Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2024118713
Vol. 11, No. 6, Desember 2024, hlm. 1253-1264 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

ANALISIS SENTIMEN PADA SOSIAL MEDIA TWITTER TERHADAP KUALITAS JARINGAN INTERNET TELKOMSEL MENGGUNAKAN *ENSEMBLE K-NEAREST NEIGHBOUR -SUPPORT VECTOR MACHINE*

Muchammad Farchan Fachrudin[1], Cucun Very Angkoso[2], Doni Abdul Fatah[3]

[1,2,3] Universitas Trunojoyo Madura, Bangkalan
Email: [1]farchanaan345@gmail.ac.id, [2]cucunvery@trunojoyo.ac.id, [3]doni.fatah@trunojoyo.ac.id
*Penulis Korespondensi

(Naskah masuk: 06 Februari 2024, diterima untuk diterbitkan: 20 November 2024)

Abstrak

Di Indonesia, PT Telekomunikasi Seluler, yang mengoperasikan layanan jaringan internet seluler melalui Telkomsel, merupakan salah satu perusahaan penyedia layanan internet. Opini pengguna Telkomsel mengenai kualitas layanan jaringan internet sering dijadikan representasi kepuasan pengguna, yang menjadi indikator penilaian dan evaluasi bagi perusahaan. Analisis sentimen, dengan melakukan klasifikasi opini pengguna ke dalam kelas positif, negatif, atau netral, dapat digunakan sebagai metode untuk mengukur kepuasan pengguna terhadap layanan tersebut. Dalam penelitian analisis sentimen ini menggunakan model algoritma *machine learning* yaitu *K-Nearest Neighbour*, *Support Vector Machine*, dan *Ensemble* KNN-SVM yang berbasis *majority vote* dan berbasis *average*. Dalam penelitian ini data yang diambil berasal dari Twitter dengan rentang waktu 7 Juli 2020 hingga 31 Desember 2022 dengan total jumlah data sebesar 30004 data dan diambil sampel yang diberi label sebesar 3900 data. Dari penggunaan data sampel tersebut, nilai akurasi pada model KNN pada K=15 memberikan hasil akurasi sebesar 83.21%, model SVM pada C=100 memberikan hasil akurasi sebesar 84.33%, model *Ensemble* KNN-SVM *Majority Vote* atau *Hard Vote* memberikan hasil akurasi sebesar 83.26%, dan model *Ensemble* KNN-SVM Average atau *Soft Vote* memberikan hasil akurasi sebesar 84.79%. Selain itu keempat model tersebut melakukan prediksi sentimen terhadap data yang belum dilabel dan keempat model tersebut memprediksi mayoritas sentimennya yaitu negatif. Sehingga dapat disimpulkan bahwa opini masyarakat terhadap kualitas layanan jaringan internet telkomsel adalah negatif. Secara keseluruhan, penggunaan model klasifikasi KNN, SVM, dan *Ensemble* KNN-SVM dalam melakukan analisis sentimen dapat dikatakan baik dan mampu untuk memprediksi sentimen pada sebuah data yang belum berlabel dan yang berlabel.

Kata kunci: *K-Nearest Neighbour, Support Vector Machine, Analisis Sentimen, Ensemble Learning, Jaringan Internet, Telkomsel*

SENTIMENT ANALYSIS ON TWITTER SOCIAL MEDIA ON TELKOMSEL'S INTERNET NETWORK QUALITY USING ENSEMBLE K-NEAREST NEIGHBOR - SUPPORT VECTOR MACHINE

Abstract

In Indonesia, PT Telekomunikasi Cellular, which operates cellular internet network services through Telkomsel, is one of the internet service provider companies. Telkomsel users' opinions regarding the quality of internet network services are often used as a representation of user satisfaction, which is an indicator of assessment and evaluation for the company. Sentiment analysis, by classifying user opinions into positive, negative, or neutral classes, can be used as a method to measure user satisfaction with the service. This sentiment analysis research uses machine learning algorithm models, namely K-Nearest Neighbor, Support Vector Machine, and KNN-SVM Ensemble which are majority vote-based and average-based. In this study, the data taken came from Twitter with a period of July 7, 2020, to December 31, 2022, with a total amount of data of 30004 data and a labeled sample of 3900 data was taken. From the use of the sample data, the accuracy value of the KNN model at K = 15 gave an accuracy result of 83.21%, the SVM model at C = 100 gave an accuracy result of 84.33%, the KNN-SVM Majority Vote or Hard Vote Ensemble model gave an accuracy result of 83.26%, and the KNN-SVM Average or Soft Vote Ensemble model gave an accuracy result of 84.79%. In addition, the four models predict sentiment against unlabeled data and all four models predict the majority of sentiment is negative. So it can be concluded that public opinion on the quality of Telkomsel's internet network services is negative. Overall, the use

of KNN, SVM, and Ensemble KNN-SVM classification models in conducting sentiment analysis can be said to be good and able to predict sentiment on unlabeled and labeled data.

Keywords: *K-Nearest Neighbor, Support Vector Machine, Sentiment Analysis, Ensemble Learning, Internet Network, Telkomsel*

1. PENDAHULUAN

Dalam Kamus Besar Bahasa Indonesia, teknologi adalah sarana dalam mempersiapkan suatu barang yang dibutuhkan untuk keberlangsungan hidup manusia (BADAN PENGEMBANGAN DAN PEMBINAAN BAHASA & KEMENDIKBUDRISTEK, 2016). Teknologi informasi merupakan salah satu cabang dari teknologi. Pada zaman teknologi informasi masa kini, manusia menggunakan media internet untuk berbagai tujuan. Salah satu negara yang memiliki jumlah pengguna internet aktif cukup banyak di dunia adalah Indonesia yaitu sebanyak 212.9 juta pengguna atau sekitar 77% dari jumlah total populasi di Indonesia pada tahun 2023 (RIZATY, 2023). Para pengguna tersebut menggunakan internet karena internet adalah media yang mudah untuk mendapatkan sebuah informasi dikarenakan perkembangannya bersamaan dengan teknologi yang juga semakin berkembang. Terdapat berbagai alasan utama para pengguna menggunakan internet, beberapa diantaranya yaitu untuk mencari sebuah informasi, berita terkini, dan membagikan opini atau pendapat kita.

Para pengguna dapat menyampaikan pendapat atau opini mereka terhadap suatu hal yang dia rasakan dan yang terjadi di masyarakat. Dengan semakin banyaknya pengguna media sosial dan internet browser, dapat memberikan dampak positif maupun negatif pada informasi serta opini publik. Para pengguna akan mendapatkan informasi maupun berita terkini yang mereka butuhkan dengan serba cepat dan tepat. Untuk mencari beberapa hal tersebut para pengguna dapat menggunakan berbagai macam platform media sosial seperti Instagram, Tiktok, Facebook, Youtube, Whatsapp, dan salah satunya yaitu Twitter.

