Dike : Jurnal Ilmu Multidisiplin
Volume 1, Nomor 2, Agustus 2023, Page 37-43
ISSN 2986-884X (media online)

Analisis Sentimen Tweet COVID-19 Menggunakan Metode K-Nearest Neighbors dengan Ekstraksi Fitur TF-IDF dan CountVectorizer

Muhammad Hafizh Mahendra[1], Danang Triantoro Murdiansyah[*2], Kemas Muslim Lhaksmana[3]

[1,2,3)]Fakultas Informatika, Universitas Telkom, Bandung, Indonesia
Email: [1]madfrog@students.telkomuniversity.ac.id, [*2]danangtri@telkomuniversity.ac.id, [3]kemasmuslim@telkomuniversity.ac.id
Email Corresponding : danangtri@telkomuniversity.ac.id

Abstrak- Memasuki abad ke-21 seiring berkembangnya teknologi dan informasi, jumlah data yang ada di internet berkembang pesat. Hal ini menarik bagi orang-orang untuk mendapatkan data dan informasi untuk berbagai kebutuhan, seperti misalnya untuk kebutuhan penelitian akademik maupun penggunaan komersial. Banyak sekali data yang ada pada internet, diantaranya data terkait wabah penyakit COVID-19 (coronavirus) yang disebabkan oleh virus SARS-CoV-2 yang tersebar diseluruh dunia. Karena penyebaran penyakit tersebut yang cepat di berbagai daerah, WHO (World Health Organization) sempat menggolongkan wabah penyakit tersebut sebagai pandemi. Banyak orang dan berbagai pihak, termasuk media organisasi dan pemerintah, menghadirkan berita terbaru dan opini mengenai COVID-19. Dengan menganalisis opini publik terhadap COVID-19, kita dapat menyimpulkan sentimen publik terhadap COVID-19. Data kesimpulan tentang sentimen publik tersebut dapat digunakan untuk menentukan tindakan atau kebijakan penting terkait COVID-19. Dataset yang digunakan pada studi ini adalah berupa tweet terkait COVID-19 dari Tweeter yang diambil dari website Kaggle. Pada studi ini digunakan KNN (K-Nearest Neighbor) yang memiliki kompleksitas komputasi rendah untuk mengklasifikasikan tweet. Kemudian ekstraksi fitur yang digunakan adalah TF-IDF (Term Frequency - Inverse Document Frequency) dan CountVectorizer. Hasil pengujian pada studi ini menghasilkan hasil akurasi terbaik 73,2% dengan menggunakan TF-IDF.

Kata Kunci: COVID-19, analisis sentimen, KNN, TF-IDF, CountVectorizer.

Abstract-Entering the 21st century as technology and information develop, the amount of data on the internet is growing rapidly. It is interesting for people to get data and information for various needs, such as for academic research needs and commercial use. There are a lot of data on the internet, including data related to the outbreak of the COVID-19 disease (coronavirus) caused by the SARS-CoV-2 virus that spread throughout the world. Due to the rapid spread of the disease in various regions, WHO (World Health Organization) had classified the disease outbreak as a pandemic. Many people and various parties, including media organizations and governments, present the latest news and opinions about COVID-19. By analyzing public opinion on COVID-19, we can conclude the public's sentiment towards COVID-19. The conclusion data on public sentiment can be used to determine important actions or policies related to COVID-19. The dataset used in this study is in the form of tweets related to COVID-19 from Tweeters taken from the Kaggle website. This study uses KNN (K-Nearest Neighbor) which has a low computational complexity to classify tweets. Then the extraction features used are TF-IDF (Term Frequency - Inverse Document Frequency) and CountVectorizer. The test results in this study resulted in the best accuracy of 73.2% by using TF-IDF.

Keywords: COVID-19, sentiment analysis, KNN, TF-IDF, CountVectorizer.

1. PENDAHULUAN

Analisis sentimen saat ini merupakan salah satu topik penelitian yang populer di bidang pemrosesan bahasa alami. Saat ini, platform media sosial seperti Twitter, Facebook, dan YouTube menjadi sumber informasi yang disebut data sosial [1]. Mulai dari membahas peristiwa dalam kehidupan sehari-hari di media sosial, setiap orang dapat dengan bebas berdiskusi dan mengungkapkan pandangannya terhadap suatu peristiwa. Wabah COVID-19 yang mulai muncul di Wuhan, China pada akhir tahun lalu menjadi salah satu penyakit yang paling mengkhawatirkan dan menyebar secara global.

