM can classify neutral classes well. Multiclass SVM can classify 365 sentences out of 402 neutral sentences. Therefore, if using binary class classification, neutral is difficult to be classified.

Keyword : Text Classification, Multi-class Support Vector Machine, Term Frequency-Inverse Document Frequency, Transportation.

1. Pendahuluan

Latar Belakang

Informasi dalam bentuk dokumen berbasis teks yang tidak terstruktur semakin banyak dan menjadi hal biasa keberadaannya di internet. Hal tersebut terjadi karena meningkatnya pengguna internet dari tahun ke tahun[1]. Informasi tersebut sering ditemukan secara mudah dan dimanfaatkan oleh pelaku bisnis atau perusahaan melalui media sosial salah satunya Twitter. Twitter menempati peringkat ke-6 sebagai media sosial yang banyak diakses saat ini yaitu sebanyak 52 % pengguna di Indonesia[2]. Penggunaan media sosial Twitter yang banyak di Indonesia membuat pelaku bisnis memanfaatkannya sebagai media komunikasi untuk menyalurkan keluhan, saran, ataupun pertanyaan terhadap suatu pelayanan yang diberikan agar semakin baik di masa yang akan datang.

Pemakaian Twitter memiliki kelemahan yaitu data teks tidak terstruktur. Twitter berisikan keluhan tentang fasilitas, pertanyaan yang berkaitan dengan pelayanan ataupun apresiasi kepuasan pelanggan. Hal ini mempersulit pelaku bisnis atau perusahaan mengetahui sentimen publik terhadap suatu layanan dengan sumber daya terbatas. Terdapat kelemahan pada penelitian yang dilakukan oleh Windasari dkk.[3] yang membahas pengklasifikasian data Twitter Gojek yaitu klasifikasi hanya kedalam positif dan negatif sedangkan banyak ditemukan cuitan yang bersifat netral sehingga data perlu diklasifikasikan kedalam sentimen netral. Sentimen publik pada Twitter perlu diklasifikasikan ke dalam sentimen positif, negatif, dan netral agar perusahaan dapat dengan mudah mengetahui sentimen publik untuk pelayanan yang lebih baik di masa yang akan datang. Pengklasifikasian dilakukan dengan pendekatan TF-IDF serta *machine learning* seperti metode *Support Vector Machine* untuk memudahkan admin mengetahui informasi/respon dari pelanggan.

Term Frequency Inverse Document Frequency (TF-IDF) merupakan suatu cara atau metode untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. *Term Frequency* berarti banyaknya term dalam satu kalimat dan *Invers Document Frequency* merupakan nilai invers dari probabilitas kemunculan term pada suatu dokumen. TF-IDF dapat disesuaikan fiturnya dengan bentuk data dan dikombinasikan dengan metode *machine learning* untuk menyeleksi fitur terbaik dan akurat dalam klasifikasi data *tweet*.

Pada penelitian sebelumnya, Metode *Support Vector Machine* (SVM) merupakan metode klasifikasi yang optimal dibandingkan metode Naïve Bayes. Kekurangan dari metode Support Vector Machine (SVM) yaitu menggunakan fungsi pemisah yang memisahkan data ke dalam dua kelas, jika kelas yang ingin dipisahkan lebih dari dua maka dibutuhkan modifikasi dan mempengaruhi waktu pelatihan dan ukuran memory yang dibutuhkan. Untuk menangani kasus klasifikasi *non-biner* pada penelitian ini diperlukan pendekatan *multi-class Support Vector Machine* (SVM) yang menangani klasifikasi tiga kelas Terdapat dua pendekatan untuk mengimplementasi metode *multiclass Support Vector Machine* dengan menggabungkan beberapa SVM biner yaitu *One Againts All (* OAA) dan *One Againts One (*OAO). Pendekatan multi kelas SVM OAA dalam penelitian Hejazi dkk.[4], Pratama dkk.[5], dan Mustakim dkk.[6] memiliki keunggulan hasil akurasinya dibandingkan OAO. Berdasarkan pernyataan tersebut, penelitian ini akan mengklasifikasikan kalimat pengguna *twitter* kedalam positif, netral dan negatif menggunakan metode klasifikasi *Multi-Class Support Vector Machine* (SVM) *One Againts All* (OAA) yang menggunakan kernel fungsi basis radial dengan lima pendekatan pembobotan fitur TF-IDF yaitu unigram[3], bigram[7], trigram[8], unigram+bigram, dan wordcloud[9] untuk memetakan sentimen masyarakat kedalam positif, negatif, atau netral. Dari kombinasi lima fitur yang berbeda akan dicari nilai terbaik yang paling tepat. Masukan dari penelitian ini yaitu kumpulan data hasil *scrapping* data Twitter dan keluarannya berupa kinerja klasifikasi sentimen masyarakat.

Penelitian ini bertujuan mengetahui kinerja metode *multiclass Support Vector Machine* (SVM) untuk klasifikasi data Twitter @KAI121 dan mengetahui kelompok fitur terbaik dilihat dari nilai akurasi yang didapatkan. Sehingga, dapat mengetahui informasi berupa sentimen masyarakat dalam layanan fasilitas, pertanyaan, maupun keluhan terhadap Kereta Api Indonesia dengan mudah.

2. Studi Terkait

2.1. Twitter

Twitter merupakan salah satu media sosial yang banyak digunakan saat ini yang terdiri dari situs komunikasi yang saling terhubung antar pengguna dari berbagai latar belakang. Pesatnya penggunaan Twitter menyebabkan peningkatan jumlah data teks. Data teks adalah contoh informasi teks yang tidak terstruktur dan berbentuk sederhana yang mudah dipahami oleh manusia namun sulit dipahami oleh komputer. Teknik dan algoritma yang efektif dan efisien sangat diperlukan untuk menemukan suatu pola agar data teks mudah dipahami oleh komputer dan melakukan analisis untuk menghasilkan suatu keputusan. Penambangan teks adalah bidang penelitian yang mengekstraksi informasi yang bermakna dari teks. Bidang penelitian penambangan teks meliputi *text preprocessing* yang terdiri dari *tokenization, filtering, lemmatization, dan stemming* dan juga klasifikasi/ klustering yang terdiri dari beberapa metode seperti *Naïve Bayes, KNN, Decision Tree, Support Vector Machines, hierarchical clustering , k-means , probabilistic clustering.*[10]

Penelitian ini berkaitan dengan riset yang pernah dilakukan Windasari dkk. [3] tentang analisis sentimen milik perusahaan Gojek menggunakan metode n-gram unigram pembobotan TF-IDF dan *Support Vector Machine* (SVM). Hasil penelitian berupa prediksi data *tweet* yang dianggap sebagai sentimen positif atau negatif terhadap layanan GoJek dengan nilai akurasi 86%. Penelitian ini juga berkaitan dengan riset Arsya Monica dkk. [11] yang melakukan analisis sentimen opini maskapai penerbangan dengan fitur *Lexicon Based* dan menggunakan metode *Support Vector Machine* (SVM). Hasil dari penelitian ini menunjukkan parameter optimal dan pengaruh penggunaan Lexicon Based Features. Dengan digunakan parameter C bernilai 10 dan learning rate bernilai 0,03 serta digunakan Lexicon Based Features dengan iterasi sebanyak 50 kali memberikan hasil accuracy sebesar 40%, precision 40%, 100% recall, dan f-measure sebesar 57,14%.