Di Indonesia, ada beberapa perusahaan penyedia layanan internet, salah satunya adalah PT Telekomunikasi Seluler yang mengoperasikan layanan jaringan internet seluler melalui Telkomsel. Dikarenakan banyak yang menggunakan operator seluler Telkomsel maka umumnya pengguna dari operator tersebut akan menyampaikan opini mereka terhadap kualitas layanan jaringan internet yang diberikan. Opini tersebut dapat merepresentasikan tentang kepuasan pengguna pada layanan jasa tersebut. Kepuasan pengguna seringkali digunakan oleh sebuah perusahaan sebagai bahan penilaian dan evaluasi dari layanan mereka. Untuk mengukur kepuasan pengguna tersebut, salah satu cara yang dapat diterapkan yaitu analisis sentimen dengan cara melakukan klasifikasi terhadap opini pengguna dan mengelompokkan kedalam kelas positif, negatif, atau netral.

Penelitian ini menggunakan analisis sentimen yang bertujuan untuk memberikan wawasan yang lebih dalam mengenai kualitas layanan jaringan internet Telkomsel, dikarenakan Telkomsel merupakan operator seluler yang memiliki persentase pengguna terbanyak yaitu sebanyak 41.94% untuk mengakses internet melalui smartphone pada tahun 2021-2022, persentase tersebut didapatkan dari survei yang dilaksanakan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) (DIHNI, 2022). Sehingga dengan adanya informasi ini dapat membantu pihak Telkomsel untuk memberikan kritik dan saran yang membangun untuk dapat meningkatkan atau mempertahankan persentase pengguna terbanyak tersebut.

Analisis sentimen ini dilakukan untuk mengetahui tentang opini pribadi masyarakat tersebut lebih cenderung kepada opini yang positif atau opini yang negatif atau opini yang netral (Salsabila, Sihombing and Sitorus, 2022). Sehingga dengan adanya analisis sentimen ini dapat membantu perusahaan untuk melihat bagaimana respon atau penilaian dari pengguna atas barang atau layanan mereka, maka dari itu perusahaan dapat menciptakan rencana-rencana yang aktual untuk masa yang akan datang (Ananda and Pristyanto, 2021).

Algoritma *K-Nearest Neighbour* (KNN) merupakan sebuah algoritma yang termasuk dalam algoritma *supervised learning* karena KNN memanfaatkan data latih yang telah diberikan label untuk melakukan sebuah prediksi pada data uji. Algoritma KNN ini menggunakan prinsip kerja yang mendapatkan jarak paling dekat dengan data yang diuji dengan K tetangga (*neighbour*) terdekatnya pada data latih. Algoritma KNN ini ditemukan oleh Fix dan Hodges pada tahun 1951. Pada penelitian lain yang menggunakan algoritma KNN terkait analisis sentimen, data yang digunakan pada penelitian ini yaitu mengenai sentimen masyarakat terhadap larangan mudik 2021 pada twitter dengan jumlah data sebanyak 4.799 data dengan nilai k = 3 menghasilkan nilai akurasi sebesar 86.67% (LESTARI & MAHDIANA, 2021).

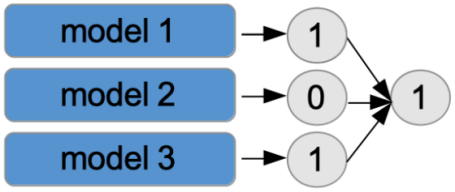
Support Vector Machine (SVM) adalah salah satu algoritma terbaru yang dapat digunakan untuk klasifikasi data baik linear maupun non-linear. Algoritma ini adalah salah satu algoritma supervised learning yang berarti algoritma ini dapat mengetahui suatu data dari model label tertentu yang ada pada dataset yang digunakan. SVM ini banyak diimplementasikan dalam penelitian dalam studi

klasifikasi teks, salah satu kelebihan dari SVM ini yaitu mampu menghasilkan nilai yang lebih baik pada sebuah teks yang berisi opini (SALSABILA et al., 2022). Pada penelitian terkait analisis sentimen, menggunakan data yang berasal dari twitter mengenai sentimen masyarakat Indonesia terkait vaksin covid-19 pada media sosial twitter, dengan jumlah data yang digunakan sebanyak 5000 data, dengan menggunakan kernel *linear*, *RBF*, *sigmoid*, dan *polynomial*, dari empat kernel tersebut menghasilkan nilai akurasi pada kernel *linear* sebesar 88.3%, pada kernel *RBF* menghasilkan nilai akurasi sebesar 88.8%, pada kernel *sigmoid* menghasilkan nilai akurasi sebesar 87%, pada kernel *polynomial* menghasilkan nilai akurasi sebesar 85.5% (ARFAT et al., 2022).

Pada penelitian yang lain yang juga menggunakan data yang berasal dari media sosial twitter yaitu mengenai vaksin *Sinovac*, pada penelitian ini menggunakan algoritma SVM dan KNN dengan menggunakan 2105 data yang menggunakan bahasa inggris, kernel yang digunakan pada algoritma SVM yaitu kernel *Linear*, kernel *Polynomial*, dan kernel RBF, sedangkan algoritma KNN menggunakan jumlah n sebanyak 3, 5, dan 7. Hasil pengujian dari masing masing algoritma tersebut dievaluasi dengan memakai *K-Fold Cross Validation* dengan jumlah *Fold* sebesar 10-*fold*. Didapatkan nilai akurasi untuk algoritma SVM untuk kernel *Linear* dengan nilai 0.70, untuk kernel *Polynomial* dengan nilai 0.57, dan untuk kernel RBF dengan nilai 0.66, sedangkan pada algoritma KNN pada jumlah *n* 3 menghasilkan nilai akurasi sebesar 0.55, pada jumlah *n* 5 menghasilkan nilai akurasi sebesar 0.55, dan pada jumlah *n* 7 menghasilkan nilai akurasi 0.56. Sehingga dapat disimpulkan bahwa algoritma SVM menghasilkan nilai akurasi terbaik sebesar 0.7 sedangkan *KNN* menghasilkan nilai akurasi sebesar 0.56 (BAITA et al., 2021).

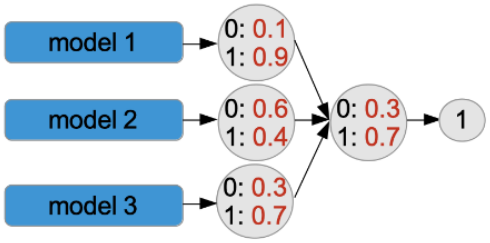
Peneliti menambahkan metode *ensemble* untuk menggabungkan kedua algoritma tersebut. Metode *ensemble* merupakan sebuah metode pembelajaran mesin yang menggabungkan dari beberapa model untuk menghasilkan sebuah solusi untuk memecahkan masalah tertentu dan akan memberikan hasil yang lebih baik daripada hanya menggunakan satu model saja (DOGAN & BIRANT, 2019; HUANG, 2023). Metode ensemble adalah metode yang menggabungkan beberapa classifier tunggal untuk membentuk classifier baru yang dapat meningkatkan akurasi (INDRIATI & KUSYANTI, 2019). Metode *ensemble* memiliki beberapa macam contohnya seperti *ensemble hard vote* atau *Majority Vote* dan *ensemble soft vote* atau *average*.

Ensemble majority voting atau *ensemble* mayoritas merupakan teknik pembelajaran dasar *ensemble* yang menggunakan voting dari beberapa model klasifikasi *machine learning*. *Ensemble majority voting* mengkombinasikan model classifier dengan cara *hard voting* atau *majority voting* akan memprediksi hasil kelas dengan menggunakan suara mayoritas yang diberikan oleh masing-masing model *classifier*. Dalam proses *majority voting*, hasil voting terbanyak dari model klasifikasi tersebut akan dijadikan sebagai hasil akhir label kelas, jika tidak tersedia hasil akhir label kelas yang memenuhi maka *ensemble learning* akan memberikan pilihan *rejection* atau tidak menghasilkan prediksi label (KHADIJAH & KUSUMANINGRUM, 2019). Proses *Ensemble Majority Vote* ditampilkan pada Gambar 1.