Menurut organisasi kesehatan dunia (WHO) lebih dari 20,000,000 orang telah tertular penyakit ini dan lebih dari 157,000 orang telah meninggal di seluruh dunia. Berdasarkan data tersebut, kita dapat melihat bahwa itu adalah salah satu epidemi virus biologis paling banyak dalam dua dekade terakhir abad ini. Dari banyak peneliti dan profesional dalam bidang kesehatan yang menganalisis dan memperoleh informasi baru dari data tersebut di platform media sosial, mereka dapat membantu para profesional kesehatan dan organisasi pemerintah mendapatkan manfaat dari data tersebut dan memahami reaksi masyarakat dan mengungkapkan perasaan mereka. Dalam penelitian ini, dataset tweet diambil dari Kaggle. Tweet yang digunakan mengandalkan dua kata kunci pencarian spesifik, yaitu #coronavirus dan #COVID-19.

Untuk ekstraksi fitur (feature extraction), berdasarkan paper [3] telah dilakukan pengujian berbagai feature extraction untuk task mengidentifikasi berbagai kasus yang relevan. Feature extraction yang digunakan pada paper tersebut yaitu Word2vec, Glove, TF-IDF. Hasil Pengujian menunjukkan Tf-Idf berkinerja lebih baik daripada metode vektorisasi lainnya. Atas dasar hal tersebut, pada penelitian ini juga akan dicoba digunakan ekstraksi fitur TF-IDF.

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**

Untuk algoritma classifier, berdasarkan paper [4] telah dilakukan pengujian terhadap metode K-Nearest Neighbors , Naive Bayes, Support Vector Machine, dan Random Forest. Pengujian ini dilakukan dengan IG (Information Gain) sebagai feature selection. Hasil pengujian menunjukkan metode KNN menghasilkan akurasi tertinggi dengan K=3 dan akurasi rata-rata sebesar 83,45%. Atas pertimbangan tersebut, pada penelitian ini akan dicoba juga digunakan metode K- Nearest Neighbors, dengan harapan akan menghasilkan akurasi yang baik.

Terdapat penelitian terkait analisis sentiment yang telah dilakukan peneliti lain sebelumnya. Pada paper [5], penulis meninjau blog twitter. Penulis tersebut membuat model untuk task classification. Terdapat task binary untuk mengklasifikasikan sentimen menjadi positif, negatif dan three way task, dimana mengklasifikasikan sentimen menjadi positif, negatif, dan netral. Pada task classification sentimen menjadi positif dan negatif telah digunakan sebanyak 1709 data untuk setiap kelas dan peluang dasarnya 50%. Sedangkan task classification sentimen menjadi positif, negatif, dan netral telah digunakan sebanyak 1709 data untuk setiap kelas dan peluang dasar untuk masing-masing kelas sebesar 33,33%.

Pada paper [6], penulis mengusulkan pendekatan otomatis mendeteksi sentimen pada pesan Twitter (Tweet) dan juga mengusulkan klasifikasi analisis sentimen dua langkah metode untuk Twitter. Pertama, penulis mengklasifikasikan pesan sebagai kategori subjektif dan objektif, selanjutnya membedakan tweet subjektif sebagai positif atau negatif. Untuk mengimplementasikan hal tersebut, penulis alih-alih menggunakan data yang dianotasi secara manual untuk menyusun data pelatihan sebagai pendekatan yang diawasi (supervised) secara teratur, penulis memanfaatkan sumber noisy label sebagai data pelatihan mereka. Noisy label didapatkan dari beberapa deteksi sentimen situs web melalui data Twitter. Penulis menggolongkan emoticon sebagai noisy label, karena dapat menyebabkan tidak sempurnanya saat mengklasifikasikan sentimen dari sebuah tweet. Sebagai contoh," @BATMANNN :(I love chutney ". Tanpa emoticon, kebanyakan orang akan mempertimbangkan tweet ini menjadi positif. Tweet dengan jenis emoticon tidak cocok digunakan untuk melatih classifier.