2.2. Pembobotan Fitur/Kata

Pembobotan Fitur/Kata dalam teks memainkan peranan penting dalam klasifikasi teks karena dapat memengaruhi ketepatan klasifikasi. Pembobotan Fitur didasarkan pada model ruang vektor, di mana kata/fitur dipandang sebagai titik dalam ruang dimensi-N. Setiap dimensi titik mewakili satu fitur teks. Algoritma ektraksi fitur biasanya menggunakan kumpulan kata kunci. Berdasarkan kumpulan kata kunci yang telah didapatkan, algoritma pembobotan fitur menghitung bobot kata dalam teks atau dokumen kemudian membentuk vektor digital yang merupakan vektor fitur teks[12]. *Term Frequency Inverse Document Frequency* (TF-IDF) merupakan suatu cara atau metode untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. *Term Frequency* (TF) berarti banyaknya suatu istilah muncul dalam sebuah teks, dan IDF adalah singkatan dari *Inverse Document Frequency*, suatu algoritma yang digunakan untuk menghitung nilai invers dari probabilitas menemukan kata dalam sebuah teks [13].

Penelitian ini berkaitan dengan riset dalam bidang TF-IDF dengan beberapa fitur seperti unigram[3], bigram[7] dan trigram[8]. Algoritma FCDC TF-IDF yaitu pengembangan dari algoritma TF-IDF dengan fitur n-gram[14] yang pernah dilakukan beberapa peneliti sebelumnya. Pembobotan fitur unigram pernah diteliti oleh Windasari dkk. [3], pada penelitian tersebut membahas analisis sentimen positif dan negatif dari akun twitter gojek dengan menggunakan metode klasifikasi SVM dan TF-IDF dengan fitur n-gram unigram. Penelitian tersebut menghasilkan nilai akurasi 86%, tingkat kesalahan prediksi 14%, tingkat prediksi yang benar untuk 100% sentimen positif, dan tingkat prediksi yang benar untuk sentimen negatif 67,44%.

Penelitian sebelumnya tentang pembobotan fitur bigram pernah diteliti oleh Gleen dkk. [7] yang menjelaskan bahwa metode TF-IDF yang mengintegrasikan kolokasi sebagai fiturnya. Penelitian ini bertujuan untuk mengatasi kelemahan *Term Frequency-Inverse Document Frequency* (TF-IDF) dalam berurusan dengan istilah tunggal. Hasilnya menunjukkan bahwa adanya peningkatan akurasi sebesar 10 % dibandingkan TF-IDF tanpa integrase kolokasi.

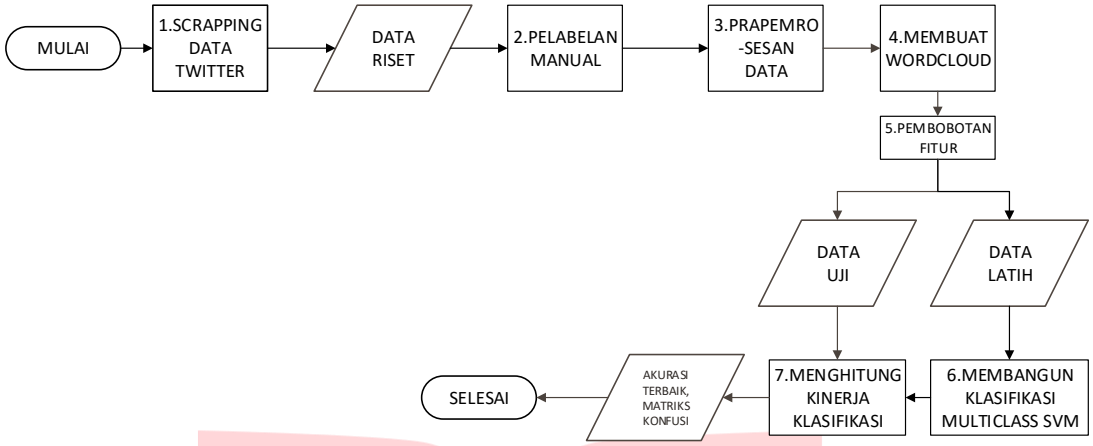
Penelitian sebelumnya tentang pembobotan fitur juga pernah diteliti oleh Wu dan Yuan [14] yang menjelaskan tentang algoritma FDCD TF-IDF untuk meningkatkan performa TF-IDF. Algoritma FDCD TF-IDF memperkenalkan konsep distribusi frekuensi kata dan distribusi kategori, algoritma ini mencerminkan korelasi antara item fitur dan kategori. Hasil percobaan menunjukkan bahwa algoritma FDCD TF-IDF dapat mencapai hasil klasifikasi yang ideal dan menunjukkan keefektifannya. Penelitian ini juga berfokus pada bagaimana mempercepat efisiensi algoritma sambil memastikan akurasi.

Penelitian yang berhubungan dengan fitur TF-IDF juga pernah diteliti oleh George dkk. [15]. Penelitian tersebut membahas klasifikasi sentimen otomatis dengan menggunakan metode SVM untuk ulasan hotel dan membandingkan dua metode yang berbeda yaitu TF-IDF *Bag of words* dan pendekatan *Term Occurrence.* Hasilnya. menunjukkan bahwa metode TF-IDF *Bag of words* efektif yang dapat dibandingkan dengan pendekatan mutakhir.

2.3. Multi-Class Support Vector Machine (SVM)

Penelitian sebelumnya terkait *multiclass Support Vector Machine* yang dilakukan oleh Hejazi dkk.[4] menjelaskan bahwa pendekatan *One Againts All* memiliki keunggulan dibandingkan *One Againts One*. Penelitian ini membahas masalah multi kelas dengan metode *Support Vector Machine* berdasarkan metode kernel RBF untuk pendekatan OAA dan OAO pada dataset aritmia jantung dengan nilai atribut yang kosong atau hilang. Hasil penelitian menunjukkan kesesuaian metode SVM OAA untuk analisis data EKG untuk aplikasi diagnostic karena adanya generalisasi. Penelitian lainnya yang terkait dengan pendekatan multi kelas svm dilakukan oleh Pratama dkk.[16] menyatakan bahwa nilai akurasi OAA semakin unggul saat dilakukan pengujian hingga tiga kali pada klasifikasi jenis dan fase malaria. Penelitian terkait metode klasifikasi multikelas SVM juga pernah diteliti oleh Mustakim dkk. [6]. Penelitian ini merancang sistem pengenalan ekspresi wajah untuk mengenali ekspresi dasar dengan menggunakan metode klasifikasi *multi-class* SVM dan membandingkan pendekatan OAA dan OAO. Hasilnya pengenalan ekspresi wajah menggunakan JAFFE dan Ekspresi Wajah Orang Indonesia menggunakan variasi 5 panjang gelombang, 8 sudut orientasi dan multikelas *One-Against-All* menghasilkan nilai akurasi yang terbaik yaitu sebesar 85,92%, dan 80,36%.