Gambar 1. Proses *Ensemble Majority Vote*

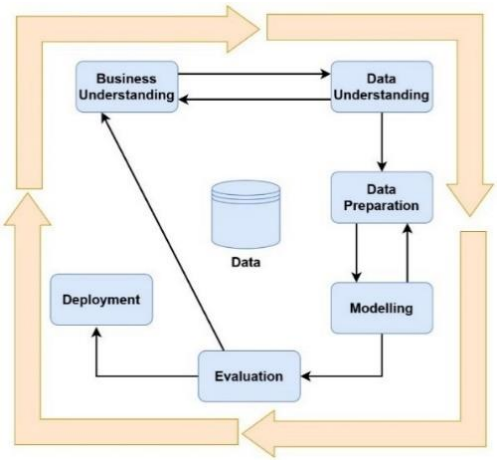
Sedangkan *Ensemble soft-voting* atau *average* merupakan sebuah metode untuk mengkombinasikan hasil dari beberapa model klasifikasi yang berbeda. *Ensemble average* ini menghasilkan sebuah keputusan kelas atau label dengan cara menghitung rata-rata probabilitas dari setiap kelas yang dibuat oleh setiap model klasifikasi dan akan hasilnya dipilih berdasarkan nilai rata-rata probabilitas tertinggi (MANCONI et al., 2022). Proses *Ensemble Average* ditampilkan pada Gambar 2.



Gambar 2. Proses *Ensemble Average*

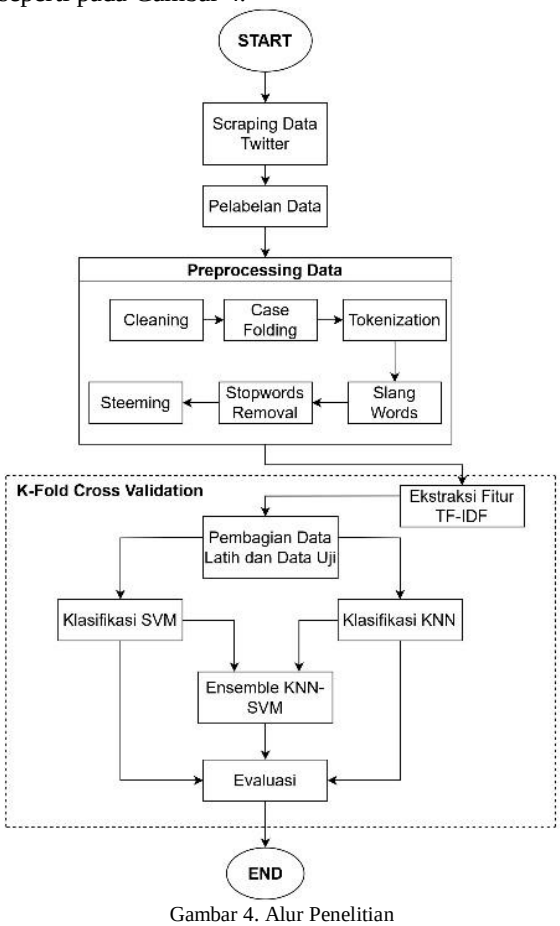
2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini yaitu metode *Cross Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM merupakan sebuah pedoman untuk membangun model data mining. Model CRISP-DM ini dibuat pada tahun 1996 oleh analis yang mewakili DaimlerChrysler, SPSS, dan NCR (UMAIDAH & PURWANTORO, 2019). Model CRISP-DM ini merupakan model yang sering digunakan karena efektif dan langkah-langkah penerapannya yang mudah untuk diaplikasikan (ATMAJA & YUSTANTI, 2021). Adapun langkah-langkah penerapan CRISP-DM pada Gambar 3.



Gambar 3. Tahapan CRISP-DM

Untuk alur penelitian yang dilakukan adalah seperti pada Gambar 4.



Gambar 4. Alur Penelitian

Tahap pertama yang dilakukan untuk penelitian ini yaitu *scraping data twitter* dengan menggunakan *library snscrape*, tahap kedua dilakukan pelabelan secara manual oleh peneliti yang kemudian dilakukan validasi oleh dosen Ahli Bahasa dan Sastra Indonesia, tahap ketiga yaitu proses *cleaning*, tahap keempat dilakukan proses *preprocessing data* yang terdiri dari *cleaning*, *case folding*, *tokenization*, *slangwords*, *stopwords removal*, dan *steeming*, tahap kelimat dilakukan ekstraksi fitur atau pembobotan kata dengan menggunakan TF-IDF, tahap keenam dilakukan pembagian data latih dan data uji dengan perbandingan 90% untuk data latih dan 10% untuk data uji yang diimplementasikan kedalam *k-fold cross validation*, tahap ketujuh dilakukan proses klasifikasi dengan menggunakan algoritma KNN yang menghasilkan hasil prediksi dan nilai metrik evaluasi, tahap kedelapan dilakukan proses klasifikasi dengan menggunakan algoritma SVM yang menghasilkan hasil prediksi dan nilai metrik evaluasi, tahap kesembilan setelah mendapatkan hasil prediksi KNN dan SVM tersebut digunakan sebagai data untuk algoritma *Ensemble Majority Vote* dan *Ensemble Average*. *Ensemble Majority Vote* akan menggunakan nilai prediksi mayoritas yang telah dihasilkan oleh setiap model dan *Ensemble Average* menggunakan nilai rata-rata dari probabilitas untuk dijadikan sebagai hasil prediksi dan *Ensemble Majority Vote* serta *Ensemble Average* menghasilkan prediksi kelas dan nilai evaluasi.

2.1 Business Understanding

Pada *Business Understanding* merupakan tahapan yang bertujuan untuk mendapatkan dan memahami tujuan bisnis dari proyek, Pada tahapan ini, topik yang digunakan yaitu analisis sentimen twitter terhadap opini publik mengenai kualitas layanan jaringan internet Telkomsel, tujuan dari digunakannya topik ini yaitu untuk memperoleh wawasan tentang opini publik terhadap kualitas layanan jaringan internet Telkomsel secara keseluruhan (paket data yang digunakan hanya di Iindonesia), sehingga dengan didapatkannya wawasan tentang opini publik tersebut dapat membantu pihak Telkomsel untuk merancang strategi baru yang lebih efektif dalam meningkatkan layanan dan memuaskan pelanggan. Sehingga Telkomsel dapat melakukan evaluasi terhadap pasar yang telah mereka miliki dan melakukan tindakan yang tepat untuk meningkatkan kualitas layanan mereka.