Pada paper [5], menggunakan metode KNN dengan normalisasi (4 feature) dan KNN dengan normalisasi dan keyword base (5 feature) terhadap 1000 tweet. Hasil yang didapat dari penelitian ini adalah algoritma KNN dengan normalisasi (4 feature) menghasilkan akurasi 80,80%. Sedangkan algoritma KNN dengan keyword base menghasilkan 84,32%.

Selanjutnya dalam penelitian tentang sentimen pada media sosial twitter untuk klasifikasi opini masyarakat indonesia terhadap kebijakan PPKM [7], penulis melakukan analisis sentimen terhadap opini masyarakat mengenai kebijakan PPKM di media sosial Twitter. Analisis sentimen ini menggunakan metode klasifikasi Naive Bayes Classifier. Hasil yang didapat dari disimpulkan penelitian ini adalah opini masyarakat mengenai kebijakan PPKM berupa 99% sentimen positif dan 1% sentimen negatif dari 1000 data ulasan..

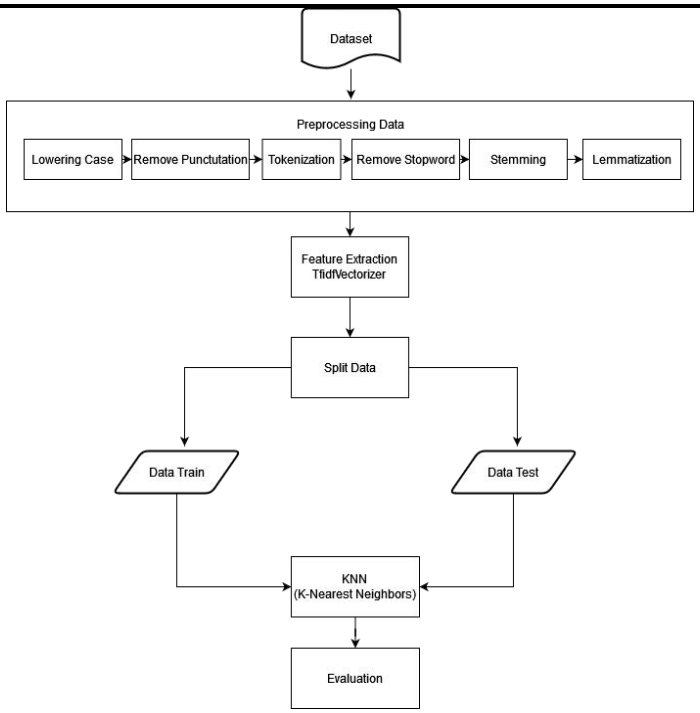
Selanjutnya dalam penelitian mengenai sentimen pada media sosial Twitter tentang topik Pilkada DKI 2017 [8], penulis melakukan analisis sentimen terhadap pengguna twitter mengenai Pilkada DKI 2017 di media sosial Twitter. Analisis Sentimen ini menggunakan metode klasifikasi K-Nearest Neighbor dengan pembobotan kata TF-IDF dan fungsi Cosine Similarity terhadap 2000 data tweet. Hasil yang didapat dari pengujian ini adalah akurasi sebesar 67,2% ketika k=5, precision tertinggi ketika k=5, dan recall 78,24% dengan k=15.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk mengetahui dan menganalisis sentiment masyarakat mengenai wabah penyakit COVID-19. Sentiment pada wabah tersebut dikelompokan menjadi dua kelas yaitu kelas positif dan negatif. Sistem yang dibangun pada penelitian ini menggunakan feature extraction Tf-IDF dan CountVectorizer pada tweet COVID-19. Gambaran sistem yang dibangun dapat dilihat pada Gambar 1.

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**

Dike : Jurnal Ilmu Multidisiplin
Volume 1, Nomor 2, Agustus 2023, Page 37-43
ISSN 2986-884X (media online)



Gambar 1. Alur Sistem yang Dibangun.