3. Sistem Klasifikasi yang Dibangun



Gambar 3-1 Sistem Klasifikasi Data Tweet dengan Metode Klasifikasi Multi-Class Support Vector Machine (SVM)

Berdasarkan Gambar 3-1 diatas, sistem klasifikasi yang dibangun pada penelitian ini adalah adalah sistem yang dapat mengklasifikasikan sentimen masyarakat terhadap pelayanan PT.KAI. Dalam penelitian ini analisis diambil berdasarkan *tweet* masyarakat pada akun Twitter @KAI121 melalui Twitter *scrapper*. Kumpulan dari *tweet* digunakan sebagai data latih dengan sebuah label dan selanjutnya dilakukan pengujian dengan data uji. Hasil dari performansi metode *multi-class Support Vector Machine* dan lima pendekatan TF-IDF yang berbeda mengenali *tweet* positif, negatif, dan netral menjadi fokus penelitian.

3.1. Pengumpulan Data

Pengumpulan data dilakukan dengan pengambilan data melalui Twitter *scrapper* dari akun @KAI121 sejak Januari 2018 hingga Januari 2020 sebanyak 7000 data yang akan menjadi data latih dan data uji dengan penentuan label sentimen secara manual. Penentuan label dilakukan dengan menganalisis *tweet* dan mengelompokkan kalimat yang mengandung kata-kata positif seperti bagus, keren, senang dan lain-lain kedalam kelas positif atau sebaliknya. Selanjutnya, dilakukan prapemrosesan data untuk mengoptimalkan fitur dari data yang maknanya sama jumlah lebih kecil agar mudah diproses. Contoh pelabelan kelas dapat dilihat pada Tabel 3-1 :

Tabel 3-1 Contoh Tweet dan Label/Kelas

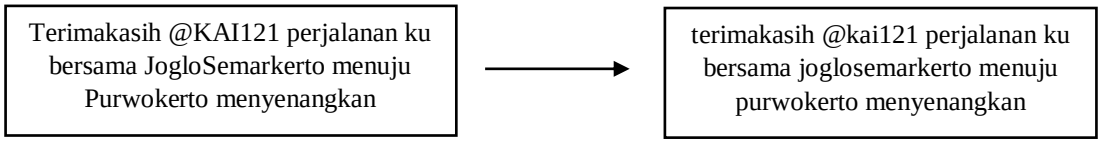
No	*Tweet*	Kelas
1	Terimakasih @KAI121 perjalanan ku bersama JogloSemarkerto menuju Purwokerto menyenangkan	Positif
2	Prosedurnya bagaimana ?	Netral
3	Adminnya tidak profesional	Negatif

3.2. Prapemrosesan Data

Pada data latih dan data uji dilakukan prapemrosesan data yang bertujuan untuk mengoptimalkan fitur dari data yang maknanya sama jumlah lebih kecil agar mudah diproses. Tahapan prapemrosesan data adalah sebagai berikut :

- *Case Folding*

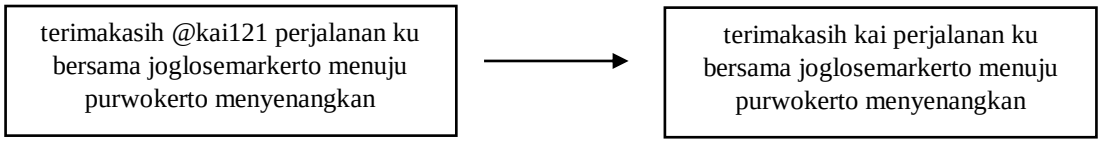
Case Folding merupakan langkah pada prapemrosesan data yang bertujuan untuk mengubah atau menghilangkan seluruh huruf kapital yang ada pada dokumen menjadi huruf kecil [11]. Data yang sudah dikumpulkan dari twitter dilakukan proses *Case Folding* terlebih dahulu seperti pada gambar berikut:



- *Remove Punctuation*

Remove Punctuation merupakan langkah yang dilakukan pada dokumen untuk menghapus atau menghilangkan beberapa tanda baca atau angka yang tidak memiliki hubungan terhadap dokumen. Tanda baca

atau angka yang tidak memiliki hubungan akan mengurangi nilai performance proses klasifikasi. Tahap *remove punctuation* digambarkan sebagai berikut :



- Pembakuan Kata

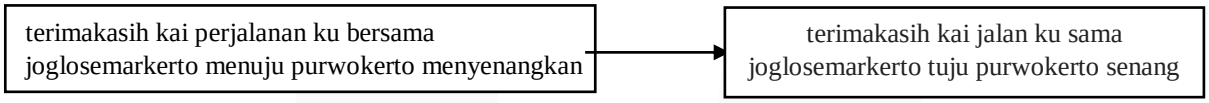
Pembakuan kata merupakan langkah yang dilakukan untuk mengubah singkatan, akronim, maupun kata ambigu pada dokumen *tweet*. Pembakuan kata dapat menangani data yang tidak seimbang. Tahap pembakuan mengkonversikan 530 kata yang peneliti dapatkan dengan menganalisis dan mengubah menjadi kata baku sesuai KBBI. Beberapa kata dalam daftar normalisasi dapat dilihat pada Tabel 3-2 :

Tabel 3-2 Contoh Normalisasi

aja	saja
aj	saja
gak	tidak
yg	yang
st	stasiun

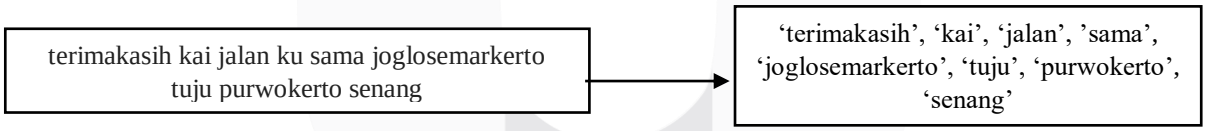
- *Stemming*

Stemming adalah proses menghilangkan awalan dan akhiran pada sebuah kata untuk mendapatkan *root* atau kata dasar dari suatu dokumen. Proses *stemming* pada penelitian ini menggunakan *library* sastrawi *stemming* berbahasa Indonesia dimana library tersebut menerapkan algoritma Nazief dan Andriani. Hasil dari tahap normalisasi sebelumnya diproses untuk melakukan *stemming* sebagai berikut:



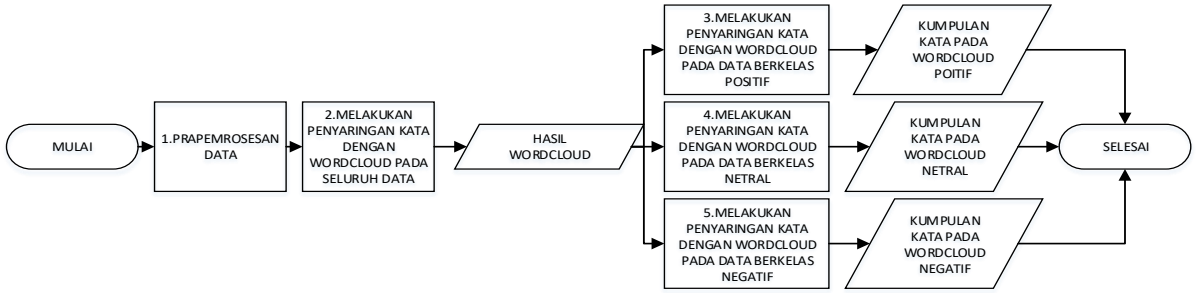
- Tokenisasi

Tokenisasi merupakan proses memecah urutan karakter menjadi beberapa bagian (kata/frasa) yang disebut dengan token [10]. Tokenisasi dilakukan untuk menghitung bobot fitur/kata pada tiap kalimat.dan digunakan untuk proses klasifikasi data teks lebih lanjut. Tahap tokenisasi digambarkan sebagai berikut:



3.3. Wordcloud

Wordcloud hadir sebagai metode visualisasi teks secara langsung dan menarik. *Wordcloud* biasanya digunakan dalam berbagai konteks sebagai sarana untuk memberikan ikhtisar dengan menyaring teks berupa kata-kata dengan nilai frekuensi yang tinggi[9]. Penelitian ini memanfaatkan wordcloud sebagai teknik untuk menyaring kata pada setiap sentimen yang selanjutnya akan dijadikan fitur pada saat proses TF-IDF. Proses wordcloud dilakukan dengan beberapa tahap. Gambaran untuk mendapatkan fitur *wordcloud* dapat dilihat pada Gambar 3-2:



Gambar 3-2 Proses untuk Mendapatkan Fitur Wordcloud

Hasil dari proses Wordcloud pada Gambar 3.2 yang telah dilakukan dapat dilihat pada Tabel 3-3 :

Tabel 3-3 Hasil Wordcloud

Stopword	Wordcloud Positif	Wordcloud Netral	Wordcloud Negatif
kai access, ini, nya, sampai, dari, di, saya mau, beli tiket, tidak, baru, kalau, saya, terima kasih, tidak bisa, jalan, yang, haru, tapi, karena. kakak, juga, ke, kereta, admin kai, sama, ya admin, tuju, mau tanya, bagaimana ya, tiket kereta, kereta api, dan, di stasiun, atau, dengan, sekarang, buat, jadi, lagi, saja, ya, itu, sudah, ada, admin mau, apakah, tidak ada, jam, apa, untuk.	aman, guna, moga, eksekutif, bersih, naik, sangat, lebih, jalan, alhamdulillah, layan, ekonomi, malam, dapat, mantap, gerbong, kembali, terima kasih, sekali, suka, lokal, enak, bagus, masih, kursi, bisa, banget, tugas, bagus, makin, bikin, hari, banyak, semua, biar, seperti, tambah, nyaman, selalu, tumpang, bapak, haru, sedia, stasiun, aku, dong, makan, keren, pakai, thank	dapat, habis, aku, berapa, sedia, aplikasi, terus, belum, masih, ktp, bisa, harga, gambir, haru, loket, hari, pasar senen, bagaimana, mesan, nanya, sore, jadwal, batal, bandung, lewat, apakah, bayar, mana, bagaimana cara, pada, tariff khusus, berangkat, mohon info, gerbong, jalan, tumpang, kenapa, naik, apakah bisa, pesan, kursi, lokal, pakai, kapan, buka, malangekonomi, seperti, Jakarta, tanya	gerbong, terus, apakah, aplikasi, lalu, coba, jalan, tolong, lebih, masih, tanya, seperti, bayar, cek, kali, dapat, tumpang, pesan, banyak, belum, selalu, lokal, malah, benar, padahal, kenapa, hari, mohon, bisa, sih, tadi, masuk, telat, bagaimana, jadwal, pakai, harga, naik, nih, pas, berangkat, error, kursi, lama, semua, gabisa, harus, muncul, saat, banget.

3.4. Pembobotan TF-IDF

Pembobotan fitur merupakan suatu cara atau metode untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Metode yang diterapkan pada penelitian ini adalah pembobotan fitur unigram, bigram, trigram, unigram+bigram, dan fitur wordcloud. Bentuk model n-gram berbasis kata selanjutnya akan dilakukan pembobotan pada setiap kata yang membentuk sebuah kalimat *tweet*. Wordcloud merupakan fitur yang sering banyak muncul pada *tweet* berkelas positif, negatif, dan netral untuk mengetahui pengaruh terhadap proses klasifikasi. *Tweet* yang berisi kata-kata langka memiliki bobot lebih tinggi daripada *tweet* yang mengandung kata-kata umum dan memiliki efek lebih besar pada klasifikasi.

Penentuan nilai bobot dalam metode TF-IDF didasarkan pada frekuensi kemunculan istilah dalam data riset. Metode ini dapat menghasilkan vektor fitur dengan jumlah besar pada corpus teks yang besar yang berpotensi dapat meningkatkan peluang untuk menyesuaikan model klasifikasi. Perhitungan TF dan IDF dapat dilihat pada persamaan 3.1 dan 3.2 :

$$Wi = TF(\omega i, d)\ x\ IDF\ (\omega i) \quad (3.1)$$

$$IDF(\omega i) = \log\left(\frac{|D|}{DF(\omega i)}\right) \quad (3.2)$$

Keterangan:

Wi = bobot kata term (ωi) dalam sebuah dokumen (d).
TF = *Term Frequency*, banyaknya term dalam satu kalimat.
DF = *Document Frequency*, banyaknya term/kata dalam satu dokumen.
|D| = Banyaknya kalimat dalam satu dokumen.
IDF(ωi) = *Invers Document Frequency*. Nilai invers dari probabilitas kemunculan term/kata (ωi) dalam dokumen. Nilai IDF terbesar muncul ketika ωi hanya muncul dalam satu dokumen.

Ilustrasi perhitungan nilai Wi, TF, DF, D, dan IDF dapat dilihat sebagai berikut (untuk TI adalah kalimat pernyataan positif, T2 adalah kalimat pernyataan netral, dan T3 adalah kalimat pernyataan negatif berasal dari Tabel 3-1)