2.2 Data Understanding

Pada tahap *Data Understanding* merupakan tahap pemahaman data terhadap data yang digunakan dalam penelitian. Data yang digunakan mencakup opini publik mengenai kualitas layanan jaringan internet Telkomsel pada media sosial Twitter. Pada tahap ini dilakukan pengambilan data dengan rentang waktu 7 Juli 2020 hingga 31 Desember 2022 dengan menggunakan dua kata kunci yaitu jaringan telkomsel dan sinyal telkomsel yang masing-masing kata kunci memperoleh data berjumlah 15,002 data. Proses pengambilan data ini menggunakan bahasa pemrograman Python dengan menggunakan *library snscrape*. Data yang diambil hanya yang berbahasa Indonesia. Pada tahap ini juga dilakukan proses *filtering* atau penghapusan data yang berasal dari akun milik lembaga negara, organisasi, situs berita, dan tweet yang terduplikat data tweet yang menggunakan selain Bahasa Indonesia akan dihapus.

Tabel 1 merupakan contoh sampel data opini publik tersebut.

Tabel 1. Sampel Data Opini Publik

No.	Tweet
1.	b'@hvtamaa Jaringan Telkomsel elek ngab'
2.	b'sinyal telkomsel pas ujan https://t.co/bl5rqNEOuG'
3.	b'Pas di RS lancar jaya nih sinyal Telkomsel \xf0\x9f\x98\x91'

2.3 Data Preparation

Pada tahap *Data Preparation* terdiri dari beberapa tahapan yaitu pengambilan sampel, pelabelan data, *preprocessing data*, dan teknik TF-IDF. Tahap pengambilan sampel menggunakan *sampel slovin.* Dari hasil *sampel slovin* tersebut didapatkan untuk pengambilan sampel dengan jumlah minimal data tweet yang digunakan sebagai data pelatihan model. Kemudian untuk pelabelan data dilakukan secara manual oleh peneliti dengan mengamati secara satu per satu dan hasil dari pelabelan ini akan divalidasi oleh dosen Ahli Bahasa dan Sastra Indonesia. adalah hasil pelabelan manual yaitu positif, netral, dan negatif.

Tabel 2. Contoh Sampel Sentimen

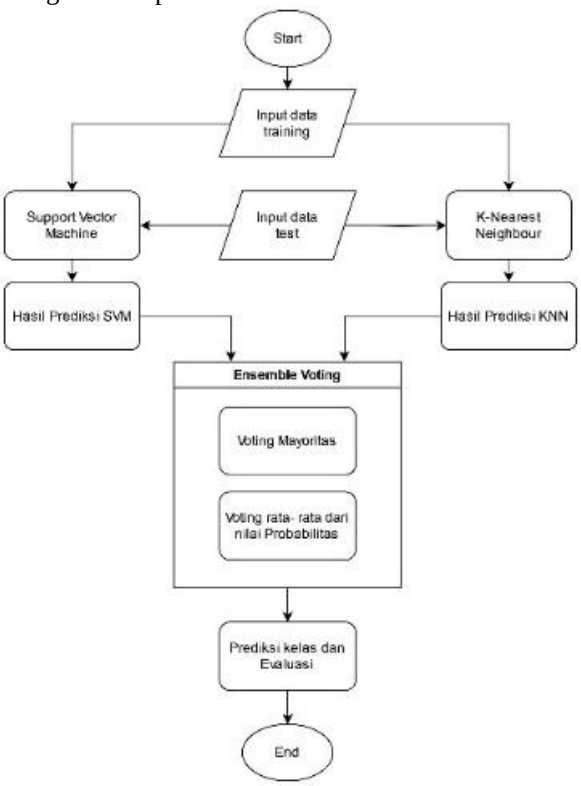
Tweet	Sentimen
b'@hvtamaa Jaringan Telkomsel elek ngab'	Negatif
b'@Telkomsel Semoga di G20 lancar jaringan telkomsel.'	Netral
b'Gilee jaringan telkomsel paten kali ahh'	Positif

Setelah dilakukan pelabelan, tahap selanjutnya yaitu *preprocessing data*. Tahap ini terdiri dari: a) *cleaning* adalah menghilangkan karakter atau tanda baca yang tidak dibutuhkan seperti menghilangkan username (@), menghilangkan hyperlink, menghilangkan tanda baca, menghilangkan karakter *newline*, menghilangkan angka, menghilangkan *emoticon*, dan menghilangkan hastag (#), b) *case folding* adalah mengubah karakter huruf besar menjadi huruf kecil pada suatu dokumen, c) *tokenization* merupakan proses untuk memecah sebuah kalimat atau teks menjadi beberapa bagian kecil yang dikenal sebagai token, d) *slangwords* merupakan memperbaiki dan mengubah kata-kata gaul dan kata tidak baku menjadi kata baku sesuai dengan ejaan Bahasa Indonesia, e) *stopwords removal* merupakan tahap dimana kata-kata yang seringkali muncul di dalam sebuah teks namun cenderung tidak memiliki arti terhadap konteks teks tersebut akan dihapus. Beberapa contoh kata *stopwords removal* Indonesia yaitu "ini", "apa", "terus", "yang", "juga", "harus" dan masih banyak lagi, dan f) *steeming* adalah proses untuk mengubah kata dalam suatu teks menjadi kata dasar dengan menghilangkan imbuhan awalan dan akhiran.

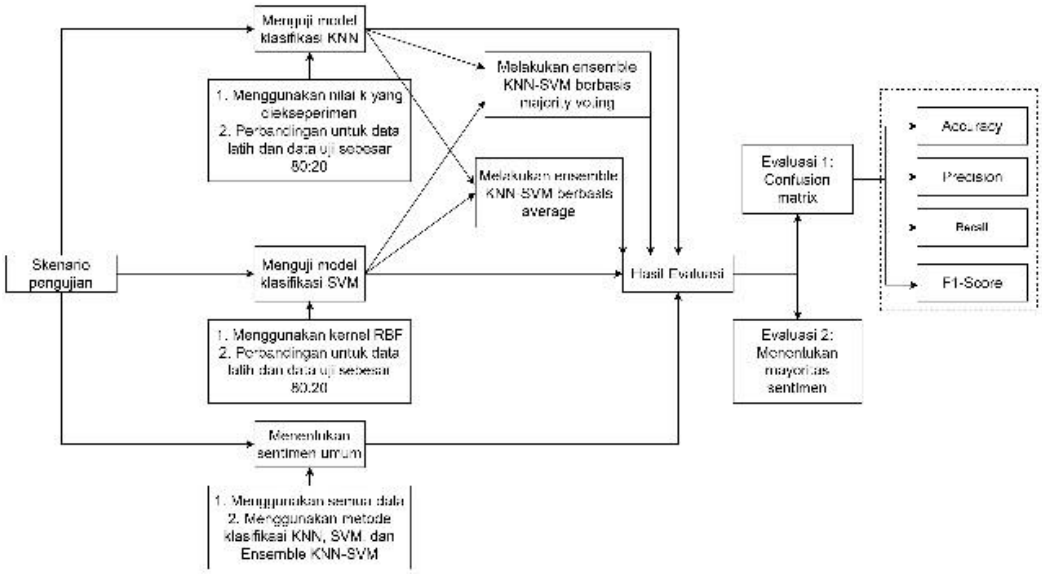
Setelah melalui tahap *preprocessing data*, tahap selanjutnya yaitu pembobotan kata dengan menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode pembobotan TF-IDF ini digunakan untuk memberikan perhitungan bobot setiap kata yang memiliki frekuensi paling sering muncul dalam sebuah dokumen.