2.1 *Dataset*

Pada penelitian ini data yang digunakan diambil dari Twitter yang tersedia pada website Kaggle dalam bentuk file csv. Data berjumlah 36623 baris data. Sentimen dari tweet pengguna terdapat pada kolom sentiment. Atribut yang digunakan pada penelitian ini adalah OriginalTweet dan Sentiment. Pada atribut OriginalTweet ini dilakukan preprocessing data. Ada 2 kelas sentimen yaitu positif dan negatif.

2.2 *Preprocessing*

Preprocessing adalah proses yang dilakukan untuk menyiapkan data dalam format yang sesuai agar dapat digunakan oleh sistem atau algoritma, sehingga proses knowledge extraction dapat diterapkan [9]. Proses yang dilakukan dalam preprocessing adalah lowering case (case folding), removing punctuation, tokenization, stopword removal, stemming, dan lemmatization.

Lowering case atau case folding adalah proses mengubah semua huruf kapital menjadi huruf kecil. Tokenization akan semakin sulit jika harus memerhatikan struktur huruf lower-case dan upper-case [10]. Removing punctuation adalah proses untuk menghilangkan tanda baca atau simbol. Tokenization adalah proses memisahkan kalimat menjadi kata-kata [10]. Stopword removal adalah proses membuang kata yang dianggap tidak memiliki makna [10]. Stemming adalah proses pemetaan dan penguraian bentuk dari suatu kata menjadi bentuk kata dasarnya [11]. Lemmatization adalah proses menemukan bentuk kata terkait dalam kamus [11].

2.3 Ekstraksi Fitur *(Feature Extraction)*

Terdapat 2 Feature Extraction yang digunakan pada penelitian ini, yaitu TF-IDF (Term Frequency – Inverse Document Frequency) dan CountVectorizer.

2.3.1 TF-IDF (Term Frequency – Inverse Document Frequency)

Tahap selanjutknya yang dilakukan setelah proses preprocessing adalah proses feature extraction. Dokumen teks diubah menjadi bentuk vektor fitur dan kumpulan data yang ada digunakan untuk membuat fitur baru. Feature extration digunakan adalah TF-IDF. Term Frequency – Inverse Document Frequency (TF-IDF) adalah metode pembobotan kata dengan menghitung nilai Term Frequency dan menghitung kemunculan sebuah kata pada koleksi dokumen teks secara keseluruhan [12]. Term Frequency adalah jumlah kemunculan sebuah kata di dalam sebuah dokumen tertentu [13], semakin sering kata yang muncul maka semakin besar nilai Term Frequency [14]. Inverse Document Frequency adalah jumlah dokumen yang mengandung sebuah kata didasarkan pada seluruh dokumen yang ada pada dataset [13], semakin jarang kata yang muncul maka semakin besar nilai Inverse Document Frequency [14]. Hasil dari pembobotan kata adalah perkalian dari nilai Term Frequency dan Inverse Document Frequency yang akan

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**

menghasilkan bobot lebih kecil jika kata yang muncul lebih sering dan akan menghasilkan bobot lebih besar jika kata yang muncul lebih jarang [14].
Persamaan untuk Inverse Document Frequency adalah sebagai berikut [15].

$$IDF = \log(N/DFi) + 1 \quad (1)$$

Keterangan:
N = jumlah keseluruhan data
DFi = jumlah dokumen d yang mengandung kata wi

Sedangkan persamaan untuk Term Frequency – Inverse Document Frequency adalah sebagai berikut [15].

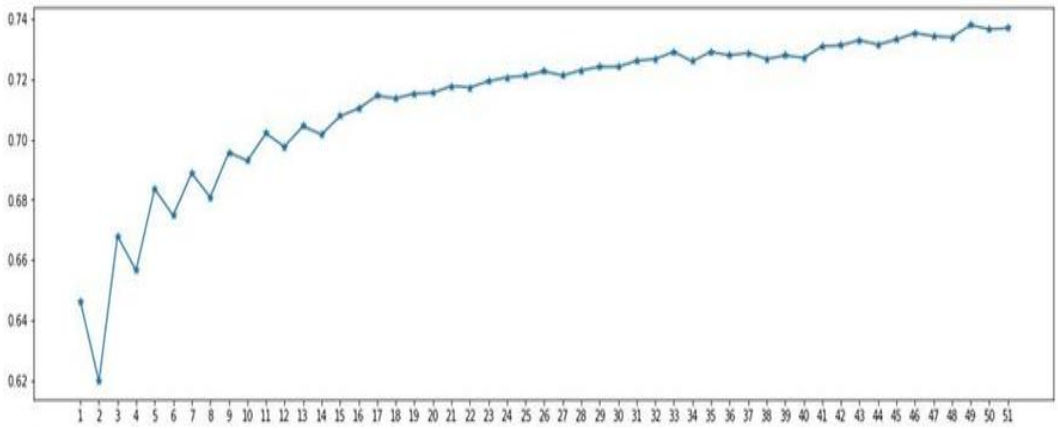
$$TF\text{-}IDF = TF(wi, d) \times IDF(wi) \quad (2)$$