Tabel 3-4 Tabel Contoh Hasil TF-IDF

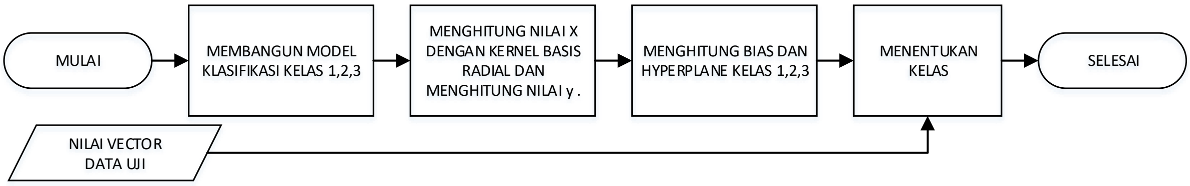
Term Unigram	Nilai TF			DF	D	IDF	W		
	T1	T2	T3				W(T1)	W(T2)	W(T3)
terima	1	0	0	1	3	0.477	0.477	0	0
kasih	1	0	0	1	3	0.477	0.477	0	0
kai	1	0	0	1	3	0.477	0.477	0	0
jalan	1	0	0	1	3	0.477	0.477	0	0
sama	1	0	0	1	3	0.477	0.477	0	0
Joglosemarkerto	1	0	0	1	3	0.477	0.477	0	0
tuju	1	0	0	1	3	0.477	0.477	0	0
purwokerto	1	0	0	1	3	0.477	0.477	0	0
senang	1	0	0	1	3	0.477	0.477	0	0
prosedur	0	1	0	1	3	0.477	0	0.477	0
bagaimana	0	1	0	1	3	0.477	0	0.477	0
admin	0	0	1	1	3	0.477	0	0	0.477
tidak	0	0	1	1	3	0.477	0	0	0.477
profesional	0	0	1	1	3	0.477	0	0	0.477
Term Bigram	TF			DF	D	IDF	W		
	T1	T2	T3				T1	T2	D3
terima kasih	1	0	0	1	3	0.477	0.477	0	0
kasih kai	1	0	0	1	3	0.477	0.477	0	0
kai jalan	1	0	0	1	3	0.477	0.477	0	0
jalan sama	1	0	0	1	3	0.477	0.477	0	0
sama joglosemarkerto	1	0	0	1	3	0.477	0.477	0	0
joglosemarkerto tuju	1	0	0	1	3	0.477	0.477	0	0
tuju purwokerto	1	0	0	1	3	0.477	0.477	0	0
purwokerto senang	1	0	0	1	3	0.477	0.477	0	0
prosedur gimana	0	1	0	1	3	0.477	0	0.477	0
Admin tidak	0	0	1	1	3	0.477	0	0	0.477
Tidak profesional	0	0	1	1	3	0.477	0	0	0.477

Pada Tabel 3.4 berisi perhitungan pembobotan fitur TF-IDF unigram dan bigram. Hasil dari pembobotan tersebut akan dijadikan vektor masukan untuk diproses menggunakan algoritma *Multiclass Support Vector Machine* (SVM). Ukuran matriks vektor masukan, ditentukan oleh jumlah record dan jumlah fitur dari masing-masing skema.

3.5. Model yang Dibangun

Metode *Support Vector Machine* (SVM) adalah metode untuk melakukan pengklasifikasian data linear dan *non* linear. Cara kerja dari algoritam SVM adalah dengan menggunakan pemetaan *non* linear untuk mengubah data latih ke dimensi yang lebih tinggi dan mencari *hyperplane* pemisah yang paling optimal. Data yang berada pada *hyperplane* disebut *support vector*[17]. Metode *Support Vector Machine* merupakan metode klasifikasi *supervised learning* yang menangani kasus klasifikasi biner. Untuk kasus klasifikasi *non-biner* seperti klasifikasi positif, negatif, dan netral diperlukan pendekatan *Multi-Class Support Vector Machine* (SVM) yang menangani klasifikasi lebih dari dua kelas. Terdapat dua pendekatan untuk mengimplementasi metode *multiclass Support Vector Machine* dengan menggabungkan beberapa SVM biner yaitu *One Againts All (* OAA) dan *One Againts One (*OAO) atau menggabungkan optimasi dari semua data. Pendekatan OAA menyelesaikan masalah *multi class* atau lebih dari dua kelas (N kelas) dengan N *decision boundary*. *Decision boundary* yang dihasilkan merupakan hasil dari pencarian hyperplane dari setiap kelas ke-*i* dengan kelas sisa lainnya. Pendekatan OAO menyelesaikan masalah *multi-class* atau lebih dari dua kelas (N kelas) dengan N(N-1)/2 *decision boundary*. *Decision boundary* yang dihasilkan merupakan hasil dari pencarian *hyperplane* dari setiap kelas dengan setiap satu kelas lainnya. Penelitian ini menggunakan pendekatan OAA sebagai model untuk menentukan kelas yang tepat. Pendekatan OAA memiliki performa yang lebih baik dibandingkan dengan pendekatan OAO dan juga lebih sederhana dibandingkan menggabungkan optimasi dari semua kelas data.

Proses klasifikasi *tweet* dalam penelitian ini terbagi menjadi dua tahap yaitu *training* untuk pembentukan model menggunakan metode *Multi-Class Support Vector Machine* dan tahap *testing* . Tahap *training* berisi proses pembentukan tiga model klasifikasi biner yang nantinya akan digunakan untuk mengklasifikasikan *tweet* kedalam tiga kelas yaitu positif, negatif, dan netral. Ilustrasi proses *training* dan *testing* adalah sebagai berikut:



Gambar 3-3. Gambaran Membangun Model Klasifikasi

Berikut penjabaran dan contoh proses alur proses training dan testing pada Gambar 3-3 :

1. Formulasi (W) yang digunakan adalah dualitas Langrange Multipler yang dimodifikasi untuk x dengan kernelnya.
2. Melakukan kernelisasi pada set data K(*x,xi*) dari fitur dimensi lama sehingga mendapatkan set data dengan fitur baru dimensi tinggi. Kernel yang digunakan adalah kernel RBF. Proses kernelisasi adalah sebagai berikut:

Tabel 3-5 Perhitungan x-xi

Data Pelatihan Kelas 1			Data Pelatihan Kelas 2			Data Pelatihan Kelas 3		
x1-x1	x1-x2	x1-x3	x2-x1	x2-x2	x2-x3	x3-x1	x3-x2	x3-x3
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	0.477	0.477	-0.477	0	0	-0.477	0	0
0	-0.477	0	0.477	0	0.477	0	-0.477	0
0	-0.477	0	0.477	0	0.477	0	-0.477	0
0	0	-0.477	0	0	-0.477	0.477	0.477	0
0	0	-0.477	0	0	-0.477	0.477	0.477	0
0	0	-0.477	0	0	-0.477	0.477	0.477	0

Setelah didapatkan pengurangan *x-xi* pada Tabel 3-5, maka selanjutnya dilakukan perhitungan untuk mendapatkan panjang vector. Hasil perhitungan panjang vektor dapat dilihat pada Tabel 3-6 :

Tabel 3-6 Perhitungan ||x-xi||

||x1-x1||	||x1-x2||	||x1-x3||	||x2-x1||	||x2-x2||	||x2-x3||	||x3-x1||	||x3-x2||	||x3-x3||
0	2.502	2.730	2.502	0	1.137	2.730	1.137	0

Hasil dari panjang vektor pada Tabel 3-6 kemudian dimasukkan ke dalam persamaan kernel RBF. Nilai gamma yang digunakan sebesar 0.5. Hasil perhitungan kernel RBF adalah sebagai berikut:

Tabel 3-7 Perhitungan exp(-y||x-xi||²)

K(1,1) $= \exp(-\gamma\|\|x1-x1\|\|^2)$ $= \exp((-0.5)(0)^2)$ $= 1$	K(1,2) $= \exp(-\gamma\|\|x1-x2\|\|^2)$ $= \exp((-0.5)(2.502)^2)$ $= 0.043$	K(1,3) $= \exp(-\gamma\|\|x1-x3\|\|^2)$ $= \exp((-0.5)(2.730)^2)$ $= 0.024$
K(2,1) $= \exp(-\gamma\|\|x2-x1\|\|^2)$ $= \exp((-0.5)(2.502)^2)$ $= 0.043$	K(2,2) $= \exp(-\gamma\|\|x2-x2\|\|^2)$ $= \exp((-0.5)(0)^2)$ $= 1$	K(2,3) $= \exp(-\gamma\|\|x2-x3\|\|^2)$ $= \exp((-0.5)(1.137)^2)$ $= 0.524$
K(3,1) $= \exp(-\gamma\|\|x3-x1\|\|^2)$ $= \exp((-0.5)(2.730)^2)$ $= 0.024$	K(3,2) $= \exp(-\gamma\|\|x3-x2\|\|^2)$ $= \exp((-0.5)(1.137)^2)$ $= 0.524$	K(3,3) $= \exp(-\gamma\|\|x3-x3\|\|^2)$ $= \exp((-0.5)(0)^2)$ $= 1$