2.4 Modelling

Pada tahap *modelling* dilakukan proses pengujian klasifikasi dengan menggunakan algoritma KNN, kernel RBF, dan *Ensemble* KNN-SVM berbasis *majority vote* dan *average*. Pada tahap ini dilakukan pembagian data dengan perbandingan 90:10 dengan menggunakan *k-fold cross validation* dengan nilai k=10. Seperti pada Gambar 5 terdapat pembagian data latih dan data uji untuk algoritma klasifikasi KNN dan SVM, kemudian algoritma KNN dan SVM menghasilkan prediksi KNN dan SVM. Dari hasil prediksi KNN dan SVM tersebut digunakan sebagai data untuk algoritma *Ensemble Majority Vote* dan *Ensemble Average*. *Ensemble Majority Vote* akan menggunakan nilai prediksi mayoritas yang telah dihasilkan oleh setiap model dan *Ensemble Average* menggunakan nilai rata-rata dari probabilitas untuk dijadikan sebagai hasil prediksi dan *Ensemble Majority Vote* serta *Ensemble Average* menghasilkan prediksi kelas dan nilai evaluasi.



Gambar 5. Pemodelan Klasifikasi



Gambar 6. Skenario Pengujian

2.5 Evaluation

Pada tahap *evaluation*, melakukan evaluasi terhadap hasil penilaian kinerja algoritma algoritma KNN, SVM kernel RBF, dan *Ensemble* KNN-SVM berbasis *majority vote* dan *average* pada data yang telah dilakukan pada tahap *modelling*. Tahap evaluasi ini dilakukan penilaian performa dari masing-masing algoritma klasifikasi tersebut dengan menggunakan *confusion matrix* yang terdiri dari nilai *accuracy*, *precision*, *recall*, F1-Score serta mengetahui hasil mayoritas sentimen yang muncul dari algoritma KNN, SVM kernel RBF, dan *Ensemble* KNN-SVM berbasis *majority vote* dan *average* seperti pada Gambar 6.

2.6 Deployment

Pada tahap *deployment*, tahap ini tujuannya yaitu digunakan untuk pemberian saran terhadap perusahaan Telkom yang memiliki layanan jaringan internet Telkomsel dengan cara melihat kata yang sering muncul dari data yang labelnya diprediksi oleh ketiga model tersebut, melihat hasil evaluasi untuk model KNN, SVM kernel RBF, dan *Ensemble* KNN-SVM berbasis *majority vote* dan *average*, dan pembuatan aplikasi untuk memprediksi kalimat dengan menggunakan model yang memiliki akurasi terbaik.

3. HASIL DAN PEMBAHASAN

3.1 Business Understanding

Pada tahap *business understanding* ini, peneliti memahami tujuan bisnis dari penelitian ini. Tujuan bisnis dari penelitian ini yaitu untuk mengetahui kualitas layanan jaringan internet Telkomsel secara keseluruhan (paket data yang digunakan hanya di Indonesia), sehingga dengan didapatkannya wawasan tentang opini publik tersebut dapat membantu pihak Telkomsel untuk merancang strategi baru yang lebih efektif dalam meningkatkan layanan dan memuaskan pelanggan.

3.2 Data Understanding

Dalam tahap *data understanding* ini, peneliti memahami data yang akan digunakan dalam penelitian. Data yang digunakan mencakup opini publik mengenai kualitas layanan jaringan internet Telkomsel pada media sosial Twitter, Pada tahap ini dilakukan pengambilan data dengan rentang waktu 7 Juli 2020 hingga 31 Desember 2022 dengan menggunakan dua kata kunci yaitu jaringan telkomsel dan sinyal telkomsel yang masing-masing kata kunci memperoleh data berjumlah 15,002 data. Proses pengambilan data ini menggunakan bahasa pemrograman Python dengan menggunakan *library snscrape*. Setelah memperoleh data tersebut, dilakukan *filtering data* atau penghapusan data yang berasal dari akun milik lembaga negara, organisasi, situs berita, dan tweet yang terduplikat selain itu data tweet yang menggunakan selain Bahasa Indonesia akan dihapus. Pada tahap ini data yang awalnya berjumlah 30,004 data setelah dilakukan proses *filtering* berubah menjadi 26,295 data.

3.3 Data Preparation

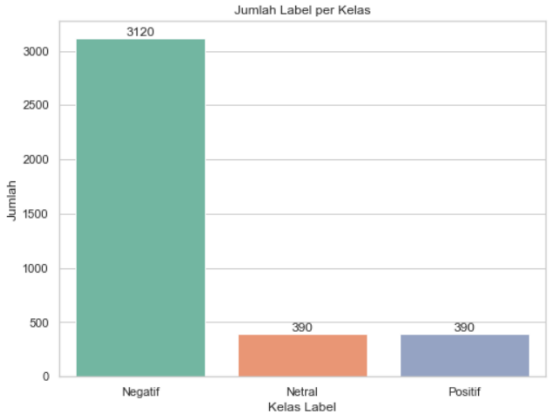
Setelah proses *filtering*, tahap selanjutnya yaitu *data preparation* yang terdiri dari pengambilan sampel, pelabelan data, tahap *preprocessing* data, dan pembobotan kata yang menggunakan TF-IDF. Pada pengambilan sampel, digunakan rumus *sampel slovin* untuk pengambilan data sampelnya, untuk pengimpelentasian rumus *sampel slovin* dapat menggunakan persamaan (1)

$$n = \frac{N}{1 + N \times e^2} \quad (1)$$

Keterangan:
N = ukuran populasi
n = ukuran sampel
e = batas toleransi kesalahan

$$n = \frac{26{,}295}{1 + 26{,}295 \times 2\%^2} = 2283$$

Data yang berjumlah 26,295 akan digunakan sebagai data yang digunakan. Data tersebut dihitung menggunakan rumus *sampel slovin* dengan batas toleransi kesalahan sebesar 2%. Didapatkan hasil berjumlah 2283, jumlah ini dijadikan sebagai nilai minimal data yang digunakan untuk pelatihan model. Peneliti menggunakan data berjumlah 3900 data. Kemudian peneliti melabeli secara manual yang kemudian divalidasi oleh dosen Ahli Bahasa dan Sastra Indonesia. Peneliti menggunakan tiga kelas sentiment yaitu positif, negatif, dan netral. Berdasarkan hasil pelabelan pada data sampel didapatkan jumlah kelas sentimen positif sebesar 3120 data, kelas sentiment negatif sebesar 390 data, dan kelas sentimen netral sebesar 390 data. Adapun jumlah sentiment yang dihasilkan seperti pada Gambar 7.



Gambar 7. Hasil Sentimen Data Sampel

Setelah melakukan pelabelan, tahap selanjutnya yaitu *preprocessing data*, tahap ini merupakan tahap untuk membersihkan data dan mempersiapkan data mentah menjadi data yang siap untuk diolah oleh algoritma model klasifikasi. Hasil dari setiap proses *preprocessing* akan mempengaruhi hasil selanjutnya. Untuk tahapan *preprocessing* seperti berikut ini:

1. *Cleaning*: Proses untuk menghilangkan karakter atau tanda baca yang tidak dibutuhkan seperti menghilangkan username (@), menghilangkan hyperlink, menghilangkan tanda baca, menghilangkan karakter *newline*, menghilangkan angka, menghilangkan *emoticon*, dan menghilangkan hastag (#).
2. *Case Folding*: Proses untuk mengubah karakter huruf besar menjadi huruf kecil pada suatu dokumen.
3. *Tokenization*: Proses untuk memecah sebuah kalimat atau teks menjadi beberapa bagian kecil yang dikenal sebagai token.
4. *Slangwords*: Proses memperbaiki dan mengubah kata-kata gaul dan kata tidak baku menjadi kata baku sesuai dengan ejaan Bahasa Indonesia.
5. *Stopwords Removal*: Proses penghapusan kata-kata yang seringkali muncul di dalam sebuah teks namun cenderung tidak memiliki arti terhadap konteks teks tersebut.
6. *Steeming*: Proses untuk mengubah kata dalam suatu teks menjadi kata dasar dengan menghilangkan imbuhan awalan dan akhiran.

