(JUSS) Jurnal Sains dan Sistem Informasi ISSN 2614-8277
Vol.7 No.1, Juli 2024

ANALISIS SENTIMEN TWEET COVID-19 DI INDONESIA MENGGUNAKAN METODE ALGORITMA TF-IDF

Sekar Arum Kusuma Jakti [1)], Anna Dina Kalifia, S.Kom., M.Cs.[2]

[1]Fakultas Sains dan Teknologi, Universitas Teknologi Yogyakarta

email: kusuma.jakti20@gmail.com[1]anna.dina.kalifia@staff.uty.ac.id[2)].

Abstract

Sentiment analysis of Twitter user reviews about COVID-19 in Indonesia has become an important research topic in natural language processing. In this research, a dataset of tweets containing the keywords #coronavirus and #COVID-19 was taken from Kaggle and processed using the TF-IDF calculation method to perform sentiment analysis. This method involves collecting data from Twitter, preprocessing the dataset, and sentiment analysis using the TF-IDF algorithm. The results of this research provide clearer insight into public sentiment regarding COVID-19 in Indonesia based on data found on Twitter. By using the RapidMiner algorithm, this research can identify positive, negative or neutral sentiment from tweets related to COVID-19 in Indonesia. The results of this analysis can provide valuable information for the government and society to understand people's views and feelings regarding the COVID-19 pandemic. In addition, WordCloud visualization was also used to capture the essence of user reviews and identify key words frequently used in reviews about COVID-19 in Indonesia. Thus, this research makes an important contribution in understanding people's reactions and their feelings towards COVID-19 in Indonesia via the Twitter platform.

Keywords: *Analisis Sentimen, COVID-19, TF-IDF*

1. PENDAHULUAN

Analisis sentimen saat ini menjadi topik sayaan yang populer dalam bidang pemrosesan bahasa alami. Platform media sosial seperti Twitter, Facebook, dan YouTube menjadi sumber data sosial yang memberikan informasi tentang pandangan dan pendapat masyarakat terhadap suatu peristiwa. Wabah COVID-19 yang mulai muncul di Wuhan, China pada akhir tahun lalu telah menyebar secara global dan menjadi salah satu penyakit yang paling mengkhawatirkan.

Organisasi Kesehatan Dunia (WHO) melaporkan bahwa lebih dari 20 juta orang telah terinfeksi COVID-19 dan lebih dari 157 ribu orang telah meninggal dunia akibat penyakit ini. Data ini menunjukkan bahwa COVID-19 adalah salah satu epidemi virus biologis terbesar dalam dua dekade terakhir.

Banyak saya dan profesional di bidang kesehatan menganalisis data dari platform media sosial untuk memahami reaksi masyarakat dan perasaan mereka terkait COVID-19. Data ini dapat membantu para profesional kesehatan dan pemerintah dalam mengambil keputusan yang tepat dan memahami kebutuhan masyarakat.

Dalam sayaan ini, dataset tweet yang mengandung kata kunci #coronavirus dan #COVID-19 diambil dari Kaggle. Data ini kemudian diproses menggunakan metode perhitungan TF-IDF untuk melakukan analisis sentimen terhadap COVID-19 di Indonesia.

COVID-19 atau Corona adalah sebuah virus yang muncul pada tahun 2019 dan berasal dari China (Amelia, 2020). Penyebaran virus ini terjadi dengan sangat cepat dan menyebar ke seluruh negara, termasuk Indonesia (Kasmad, 2020). Munculnya virus COVID-19 ini mendapatkan berbagai tanggapan dan opini dari masyarakat Indonesia dalam kehidupan sehari-hari, terutama di platform digital atau media sosial seperti Facebook, Twitter, dan Instagram. Tanggapan dan opini yang dikemukakan melalui media sosial tersebut dapat mempengaruhi masyarakat yang membacanya. Oleh karena itu, perlu dilakukan penyaringan opini agar tidak salah menerima informasi dan

menilai apakah opini tersebut bersifat negatif, positif, atau netral. Salah satu cara untuk melakukan penyaringan tersebut adalah dengan melakukan analisis sentimen. Sentimen analisis adalah sayaan komputasional terhadap opini, sentimen, dan emosi yang diekspresikan secara tekstual (Zulfa & Winarko, 2017).