Keterangan:
wi = kata ke-i
d = dokumen
TF(wi, d) = jumlah kemunculan kata wi di dalam dokumen d
IDF(wi) = jumlah dokumen d yang mengandung kata wi

2.3.2 K-Nearest Neighbors

K-Nearest Neighbors, yang disingkat dengan KNN, merupakan sebuah algoritma yang berfungsi untuk melakukan klasifikasi terhadap objek berdasarkan jarak antar data. Algoritma KNN menggunakan neighbourhood classification sebagai nilai klasifikasi yang baru [16]. Data yang menjadi acuan untuk klasifikasi adalah data sebanyak K yang jaraknya paling dekat dengan data yang akan diklasifikasikan.
Cara kerja algoritma KNN yaitu [17]

1. Tentukan jumlah tetangga (K) yang akan digunakan untuk pertimbangan penentuan (klasifikasi) kelas. Nilai K yang ditinjau pada penelitiian ini adalah 1 sampai 51. Nilai K yang dipilih adalah nilai K yang memiliki akurasi tertinggi. Grafik hasil pengujian tersebut dapat dilihat pada Gambar 2, di mana sumbu x adalah nilai K, dan sumbu y adalah akurasi.



Gambar 2. Grafik Penentuan Nilai K Terbaik.

2. Hitung jarak dari data baru ke masing-masing data point di dataset. Untuk menghitung jarak kita menggunakan rumus Euclidean distance sebagai berikut.

$$d(p, q) = \sqrt{\sum_{i=1}^{m}(p_i - q_i)^2} \quad (3)$$

keterangan:
p_i = sampel data dengan label kelas diketahui
q_i = data uji atau data testing

3. Urutkan hasil perhitungan tersebut secara ascending.

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**

Dike : Jurnal Ilmu Multidisiplin
Volume 1, Nomor 2, Agustus 2023, Page 37-43
ISSN 2986-884X (media online)

4. Ambil sejumlah K data dengan jarak terdekat, kemudian tentukan kelas dari data baru menjadi kelas dari tetangga dengan kelas terbanyak.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian

Pada penelitian ini memiliki 2 skenario pengujian, yaitu pengujian pada tahap feature extraction dan pengujian pada tahap data splitting. Pada skenario 1 dilakukan pengujian model klasifikasi terhadap feature selection. Pada skenario ini akan dilakukan pengujian terhadap hasil feature extraction menggunakan TF-IDF dan CV (CountVectorizer) (untuk level kata dan level n-gram), dengan menggunakan parameter max features yang berbeda yaitu menggunakan seluruh fitur, 8000, dan 10000. Tujuan skenario pengujian ini untuk melihat seberapa besar pengaruh feature extraction dan parameter max features terhadap akurasi model klasifikasi menggunakan K-Nearest Neighbor (KNN). Hasil akurasi pengujian skenario 1 dapat dilihat pada Tabel 1.

Tabel 1. Hasil Pengujian Skenario 1.

Max Features	Akurasi *Feature Extraction*		
	CV	TF-IDF	TF-IDF + Bi-Gram
(Seluruh fitur)	0.6533	**0.7325**	0.7204
10000	0.6740	0.7251	0.7177
8000	0.6645	0.7259	0.7196

Pada skenario 2 dilakukan pengujian pada tahap data splitting. Pada pengujian ini dilakukan menggunakan data splitting. Tujuan skenario pengujian ini untuk melihat hasil akurasi menggunakan metode K- Nearest Neighbor (KNN) yang terbaik pada pembagian data training dan data testing. Pada skenario ini, data splitting menggunakan 3 rasio berbeda yaitu menjadi

(a) Pada skenario ini menggunakan data yang rasionya 90:10. Hasil pengujian pada skenario ini dapat dilihat pada Tabel 2.