Setelah tahap *preprocessing data*, tahap selanjutnya yaitu pembobotan kata menggunakan TF-IDF. Pembobotan kata ini diambil dari kolom *steeming* yang akan diubah menjadi vektor numerik. Jumlah vektor numerik ini dihitung berdasarkan frekuensi seringnya kemunculan sebuah kata pada suatu dokumen atau kalimat. Pada penelitian ini, tahap TF-IDF dibagi menjadi dua bagian yaitu untuk data yang telah memiliki label dan data yang tidak memiliki label. Tahap ekstraksi fitur TF-IDF pada penelitian ini menggunakan *library* dari *scikit learn*.

3.4 Modelling

Pada tahap *modelling* ini digunakan untuk melakukan pemodelan dengan menggunakan model klasifikasi yang digunakan yaitu KNN, SVM kernel RBF, dan Ensemble KNN-SVM berbasis voting mayoritas dan *average* dilakukan *splitting* data dengan pembagian 90% untuk data *training* dan 10% untuk data *testing* yang diimplementasikan pada *K-Fold Cross Validation* dengan menggunakan K=10. Tujuan dari pemodelan ini yaitu untuk membuat model yang memiliki kemampuan bekerja dengan menggunakan data yang tersedia.

Pada algoritma KNN menggunakan beberapa nilai k yang dieksperimen yaitu 3, 5, 7, 9, 11, 13, 15, 17, 19, dan 21. Sedangkan pada algoritma SVM kernel RBF menggunakan beberapa nilai C yang dieksperimen yatu 0.1, 1, 10, dan 100. Kemudian hasil prediksi algoritma KNN dan SVM kernel RBF digunakan untuk *Ensemble* KNN-SVM yang berbasis *majority vote* menggunakan hasil mayoritas prediksi dari kedua algoritma individual tersebut, sedangkan untuk *Ensemble* KNN-SVM berbasis *average* menggunakan nilai rata-rata probabilitas dari algoritma KNN dan SVM tersebut.

3.5 Evaluation

Pada tahap *evaluation*, melakukan evaluasi terhadap hasil penilaian kinerja algoritma algoritma KNN, SVM kernel RBF, dan *Ensemble* KNN-SVM berbasis *majority vote* dan *average* pada data yang telah dilakukan pada tahap *modelling*. Tahap evaluasi ini dilakukan penilaian performa dari masing-masing algoritma klasifikasi tersebut dengan menggunakan

confusion matrix yang terdiri dari nilai *accuracy*, *precision*, *recall*, dan *F1-Score* yang diambil rata-rata setelah dilakukan *cross-validation* sebanyak 10-*fold*.

Berikut merupakan *output* rata-rata algoritma KNN seperti pada Tabel *3*.

Tabel 3 Hasil Metrik Evaluasi KNN

Metrik	Akurasi	Presisi	Recall	F1-Score
K=3	79.59%	76.16%	79.59%	77.28%
K=5	81.08%	77.41%	81.08%	78.02%
K=7	82.77%	79.63%	82.77%	79.38%
K=9	83.05%	79.91%	83.05%	79.37%
K=11	82.82%	79.33%	82.82%	78.84%
K=13	83.08%	80.13%	83.08%	79.02%
K=15	83.21%	81.43%	83.21%	78.84%
K=17	83.13%	80.25%	83.13%	78.40%
K=19	82.97%	80.06%	82.97%	77.93%
K=21	82.87%	79.52%	82.87%	77.72%

Berikut merupakan *output* rata-rata algoritma SVM kernel RBF seperti pada Tabel 4.

Tabel 4. Hasil Metrik Evaluasi SVM Kernel RBF

Metrik	Akurasi	Presisi	Recall	F1-Score
C=0.1	80.23%	69.75%	80.46%	71.65%
C=1	84.21%	82.49%	84.21%	79.58%
C=10	84.31%	81.53%	84.31%	81.71%
C=100	84.33%	81.69%	84.33%	81.88%

Berikut merupakan *output* rata-rata algoritma *Ensemble* KNN-SVM berbasis *majority vote* atau *hard vote* seperti pada Tabel *5*.

Tabel 5. Hasil Metrik Evaluasi Ensemble Majority Vote

Metrik	Akurasi	Presisi	Recall	F1-Score
Rata-Rata Ensemble Majority Vote	83.26%	81.99%	83.26%	78.25%

Berikut merupakan *output* rata-rata algoritma *Ensemble* KNN-SVM berbasis *average* atau *soft vote* seperti pada Tabel 6.

Tabel 6. Hasil Metrik Evaluasi Ensemble Average

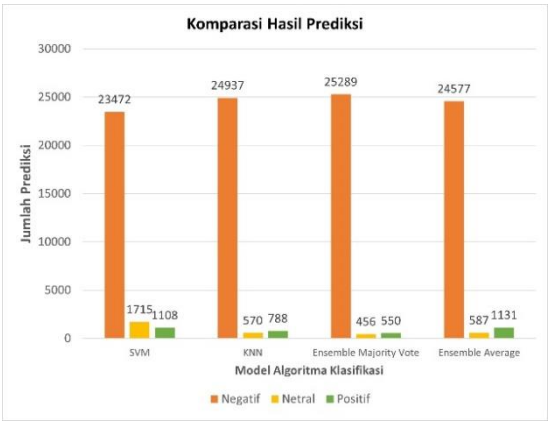
Metrik	Akurasi	Presisi	Recall	F1-Score
Rata-Rata Ensemble Average	84.79%	82.90%	84.79%	81.02%

Overfitting dalam kasus ini terjadi karena model terlalu fokus pada kelas mayoritas ("Negatif"), sehingga performanya menurun drastis pada kelas minoritas ("Netral" dan "Positif") seperti pada Tabel 7.

Tabel 7. Hasil Classification Report Model Ensemble Average

	Precision	Recall	F1-Score
Negatif	0.86	0.98	0.92
Netral	0.62	0.09	0.15
Positif	0.78	0.53	0.63

Mayoritas sentimen yang dihasilkan oleh keempat model tersebut dengan menggunakan data secara keseluruhan adalah negatif dapat dilihat pada Gambar 8.