Analisis sentimen dilakukan untuk mengetahui kecenderungan penilaian terhadap suatu permasalahan dari sudut pandang orang lain. Dalam beberapa sayaan, sentimen dapat dikategorikan sebagai big data, di mana semakin banyak data teks yang ada, semakin beragam pula makna konteksnya. Proses analisis sentimen melibatkan tweet, teks ulasan, blog, forum, dan melibatkan tahapan preprocessing data seperti penghapusan stopword, stemming, identifikasi sentimen, tokenization, dan klasifikasi sentimen. Text mining merupakan bidang ilmu yang berkembang untuk mengekstrak informasi yang bermakna dari teks alami. Proses ini melibatkan tahapan preprocessing atau pembersihan data teks sebelum dilakukan tahap pengelompokkan atau klasifikasi lebih lanjut. Text mining memiliki tugas khusus dalam pengelompokkan teks (Zulfa & Winarko, 2017).

2. METODE PENELITIAN

Metode sayaan ini meliputi pengumpulan data dari sumber primer yaitu aplikasi tweeter untuk melihat komentar mengenai covid-19 di Indonesia dan dataset tambahan dari Kaggle, pengolahan data menggunakan RapidMiner,dan analisis sentiment menggunakan algoritma TF-IDF. Metode sayaan yang dilakukan adalah sebagai berikut :

1) Pengumpulan Data

Metode dalam pengumpulan data pada sayaan ini dilakukan dengan cara melakukan crawling data untuk mengambil tweets berbahasa Indonesia tentang topik terkait melalui fasilitas searching yang disediakan oleh Twitter. Pengambilan data menggunakan aplikasi Twitter dengan memanfaatkan tools RapidMiner. Atribut yang digunakan untuk pemrosesan data sayaan ini adalah atribut "Text". Jumlah data yang digunakan sebanyak 500 record data.

2) Prepocessing Dataset (Pembersihan Data)

Preprocessing adalah proses yang dilakukan untuk menyiapkan data dalam format yang sesuai agar dapat digunakan oleh sistem atau algoritma. Tujuannya adalah untuk mempermudah proses knowledge extraction atau ekstraksi pengetahuan.

Salah satu tahapan dalam preprocessing adalah lowering case atau case folding. Tahap ini melibatkan pengubahan semua huruf kapital menjadi huruf kecil.

Selanjutnya, tahap tokenization dilakukan untuk memisahkan kalimat menjadi kata-kata. Proses ini dapat menjadi lebih sulit jika harus memperhatikan struktur huruf lower-case dan upper-case.

Pada tahap removing punctuation, tanda baca atau simbol dihilangkan dari teks.

Stopword removal adalah tahap di mana kata-kata umum yang dianggap tidak memiliki makna atau tidak relevan dihapus dari teks.

Stemming adalah proses pemetaan dan penguraian bentuk kata menjadi bentuk kata dasarnya. Tujuannya adalah untuk mengurangi kata-kata yang memiliki akar kata yang sama menjadi bentuk yang lebih sederhana.

Lemmatization adalah proses menemukan bentuk kata terkait dalam kamus. Dalam tahap ini, kata-kata diubah menjadi bentuk kata dasar yang sesuai dengan kamus yang digunakan.

Semua tahapan preprocessing ini dilakukan untuk mempersiapkan data agar lebih terstruktur, mudah dipahami, dan siap untuk dianalisis lebih lanjut

3) Analisis Sentimen

Pada tahap ini, saya melakukan analisis sentiment dengan menggunakan Algoritma TF-IDF. TF-IDF (Term Frequency - Inverse Document Frequency) adalah metode

pembobotan kata yang menghitung nilai Term Frequency dan menghitung kemunculan sebuah kata pada koleksi dokumen teks secara keseluruhan.

Dalam analisis sentiment, Algoritma TF-IDF digunakan untuk menentukan bobot kata-kata dalam teks berdasarkan seberapa sering kata tersebut muncul dalam dokumen dan seberapa umum kata tersebut dalam keseluruhan koleksi dokumen. Dengan menggunakan metode ini, kata-kata yang sering muncul dalam dokumen tetapi jarang muncul dalam koleksi dokumen secara keseluruhan akan memiliki bobot yang tinggi, karena kata-kata tersebut dianggap penting dalam konteks dokumen tersebut.

3. HASIL DAN PEMBAHASAN

Hasil sayaan ini dilakukan menggunakan bahasa pemrograman RapidMiner dengan tujuan menerapkan analisis sentiment menggunakan algoritma RapidMiner untuk mengetahui sentimen dari Twitter tentang COVID-19 di Indonesia. Dalam sayaan ini, saya menggunakan bahasa pemrograman RapidMiner sebagai alat untuk melakukan analisis sentiment terhadap data yang terkait dengan COVID-19 di Twitter.