(b) Pada skenario ini menggunakan data yang rasionya 80:20. Hasil akurasi pengujian pada skenario ini dapat dilihat pada Tabel 3.

(c) Pada skenario ini menggunakan data yang rasionya 70:30. Hasil pengujian pada skenario ini dapat dilihat pada Tabel 4.

Tabel 2. Hasil Pengujian Skenario 2(a).

Max Features	90:10		
	Feature Extraction Accuracy		
	CV	TF- IDF	TF-IDF + Bi-Gram
(Seluruh fitur)	0.653	**0.732**	0.720
10000	0.674	0.725	0.718
8000	0.664	0.726	0.720

Tabel 3. Hasil Pengujian Skenario 2(b).

Max Features	80:20		
	Feature Extraction Accuracy		
	CV	TF- IDF	TF-IDF
(Seluruh fitur)	0.671	**0.740**	0.721
10000	0.659	0.726	0.712
8000	0.667	0.732	0.718

Tabel 4. Hasil Pengujian Skenario 2(c).

Max Features	70:30		
	Feature Extraction Accuracy		
	CV	TF- IDF	TF-IDF

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**

Dike : Jurnal Ilmu Multidisiplin
Volume 1, Nomor 2, Agustus 2023, Page 37-43
ISSN 2986-884X (media online)

(Seluruh fitur)	0.661	0.735	0.715
10000	0.661	**0.744**	0.728
8000	0.656	0.735	0.718

3.2 Pembahasan (Analisis Hasil Pengujian)

Berdasarkan hasil pengujian yang telah dibahas sebelumnya, pada skenario 1 nilai akurasi yang paling tinggi adalah 0,73. Hasil akurasi tertinggi tersebut dihasilkan dari feature extraction TF-IDF dengan parameter max features None. Dari hasil pengujian skenario 1 dapat dilihat dari Tabel 1. bahwa parameter max features mempengaruhi akurasi setiap feature extraction. Dapat diperhatikan bahwa nilai akurasi dari TF-IDF jauh lebih tinggi daripada CV. Ini bisa terjadi karena CV hanya menghitung jumlah kemunculan kata, sedangkan TF-IDF melakukan pembobotan yang skornya menunjukkan seberapa pentingnya sebuah term dalam dokumen dan seluruh korpus.

Pada skenario 2 nilai akurasi tertinggi adalah 0,744. Hasil akurasi tersebut diperoleh dari feature extraction TF-IDF dengan max features 10000 pada rasio data 70:30. Jika dilihat pada hasil data splitting menggunakan rasio 80:20 dan 70:30 pada Tabel 3 dan Tabel 4 memiliki perbedaan nilai akurasi yang tidak terlalu besar. Lalu jika kita bandingkan dengan hasil data splitting pada Tabel 2 dan Tabel 3 dapat dilihat hasil akurasi pada rasio data splitting 80:20 secara umum lebih besar dari 90:10. Hal ini bisa disebabkan oleh penambahan proporsi data training yang semakin banyak tidak membantu dalam mengklasifikasikan sentimen, yang artinya hanya menambahkan keyword atau fitur yang hampir sama dan tidak menambah pengetahuan sistem dalam mengklasifikasikan sentimen.

Tabel 5. *Confusion Matrix* Skenario 2(c).

		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	323 (TP)	839 (FP)
	Negative	1947 (FN)	4962 (TN)

Pada Tabel 5 dapat dilihat confusion matrix dari skenario yang memberikan hasil terbaik, yaitu dari skenario 2(c). TP adalah true positive, TN adalah true negative, FP adalah false positive, dan FN adalah false negative. Dapat dihitung nilai performansi dari model klasifikasi tersebut, yaitu

a. Precission = 0.7943.
b. Recall = 0.6246.
c. Specifity = 0.8554.
d. Accuracy = 0.7464.