Gambar 8. Hasil Prediksi Sentimen Oleh Semua Model

Model SVM kernel RBF dengan nilai C=100 menghasilkan prediksi dengan jumlah kelas negatif berjumlah 23472 data, kelas positif berjumlah 1108 data, dan kelas netral berjumlah 1715 data. Model KNN dengan nilai k=15 menghasilkan prediksi dengan jumlah kelas negatif berjumlah 24937 data, kelas positif berjumlah 788 data, dan kelas netral berjumlah 570 data. Model *Ensemble* KNN-SVM *majority vote* menghasilkan prediksi dengan jumlah kelas negatif berjumlah 25289 data, kelas positif berjumlah 550 data, dan kelas netral berjumlah 456 data. Model *Ensemble* KNN-SVM *average* menghasilkan prediksi dengan jumlah kelas negatif berjumlah 24577 data, kelas positif berjumlah 1131, dan kelas netral berjumlah 587. Sehingga dari keempat model tersebut memberikan prediksi sentimen dengan menggunakan data secara keseluruhan terhadap opini publik mengenai kualitas layanan jaringan internet Telkomsel adalah mayoritas bersifat negatif.

3.6 Deployment

Pada tahap *deployment*, tahap ini tujuannya yaitu digunakan untuk pemberian saran terhadap perusahaan Telkom yang memiliki layanan jaringan internet Telkomsel dengan cara melihat kata yang sering muncul dari data yang labelnya diprediksi oleh model yang menghasilkan nilai akurasi tertinggi, melihat hasil evaluasi untuk model KNN, SVM kernel RBF, dan *Ensemble* KNN-SVM berbasis *majority vote* dan *average*, dan pembuatan aplikasi untuk memprediksi kalimat dengan menggunakan model yang memiliki akurasi terbaik.

Didapatkan hasil akurasi terbaik dari setiap model algoritma adalah seperti pada Tabel 8.

Tabel 8 menunjukkan data nilai metrik evaluasi terbaik dari setiap model algoritma. model individual seperti model SVM memiliki kinerja yang lebih baik jika dibandingkan dengan model algoritma KNN. Model gabungan *Ensemble* KNN-SVM berbasis *average* memiliki nilai akurasi yang lebih baik

dengan nilai akurasi 84.79% jika dibandingkan dengan model algoritma individual SVM atau KNN dan *Ensemble* KNN-SVM berbasis *majority vote*.

Tabel 8. Metrik Evaluasi Terbaik Dari Setiap Model

Model	Akurasi	Presisi	Recall	F1-Score
KNN K=15	83.21%	81.43%	83.21%	78.84%
SVM C=100	84.33%	81.69%	84.33%	**81.88%**
Ensemble Majority Vote	83.26%	81.99%	83.26%	78.25%
Ensemble Average	**84.79%**	**82.90%**	**84.79%**	81.02%

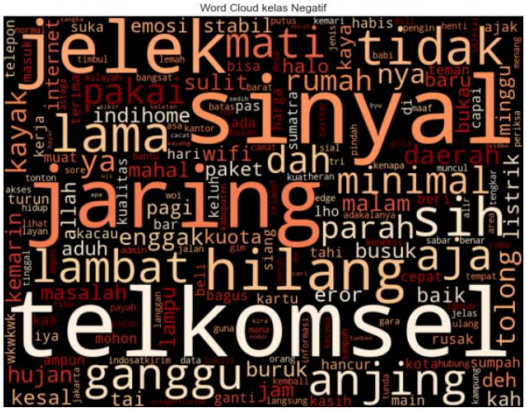
Ensemble KNN-SVM berbasis *average* dapat menghasilkan nilai yang lebih baik dari *Ensemble* KNN-SVM berbasis *majority vote* dikarenakan *Ensemble* KNN-SVM *majority vote* lebih sensitif terhadap sebuah kesalahan dalam memprediksi dikarenakan *Ensemble majority vote* bergantung pada mayoritas hasil dari kedua model klasifikasi individu sehingga dari kesalahan prediksi tersebut menyebabkan menurunnya nilai akurasi dari *Ensemble Majority Vote*, sedangkan *Ensemble Average* memiliki nilai akurasi lebih tinggi karena hasil prediksi diambil dari nilai rata-rata dari kedua model tersebut sehingga mengurangi kesalahan prediksi.

Untuk pemberian saran kepada perusahaan terkait mayoritas sentimen yang dihasilkan diambil dengan cara menghitung kata yang sering muncul pada kelas sentimen positif dan negative saja pada model yang memiliki nilai akurasi paling tinggi selama pelatihan yaitu model *Ensemble* KNN-SVM berbasis *Average* dengan nilai akurasi sebesar 84.79% dengan cara menggunakan pembobotan kata TF-IDF dan hanya diambil pada prediksi sentimen positif dan negatif. Prediksi yang bersentimen netral tidak digunakan sebagai saran dikarenakan sentimen netral tidak cenderung bersentimen positif maupun negative dan terkadang tidak bersentimen atau berisi kalimat tanya. Berikut merupakan hasil dari pembobotan kata TF-IDF untuk label sentimen negatif dan sentiment positif pada Tabel 9.

Tabel 9. Kemunculan Kata Dari Sentimen Negatif Dan Sentimen Positif

Nomor	Kata Negatif	Kata Positif
1.	Sinyal	Sinyal
2.	Jaring	Jaring
3.	Telkomsel	Telkomsel
4.	Jelek	Lancar
5.	Hilang	Kuat
6.	Lama	Kencang
7.	Lambat	Bagus

Selain itu bentuk visualisasi kata yang sering muncul dalam prediksi yang bersentimen negatif dan positif juga menggunakan pembobotan kata TF-IDF dan menghasilkan seperti pada
Gambar 9 dan
Gambar *10*.

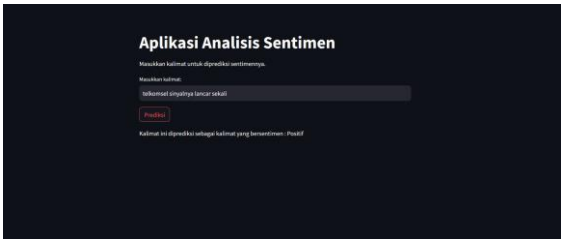


Gambar 9. Kata Yang Sering Muncul Pada Prediksi Sentimen Negatif

Gambar 10 Kata Yang Sering Muncul Pada Prediksi Sentimen Positif

Sehingga peneliti memberikan saran kepada perusahaan untuk dapat memperbaiki kualitas jaringan Telkomsel, mengoptimalkan kecepatan layanan Telkomsel, menyesuaikan antara kualitas jaringan Telkomsel dengan harga, hingga dapat mempertimbangkan penawaran paket harga yang lebih bersaing. Selain itu juga peneliti menyarankan kepada perusahaan untuk dapat mempertahankan, memperkuat, dan meningkatkan kualitas layanan jaringan Telkomsel.

Untuk aplikasi web prediksi kalimat dibuat dengan menggunakan *streamlit*. Aplikasi ini dibuat menggunakan model *machine learning* yang memiliki nilai akurasi terbaik dalam penelitian ini yaitu model *Ensemble* KNN-SVM berbasis *Average*, hal ini memungkinkan model untuk memberikan prediksi yang lebih akurat terhadap kalimat yang diberikan. Oleh karena itu, dengan memanfaatkan kinerja model ini, pengguna dapat mengimplementasikan pada kalimat yang baru untuk menentukan sentimen dari kalimat tersebut. Tampilan aplikasi web tersebut dapat dilihat pada
Gambar *11* pada saat melakukan prediksi kalimat.



Gambar 11. Aplikasi Digunakan Saat Melakukan Prediksi

4. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, didapatkan beberapa kesimpulan pada saat melakukan analisis sentimen dengan menggunakan model KNN, SVM, *Ensemble* KNN-SVM dengan data opini publik mengenai kualitas layanan jaringan internet Telkomsel. Berikut merupakan kesimpulan yang didapatkan dari penelitian.