3. Setelah melakukan perhitungan kernel pada Tabel 3-7, tahap selanjutnya adalah melakukan perhitungan terhadap nilai y. Nilai y adalah nilai label atau nilai dari kelas yang telah diberikan. Nilai y dapat dilihat pada Tabel 3.8:

Tabel 3-8 Nilai Label (y)

Nilai y pada training kelas 1			Nilai y pada training kelas 2			Nilai y pada training kelas 3		
y1	y2	y3	y1	y2	y3	y1	y2	y3
1	-1	-1	-1	1	-1	-1	-1	1

Tahap selanjutnya adalah melakukan perhitungan $yiyi^T$ yaitu perhitungan kernel. Nilai y adalah nilai label yang diberikan. Untuk tahap pelatihan kelas 1 didapatkan nilai y dapat dilihat pada Tabel 3.9:

Tabel 3-9 Nilai y Tiap Pernyataan

y1	y2	y3
-1	1	1

4. Kemudian tahap selanjutnya adalah mencari nilai a. Proses mendapatkan nilai a diawali dengan mengubah setiap pernyataan menjadi nilai vector (support vector) = $\begin{cases} \sqrt{x^2+y^2>2} \rightarrow \binom{4-y+\lvert x-y\rvert}{4-x+\lvert x-y\rvert} \\ \sqrt{x^2+y^2\leq 2} \rightarrow \binom{x}{y} \end{cases}$.
Sebagai contoh perhitungan dilakukan pada pernyataan pertama. Proses perhitungan adalah sebagai berikut:

$\sqrt{1^2+-1^2} = \sqrt{2}$ -> $\binom{1}{-1}$

5. Setelah itu masing-masing support vector diberi nilai bias 1. Untuk mendapatkan jarak tegak lurus yang optimal serta membantu mendapatkan nilai b atau *hyperplane*. Kemudia kalikan setiap kalimat menggunakan persamaan $\sum_{i=1,j=1}^{n} a_i S_i^T S_j$, sebagai contoh perhitungan pada pernyataan pertama sebagai berikut :

$a_1 \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}^T * \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = 3\ a_1$

6. Setelah dilakukan perhitungan pada seluruh pernyataan. Kemudian cari parameter *ai* menggunakan persamaan $\sum_{i=1,j=1}^{n} a_i S_i^T S_j = yi$. Sehingga bentuknya dapat dilihat sebagai berikut :

3 *a1* + 2.002 *a2* + 2.001 *a3* = 1
2.002 *a1* + 3 *a2* + 2.270 *a3* = -1
2.001 a1 + 2.270 *a2* + 3 *a3* = -1

Sedemikian sehingga didapatkan nilai *a1*, *a2* dan *a3* adalah sebagai berikut :

a1 = -418.077 *a2*= 1427.4 *a3*= -802.122

7. Setelah didapatkan nilai *ai* masukkan ke persamaan seperti berikut untuk mendapatkan nilai w dan b

$$W = -418.077\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} + 1427.4\begin{bmatrix} 0.043 \\ 1 \\ 1 \end{bmatrix} + -802.122\begin{bmatrix} 0.024 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} -418.077 \\ 418.077 \\ -418.077 \end{bmatrix} + \begin{bmatrix} 61.378 \\ 1427.4 \\ 1427.4 \end{bmatrix} + \begin{bmatrix} -19.251 \\ -802.122 \\ -802.122 \end{bmatrix} = \begin{bmatrix} -375.95 \\ 1043.355 \\ 207.201 \end{bmatrix}$$

$W1 = \begin{bmatrix} -375.95 \\ 1043.355 \end{bmatrix}$, B1 = 207.201

8. Langkah untuk menemukan hyperplane dua dan tiga sama seperti menentukan hyperplane pertama. Berikut ini merupakan nilai hyperplane dua dan tiga :

W2 = $\begin{bmatrix}-989.111\\2647.599\end{bmatrix}$, B2 = 207.201 W3 = $\begin{bmatrix}64.198\\1811.445\end{bmatrix}$, B3 = 207.201

9. Setelah mendapatkan nilai hyperplane pertama hingga ketiga, selanjutnya dapat menentukan kelas data uji masuk ke dalam kelas positif, netral atau negatif.

Misal, data uji memiliki nilai support vector (120.112,2) maka pada tahap pengujian nilai vektor disubstitusikan kedalam persamaan berikut :

$$kelas\ x = \arg\max([w^1]^T.\varphi(x) + b^1, [w^2]^T.\varphi(x) + b^2, [w^3]^T.\varphi(x) + b^3 \qquad (3.3)$$

$$kelas\ x = \arg\max(\begin{bmatrix}-375.95\\1043.355\end{bmatrix}^T.\begin{bmatrix}120.112\\2\end{bmatrix} + 207.201\ ,\begin{bmatrix}-989.111\\2647.599\end{bmatrix}^T.\begin{bmatrix}120.112\\2\end{bmatrix} + 207.201\ ,\ \begin{bmatrix}64.198\\1811.445\end{bmatrix}.\begin{bmatrix}120.112\\2\end{bmatrix} + 207.201)$$

= arg max (-42862.189 , 113301.701 , 10281.043)

Nilai hyperplane terbesar adalah 113301.701 dimana nilai hyperplane tersebut merupakan nilai kelas 2 . Berarti data uji termasuk kedalam kelas netral.

3.6. Performa Klasifikasi

Performa sistem klasifikasi menggambarkan seberapa bagus sistem tersebut dalam mengklasifikasikan data. *Confussion Matrix* merupakan salah satu metode yang digunakan untuk mengukur performa dari suatu metode klasifikasi. Pada dasarnya, *confussion matrix* berisikan perbandingan antara hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang sebenarnya [18]. Data uji yang dimasukkan kedalam matriks konfusi akan menghasilkan nilai *akurasi*, . Confussion Matrix dalam penelitian ini dapat dilihat pada Tabel 3.10 :

Tabel 3-10 Tabel Matriks Konfusi

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif	Terklasifikasi Netral
Positif	*True Positive (TP)*	*False Negative (FNe)*	*False Netral (FNt)*
Negatif	*False Positive (FP)*	*True Negative (TNe)*	*False Netral (FNt)*
Netral	*False Positive (FP)*	*False Negative (FNe)*	*True Netral (TNt)*

Berdasarkan nilai True Negative (TNe), True Netral (TNt) , False Netral (FNt), False Positive (FP), False Negative (FNe), dan True Positive (TP) dapat diperoleh nilai akurasi. Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar. Nilai akurasi dapat diperoleh dengan persamaan berikut :

$$Akurasi = \frac{TP + TNe + TNt}{TP + TNe + TNt + FP + FNe + FNt} * 100\% \qquad (3.4)$$

Variabel TP (*True Positive*) merupakan jumlah data positif yang terklasifikasi dengan benar. Variable TNe (*True Negative*) merupakan jumlah data negative yang terklasifikasi dengan benar. Variabel TNt (*True Netral*) merupakan jumlah data netral yang terklasifikasi dengan benar. Variabel FP (*False Positif*) merupakan jumlah data positif namun terklasifikasi salah oleh sistem. Variabel FNe (*False Negatif)* merupakan jumlah data negatif namun terklasifikasi salah oleh sistem.