Dari hasil perhitungan performansi diatas, perhatikan nilai recall dan specificity. Recall adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data positif. Sedangkan Specificity adalah rasio prediksi benar negatif dibandingkan dengan keseluruhan data negatif. Yang artinya, hasil prediksi dari model klasifikasi yang didapat setelah training yaitu dapat memprediksi data benar positif sebesar 62,5% dan dapat memprediksi data benar negatif sebesar 85,5%.

4. KESIMPULAN

Pada penelitian ini telah dibangun sistem klasifikasi untuk analisis sentimen tweet COVID-19 menggunakan metode K-Nearest Neighbor, dan TF-IDF serta CV sebagai feature extraction. Berdasarkan hasil pengujian yang dilakukan, penggunaan TF-IDF sebagai feature extraction dengan max features 10000, pada rasio data splitting 70:30, menghasilkan akurasi tertinggi yaitu sebesar 74,4%. Jika dilihat dari pengaruh rasio data splitting terhadap akurasi menunjukkan bahwa semakin banyak data training yang digunakan belum tentu membantu model untuk mengklasifikasikan data dengan benar. Kemudian jika feature extraction TF-IDF dibandingkan dengan CountVectorizer, hasil akurasi cukup berbeda jauh. TF-IDF menghasilkan hasil yang lebih baik dari CountVectorizer. Hal tersebut menunjukkan TF-IDF lebih cocok digunakan pada kasus di penelitian ini. Dari hasil pengujian pada penelitian ini juga dapat disimpulkan performansi dari sistem yang dibuat, yaitu sistem dapat memprediksi data benar positif sebesar 62,5% dan dapat memprediksi data benar negatif sebesar 85,5%.

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**

DAFTAR PUSTAKA

[1] Tyagi, Priyanka, and R. C. Tripathi 2019. A Review Towards the Sentiment Analysis Techniques for the Analysis of Twitter Data. Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE).
[2] Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2), 361-374.
[3] Kayalvizhi, S., Thenmozhi, D., & Aravindan, C. (2019). Legal Assistance using Word Embeddings. In FIRE (Working Notes) (pp. 36-39).
[4] Daeli, N. O. F., & Adiwijaya, A. (2020). Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. Journal of Data Science and Its Applications, 3(1), 1-7.
[5] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
[6] Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. International Journal of Advanced Computer Science and Applications, 8(6), 19-25.
[7] Krisdiyanto, T. (2021). Analisis sentimen opini masyarakat Indonesia terhadap kebijakan PPKM pada media sosial Twitter menggunakan naïve bayes clasifiers. Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi, 7(1), 32-37.
[8] Deviyanto, A., & Wahyudi, M. D. R. (2018). Penerapan analisis sentimen pada pengguna twitter menggunakan metode K-Nearest Neighbor. JISKA (Jurnal Informatika Sunan Kalijaga), 3(1), 1-13.
[9] Acuña, E. (2011). Preprocessing in Data Mining.
[10] Rimba Nuzulul Chory (2019). Analisis Sentimen pada tingkat kepuasan pengguna layanan data seluler menggunakan algoritma Support Vector Machine(SVM). OpenLibrary
[11] Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances. Um.edu.my
[12] Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research, 54-65.
[13] Muhammad Farhan Muzakki (2019). Analisis Sentimen Mahasiswa terhadap fasilitas Universitas Telkom menggunakan metode Jaringan Syaraf Tiruan dan TF-IDF. Open Library
[14] Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. Journal of Intelligent System and Computation, 1(1), 43-49.
[15] Assuja, M. A., & Saniati, S. (2016). Analisis Sentimen Tweet Menggunakan Backpropagation Neural Network. Jurnal Teknoinfo, 10(2), 48-53.
[16] Ismail, A. M. (2018). Cara Kerja Algoritma k-Nearest Neighbor (k-NN).
[17] Raza, G. M., Butt, Z. S., Latif, S., & Wahid, A. (2021, May). Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models. In 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2) (pp. 1-6). IEEE.
[18] Andini, A. S., Murdiansyah, D. T., & Lhaksmana, K. M. (2021). Topic Classification of Islamic Question and Answer Using Naïve Bayes and TF-IDF Method. Computer Engineering and Applications Journal, 10(3), 151-160.

Submitted: **20/07/2023**; Accepted: **20/08/2023**; Published: **31/08/2023**