1. Berdasarkan pengujian yang telah dilakukan, dapat dinyatakan bahwa KNN dengan nilai K=15 dengan nilai akurasi sebesar 83.21%, nilai presisi sebesar 81.43%, nilai *recall* sebesar 83.21%, dan nilai *f1-score* sebesar 78.84%. Untuk SVM dengan nilai C = 100 yaitu nilai akurasi sebesar 84.33%, untuk nilai presisi sebesar 81.69%, untuk nilai *recall* sebesar 84.33%, dan *f1-score* sebesar 81.88%. Untuk model Ensemble KNN-SVM berbasis *Majority Vote* memiliki nilai akurasi sebesar 83.26%, presisi sebesar 81.99%, *recall* sebesar 83.26%, dan *f1-score* sebesar 78.25%, sedangkan untuk model *Ensemble* KNN-SVM berbasis *Average* memperoleh nilai akurasi sebesar 84.79%, presisi sebesar 82.90%, *recall* sebesar 84.79%, dan *f1-score* sebesar 81.02%. Dapat disimpulkan bahwa keempat model algoritma tersebut memiliki nilai akurasi dan kinerja yang baik.
2. Hasil sentimen dari keempat model algoritma klasifikasi, dapat disimpulkan bahwa masing-masing algoritma memberikan hasil mayoritas prediksi terhadap pada opini publik mengenai kualitas layanan jaringan internet Telkomsel adalah negatif. Sehingga peneliti memberikan saran kepada perusahaan untuk dapat memperbaiki kualitas jaringan Telkomsel, mengoptimalkan kecepatan layanan Telkomsel, menyesuaikan antara kualitas jaringan Telkomsel dengan harga, hingga dapat mempertimbangkan penawaran paket harga yang lebih bersaing.
3. Dari penelitian yang telah dilakukan didapatkan bahwa metode *Ensemble* KNN-SVM berbasis *Average* merupakan metode yang menghasilkan nilai akurasi terbaik. Sehingga dapat disimpulkan bahwa *Ensemble* KNN-SVM berbasis *Average* mampu menggabungkan kekuatan dari kedua algoritma klasifikasi dengan keberhasilan meningkatkan akurasi. Dengan penggabungan dua metode dengan melalui metode *Ensemble* dapat mengambil keunggulan dari kedua metode yaitu KNN yang unggul dalam hal penanganan data *outliers* dan SVM yang unggul dalam hal penanganan data yang berdimensi tinggi.
4. Terjadi *overfitting* hal ini terjadi karena model terlalu fokus pada kelas mayoritas ("Negatif"), sehingga performanya menurun drastis pada kelas minoritas ("Netral" dan "Positif"). Pada penelitian selanjutnya diperlukan penyeimbangan data untuk memperbaiki kondisi *overfitting* ini.

DAFTAR PUSTAKA

ANANDA, F. D., & PRISTYANTO, Y. 2021. Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider Menggunakan Algoritma Support Vector Machine. *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *20*(2), 407–416. https://doi.org/10.30812/matrik.v20i2.1130

ARFAT, M. F., STYAWATI, NURKHOLIS, A., & KURNIAWAN, I. 2022. Analisis Sentimen Masyarakat Indonesia Terkait Vaksin Covid-19 Pada Media Sosial Twitter Menggunakan Metode Support Vector Machine (SVM). *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, *7*(2), 96–103.

ATMAJA, R. M. R. W. P. K., & YUSTANTI, W. 2021. Analisis Sentimen Customer Review Aplikasi Ruang Guru Dengan Metode BERT (Bidirectional Encoder Representations from Transformers). *Journal of Emerging Information System and Business Intelligence (JEISBI)*, *2*(3), 55–62.

BAITA, A., PRISTYANTO, Y., & CAHYONO, N. 2021. Analisis Sentimen Mengenai Vaksin Sinovac Menggunakan Algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN). *Information System Journal (INFOS)*, *Vol.4*(No. 2), 42–46. https://doi.org/http://dx.doi.org/10.30645/j-sakti.v5i2.386

DIHNI, V. A. 2022, June 13. *5 Operator Seluler Favorit Masyarakat Indonesia Versi APJII*. Kata Data Media Network (Databooks). https://databoks.katadata.co.id/datapublish/2022/06/13/5-operator-seluler-favorit-masyarakat-indonesia-versi-apjii

DOGAN, A., & BIRANT, D. 2019. A Weighted Majority Voting Ensemble Approach for Classification. *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 1–6. https://doi.org/10.1109/UBMK.2019.8907028

HUANG, Y. 2023. Improved SVM-Based Soil-Moisture-Content Prediction Model for Tea Plantation. *Plants*, *12*(12), 2309.

https://doi.org/https://doi.org/10.3390/plants12122309

INDRIATI, & KUSYANTI, A. 2019. Metode Ensemble Classifier untuk Mendeteksi Jenis Attention Deficit Hyperactivity Disorder (SDHD) pada Anak Usia Dini. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *6*(3), 301–307. https://doi.org/10.25126/jtiik.2019631313

BADAN PENGEMBANGAN DAN PEMBINAAN BAHASA, & KEMENDIKBUDRISTEK. (n.d.). *Apa itu teknologi?* Retrieved 21 August 2023, from https://kbbi.kemdikbud.go.id/entri/teknologi

KHADIJAH, & KUSUMANINGRUM, R. 2019. Ensemble Classifier untuk Klasifikasi Kanker Payudara. *IT Journal Research and Development (ITJRD)*, *4*(1), 61–71. https://doi.org/10.25299/itjrd.2019.vol4(1).3540

LESTARI, D. A., & MAHDIANA, D. 2021. Penerapan Algoritma K-Nearest Neighbor pada Twitter untuk Analisis Sentimen Masyarakat Terhadap Larangan Mudik 2021. *Informatik: Jurnal Ilmu Komputer*, *17*(2), 123–131. https://doi.org/10.52958/iftk.v17i2.3629

MANCONI, A., ARMANO, G., GNOCCHI, M., & MILANESI, L. 2022. A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19. *Applied Sciences*, *12*(15), 7554. https://doi.org/10.3390/app12157554

RIZATY, M. A. 2023, February 3. *Pengguna Internet di Indonesia Sentuh 212 Juta pada 2023*. DataIndonesia.Id. https://dataindonesia.id/digital/detail/pengguna-internet-di-indonesia-sentuh-212-juta-pada-2023

SALSABILA, A., SIHOMBING, J. J., & SITORUS, R. I. 2022. Implementasi Algoritma Support Vector Machine Untuk Analisis Sentimen Aplikasi OLX di Playstore. *Journal of Informatics and Data Science*, *1*(2), 1–6. https://doi.org/10.24114/j-ids.v1i2.42597

UMAIDAH, Y., & PURWANTORO. 2019. Penerapan Algoritma K-Nearest Neighbor (K-NN) dengan Pencarian Optimal untuk Prediksi Prestasi Siswa. *JISICOM (Journal of Information System, Informatics and Computing)*, *3*(2), 1–8. https://journal.stmikjayakarta.ac.id/index.php/jisicom/article/view/132

Halaman ini sengaja dikosongkan