4. Evaluasi

4.1. Hasil Pengujian

Pada bagian ini akan dijabarkan hasil pengujian dengan menampilkan tabel jumlah fitur setiap skema, nilai akurasi dan visualisasi matriks konfusi dari *akurasi* terbaik dengan perbandingan antara data *train* dan data *test* 90:10 dengan interval *parameter gamma adalah* 0.4 hingga 0.9. Pengujian dilakukan pada 7000 *tweet* yang didapatkan dari akun @KAI121 selama Januari 2018 hingga Januari 2020 menggunakan metode *multi-class Support Vector Machine (SVM)* dan pembobotan TF-IDF.

Berikut jumlah fitur yang dihasilkan dari setiap scenario/skema :

Tabel 4-1. Jumlah Fitur Setiap Skema.

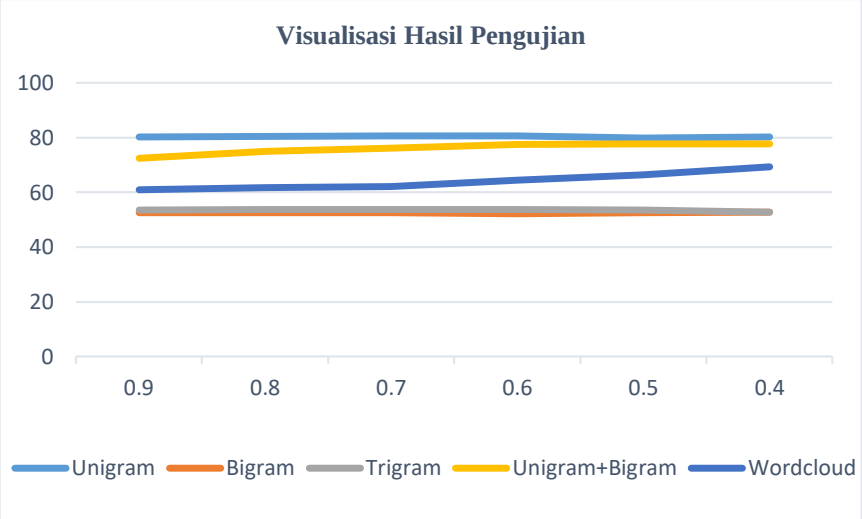
Fitur	Unigram	Bigram	Trigram	Unigram+Bigram	Wordcloud
Jumlah	7130	50655	77565	57758	116

Berdasarkan Tabel 4-1, dapat terlihat bahwa dalam penelitian ini menggunakan lima skenario fitur pembobotan TF-IDF. Setiap skenario memiliki jumlah/total fitur yang berbeda. Wordcloud memiliki jumlah fitur paling sedikit dan trigram memiliki jumlah fitur paling banyak. Setiap fitur dapat mempengaruhi hasil klasifikasi. Berikut nilai akurasi yang didapatkan dari tahap pengujian:

Tabel 4-2. Tabel Akurasi

No	Fitur	Gamma						Rata-Rata
		0.9	0.8	0.7	0.6	0.5	0.4	
1	**Unigram**	**80.31**	**80.45**	**80.59**	**80.59**	**79.88**	**80.17**	**80.33**
2	Bigram	52.6	52.54	52.56	52.12	52.54	52.83	52.53
3	Trigram	53.54	53.82	53.82	53.68	53.54	52.83	53.54
4	Unigram+Bigram	72.37	75.07	76.20	77.62	77.76	77.76	76.13
5	Wordcloud	61.04	61.75	62.18	64.45	66.43	69.26	64.19
Rata-Rata		63.97	64.72	65.07	65.69	66.03	66.57	

Berdasarkan Tabel 4-2, dengan mengubah nilai gamma didapatkan hasil seperti di atas. Hasil akurasi tertinggi berasal dari skema unigram dengan parameter gamma 0.7, akurasi yang didapatkan sebesar 80.59, rata-rata akurasi terbesar didapat pada saat nilai gamma 0.4 yaitu sebesar 66.57, dan rata rata akurasi terkecil didapat pada saat nilai gamma 0.9 yaitu sebesar 63.97. Visualisasi nilai akurasi dapat dilihat pada Gambar 4-1 :



Gambar 4-1 Kurva Nilai Akurasi

Gambar 4-1 diatas menunjukkan visualisasi nilai akurasi lima fitur pembobotan TF-IDF yang dipengaruhi oleh nilai gamma. Parameter nilai gamma pada penelitian ini adalh 0.9, 0.8, 0.7, 0.6, 0.5, dan 0.4. Dari nilai akurasi tertinggi yang didapatkan dapat diketahui matriks konfusinya pada Tabel 4-3:

Tabel 4-3 Matriks Konfusi dari Akurasi Terbaik

Kelas	Terklasifikasi Positif	Terklasifikasi Netral	Terklasifikasi Negatif
Positif	35	24	13
Netral	0	365	37
Negatif	13	37	166

Tabel 4-3 menunjukkan matriks konfusi dari nilai akurasi tertinggi. Dari Tabel 4-3 dapat dilihat 35 data terklasifikasi positif dengan benar, 365 data terklasfikasi netral dengan benar, dan 166 data terklasifikasi negatif dengan benar.

4.2. Analisis Hasil Pengujian

Pengujian model TF-IDF dan metode OAA *multiclass* SVM pada penelitian ini dilakukan dengan skenario perbandingan 90:10 antara data *train* dan data *test*. Hasil dari pengujian tersebut berupa kelas prediksi yang akan dibandingkan dengan kelas aktual pada data *test* dengan menggunakan *confusion matrix*. Analisis hasil pengujian ini bertujuan untuk mengetahui performansi *multiclass SVM* dengan kernel fungsi basis radial serta nilai akurasinya dan mengetahui fitur terbaik dari hasil klasifikasi data menggunakan lima fitur TF-IDF berbeda yang dikombinasikan dengan metode *multi-class Support Vector Machine* serta analisis pengaruh dari penggunaan fitur TF-IDF yang berbeda.