Dengan menggunakan algoritma RapidMiner, saya dapat mengidentifikasi sentimen positif, negatif, atau netral dari tweet yang berkaitan dengan COVID-19 di Indonesia. Algoritma ini memungkinkan saya untuk menganalisis teks yang terkandung dalam tweet dan menentukan apakah sentimen yang terkandung dalam teks tersebut cenderung positif, negatif, atau netral terkait dengan COVID-19.

Melalui analisis sentiment menggunakan algoritma RapidMiner, saya dapat memperoleh pemahaman yang lebih dalam tentang bagaimana masyarakat Indonesia merespons dan merasakan situasi terkait COVID-19 melalui platform Twitter. Informasi ini dapat sangat berharga dalam memahami pandangan dan perasaan masyarakat terhadap pandemi ini.

Dengan menggunakan bahasa pemrograman RapidMiner dan algoritma analisis sentiment yang diterapkan, sayaan ini dapat memberikan wawasan yang lebih jelas tentang sentimen masyarakat terkait COVID-19 di Indonesia berdasarkan data yang ditemukan di Twitter. Hal ini dapat membantu dalam pengambilan keputusan dan perencanaan strategi untuk mengatasi pandemi ini dengan lebih efektif.

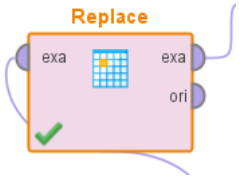
Dengan demikian, sayaan ini menggunakan bahasa pemrograman RapidMiner dan algoritma analisis sentiment untuk mengungkapkan sentimen masyarakat terkait COVID-19 di Indonesia berdasarkan data yang ditemukan di Twitter. Hasil sayaan ini dapat memberikan informasi berharga bagi pemerintah, saya, dan masyarakat untuk memahami pandangan dan perasaan masyarakat terkait pandemi ini.

a) Klasifikasi Data

Dalam langkah pertama klasifikasi data, saya melakukan preprocessing data untuk mempersiapkan data sebelum dilakukan analisis sentimen. Preprocessing data melibatkan beberapa tahapan yang bertujuan untuk meningkatkan kualitas data sebelum dilakukan analisis. Tahapan-tahapan yang dilakukan saya dalam preprocessing data antara lain:

a) Data Cleaning (Pembersihan Data)

Pada tahap pembersihan data, saya melakukan identifikasi dan penanganan nilai yang tidak valid atau outliers. Salah satu langkah yang dilakukan adalah menggunakan operator "Filter Examples" untuk memilih data yang tidak mengandung nilai yang hilang (missing value). Selain itu, data juga diubah untuk menggantikan simbol yang dianggap tidak penting dengan spasi menggunakan operator "Replace".



 ISSN 2614-8277

b) Data Transformation (Transformasi Data)

Dalam proses transformasi data, tujuannya adalah mengubah format atau struktur data agar sesuai dengan kebutuhan analisis. Salah satu operator yang digunakan untuk mengubah data kategorikal menjadi teks adalah "Nominal to Text". Hal ini dilakukan karena algoritma yang akan digunakan dalam analisis sentimen membutuhkan tipe data teks.

Dalam konteks sayaan ini, saya menggunakan operator "Nominal to Text" pada bahasa pemrograman RapidMiner untuk mengubah data kategorikal menjadi teks. Hal ini penting karena analisis sentimen yang akan dilakukan terkait dengan data Twitter tentang COVID-19 di Indonesia membutuhkan data dalam bentuk teks.

c) Data Reduction (Pengurangan Data)

Dalam tahap pengurangan data, saya melakukan langkah-langkah untuk mengurangi jumlah data atau atribut tanpa menghilangkan informasi penting. Dalam konteks sayaan ini, saya hanya akan mengambil atribut “Text” dari atribut yang ada pada dataset. Selain itu, saya juga membatasi jumlah data yang akan digunakan menjadi 500 data sebagai sampel yang cukup representatif untuk analisis.

Untuk melakukan pengurangan data, saya menggunakan operator "Select Attribute" dan "Filter Examples". Operator "Select Attribute" digunakan untuk memilih atribut yang akan digunakan dalam analisis, yaitu atribut “Text”. Sedangkan operator "Filter Examples" digunakan untuk memilih jumlah data yang akan digunakan, dalam hal ini sebanyak 500 data.