Berdasarkan hasil pengujian dengan menggunakan *confussion matrix* yang telah dilakukan, peneliti dapat menarik beberapa analisis yang dapat dilihat pada Tabel 4.1 yang menampilkan jumlah fitur yang digunakan pada pembobotan TF-IDF. Fitur unigram dengan jumlah fitur 7103 pada tahapan pembobotan fitur sangat sesuai dikombinasikan dengan metode *multiclass* SVM pada penelitian ini, terbukti dari nilai akurasi yang didapat tertinggi diantara fitur yang lain. Dilihat dari jumlah fitur unigram dapat dikatakan fitur unigram paling efisien dan efektif pada penelitian ini. Fitur trigram sangat tidak sesuai, tidak efisien, dan tidak efektif dikombinasikan dengan metode *multi-class SVM* pada dataset penelitian ini, karena dilihat dari jumlah fitur yang paling banyak yaitu 77565 dengan nilai akurasi yang terendah. Penggunaan fitur *wordcloud* pada dataset tidak lebih baik dan tidak efektif dikarenakan banyak kata yang masuk kedalam dua kelas sehingga sulit mengklasifikasikan dengan tepat. Lebih baik jika fitur yang digunakan tidak masuk kedalam dua kelas sekaligus.

Pada Tabel 4.2 yang berisikan hasil akurasi dari fitur unigram, bigram, trigram, unigram+bigram, dan wordcloud. Fitur unigram memiliki nilai rata-rata akurasi tertinggi yaitu 80.33 dibandingkan dengan empat model lainnya yaitu bigram sebesar 52.53 , trigram sebesar 53.54, Unigram+bigram sebesar 76.13, dan wordcloud sebesar 70.33 . Nilai akurasi tertinggi berasal dari pengujian model TF-IDF unigram yang dikombinasikan dengan metode klasifikasi *multi-class Support Vector Machine* (SVM) dengan nilai parameter *gamma 0.7* yaitu 80.59. Gamma yang digunakan dapat mempengaruhi hasil klasifikasi, semakin kecil nilai gamma yang digunakan, hasil akurasinya cenderung naik sesuai dengan visualisasi pada Gambar 4.1 yang menampilkan visualisasi nilai akurasi dengan variable nilai gamma.

Pada Tabel 4.3 menampilkan matriks konfusi dari akurasi terbaik. Akurasi terbaik didapatkan dari pengujian Metode Multiclass *Support Vector Machine* (SVM) dengan parameter gamma 0.7 dan pembobotan TF-IDF Unigram. Berdasarkan Tabel 4.3 dapat dilihat sebanyak 35 data terklasifikasi positif dengan benar, 365 data terklasifikasi netral dengan benar, dan 166 data terklasifikasi negatif dengan benar.

5. Kesimpulan dan Saran

Setelah menerapkan lima pendekatan pembobotan fitur TF-IDF yang berbeda dengan metode *multi-class Support Vector Machine* (SVM) untuk mengklasifikasikan data *tweet* akun PT.KAI, peneliti dapat menarik kesimpulan bahwa :

1. Hasil akurasi tertinggi yang diperoleh dengan menggunakan metode *multiclass* SVM OAA dalam menganalisis sentimen didapat pada rasio 90:10 dengan menggunakan skema unigram, pembobotan TF-IDF dan nilai parameter *gamma* 0.7, yaitu sebesar 80.59 .
2. Fitur penting pada penelitian ini adalah fitur unigram karena mewakili fitur unik dan menghasilkan nilai akurasi yang tinggi.
3. *Gamma* yang digunakan dapat mempengaruhi hasil klasifikasi, semakin kecil nilai gamma yang digunakan, hasil akurasinya cenderung naik.
4. Dari hasil penelitian yang didapatkan, akun @kai121 memiliki 11 % sentimen positif, 58% sentimen netral, dan 31% sentimen negatif. PT.KAI diharapkan untuk meningkatkan pelayanannya kepada pengguna jasa transportasi kereta api dikarenakan angka sentimen positif yang merupakan nilai kepuasan pengguna kereta api masih dibawah rata-rata.

Saran untuk penelitian selanjutnya diantaranya :

1. Disarankan untuk melakukan pelabelan otomatis dengan menggunakan deep learning atau metode lainnya
2. Menggunakan metode pembobotan fitur yang berbeda.

Daftar Pustaka

[1] "• Number of internet users in Indonesia 2023 | Statista." [Online]. Available: https://www.statista.com/statistics/254456/number-of-internet-users-in-indonesia/. [Accessed: 17-Sep-2019].

[2] "Indonesia Digital 2019 : Media Sosial - Websindo." [Online]. Available: https://websindo.com/indonesia-digital-2019-media-sosial/. [Accessed: 17-Sep-2019].

[3] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek," *Proc. - 2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2017*, vol. 2018-Janua, pp. 266–269, 2018.

[4] M. Hejazi, S. A. R. Al-Haddad, Y. P. Singh, S. J. Hashim, and A. F. A. Aziz, "Multiclass Support Vector Machines for Classification of ECG Data with Missing Values," *Appl. Artif. Intell.*, vol. 29, no. 7, pp. 660–674, 2015.

[5] M. L. Pratama, "Studi Komparasi Metode Multiclass Support Vector Machine Untuk Masalah Analisis Sentimen Pada Twitter," *Fmipa Ui*, 2014.

[6] A. Mustakim, I. Santoso, and A. A. Zahra, "Pengenalan Ekspresi Wajah Manusia Menggunakan Tapis Gabor 2-D Dan Support Vector Machine (Svm)," *Transient*, vol. 6, no. 3, p. 232, 2017.

[7] G. A. Dalaorao, A. M. Sison, and R. P. Medina, "Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy," *TSSA 2019 - 13th Int. Conf. Telecommun. Syst. Serv. Appl. Proc.*, pp. 282–285, 2019.

[8] D. De Clercq, Z. Wen, and Q. Song, "Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: A natural language processing approach," *Sci. Total Environ.*, vol. 673, pp. 402–413, 2019.

[9] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 1833–1842, 2014.

[10] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," 2017.

[11] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 3, no. 3, pp. 2789–2797, 2019.

[12] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *Eurasip J. Wirel. Commun. Netw.*, vol. 2017, no. 1, pp. 1–12, 2017.

[13] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and support vector machine for news classification," *Proc. 2nd IEEE Int. Conf. Eng. Technol. ICETECH 2016*, no. March, pp. 112–116, 2016.

[14] H. Wu and N. Yuan, "An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information," *ACM Int. Conf. Proceeding Ser.*, pp. 211–215, 2018.

[15] V. Katsoni, "Cultural tourism in a digital era," *Springer Proc. Bus. Econ.*, vol. 9, pp. 1–12, 2015.

[16] E. Permata, I. K. E. Purnama, and M. H. Purnomo, "Klasifikasi Jenis Dan Fase Parasit Malaria Plasmodium Falciparum Dan Plasmodium Vivax Dalam Sel Darah Merah Menggunakan Support Vector Machine One Against One," *Setrum*, vol. 1, no. 2, pp. 1–8, 2012.

[17] "Digital library - Perpustakaan Pusat Unikom - Knowledge Center - WELCOME | Powered by GDL4.2 | ELIB UNIKOM." [Online]. Available: https://elib.unikom.ac.id/gdl.php?mod=browse&op=read&id=jbptunikompp-gdl-citrawatii-35966&newtheme=gray&newtheme=green. [Accessed: 15-Dec-2019].

[18] "Mengukur Kinerja Algoritma Klasifikasi dengan Confusion Matrix – Achmatim.Net." [Online]. Available: https://achmatim.net/2017/03/19/mengukur-kinerja-algoritma-klasifikasi-dengan-confusion-matrix/. [Accessed: 13-Nov-2019].