Gambar 1. Proses Cleaning, Tranformation, dan Reduction.

Row No.	word
1	covid
2	https
3	positif
4	netral
5	pandemi
6	jakarta
7	positive
8	indonesia
9	police
10	orang
11	vaksinasi
12	omicron
13	vaksin
14	februari
15	vaccination

ExampleSet (500 examples,0 special

There are no sources in the current document.

Gambar 2. Hasil dari Cleaning, Transformation, dan Reduction.

d) Tokenization

Tokenisasi (tokenization) adalah proses memecah kalimat menjadi bagian kata-kata. Pada tahap ini, dilakukan tokenisasi untuk menganalisis sentimen pada tingkat kata. Tujuan dari tokenisasi adalah untuk memecah sebuah kalimat menjadi kepingan-kepingan kata. Misalnya, kalimat "Ini baju saya" akan menjadi "ini", "baju", "saya".

Tokenisasi sering digunakan dalam analisis teks untuk memahami makna dari sebuah teks. Dengan memecah kalimat menjadi kata-kata, sistem dapat melakukan pengecekan terhadap tiap-tiap kata dalam kalimat untuk analisis lebih lanjut.

e) Case-Folding

Proses mengubah semua huruf dalam teks menjadi huruf kecil dilakukan untuk memastikan konsistensi dalam analisis sentimen tanpa memperhitungkan perbedaan huruf besar atau kecil. Langkah ini penting agar data yang diolah memiliki format yang seragam dan dapat menghasilkan hasil yang lebih akurat

dalam analisis sentimen. Dengan mengubah huruf menjadi huruf kecil, informasi yang terkandung dalam teks dapat diproses dengan lebih mudah dan efisien. Hal ini membantu para peneliti atau analis untuk menginterpretasikan data dengan lebih baik dan mendapatkan wawasan yang lebih mendalam tentang sentimen yang terkandung dalam teks terkait COVID-19 di Twitter.

f) Filtering

Proses mengambil kata-kata yang penting dari hasil tokenisasi merupakan tahap penting dalam preprocessing data. Pada tahap ini, dilakukan penghapusan kata-kata umum atau stopwords yang tidak memberikan kontribusi signifikan pada analisis sentimen terkait COVID-19 di Twitter

Langkah ini dilakukan untuk memastikan bahwa kata-kata yang relevan dan memiliki makna penting dalam teks dipertahankan, sementara kata-kata yang umum dan tidak memberikan informasi yang berarti dihapus. Dengan demikian, analisis sentimen dapat dilakukan dengan lebih fokus pada kata-kata yang memiliki dampak signifikan dalam mengekspresikan sentimen terkait COVID-19 di Twitter. Proses selanjutnya menuju TF-IDF, berikut operator operator yang akan di gunakan:

Gambar 3. Proses Tokenize, Case Folding, Filtering

2) Algoritma TF-IDF

Sebuah metode yang mengubah sebuah kalimat yang berisi kata-kata menjadi sebuah matriks telah dikembangkan. Dalam metode ini, data yang dihasilkan dapat dimasukkan sebagai dataset dalam bentuk list. Token kemudian dimasukkan ke dalam sebuah variabel dan diubah menjadi sebuah Tensor yang berisi angka yang diproses melalui sebuah algoritme. Perhitungan bobot kata dimulai dengan menghitung nilai TF, di mana bobot masing-masing kata adalah 1. Selain itu, IDF juga dihitung untuk menentukan bobot kata yang lebih akurat. Metode ini memungkinkan representasi kalimat dalam bentuk matriks, yang memudahkan analisis dan pemrosesan lebih lanjut.

a) Term Frequency (TF)

Term Frequency (TF) adalah frekuensi kemunculan suatu kata dalam sebuah dokumen. Semakin tinggi nilai TF-nya, semakin besar bobot atau kepentingannya dalam konteks dokumen tersebut. Dalam kata lain, semakin sering kata muncul, semakin tinggi pula nilai TF-nya, memberikan gambaran mengenai seberapa pentingnya suatu kata dalam suatu konteks dokumen.

Dalam tahap preprocessing data, penghitungan Term Frequency (TF) merupakan salah satu langkah penting. Dengan menghitung frekuensi kemunculan kata-kata dalam teks, kita dapat menentukan seberapa sering kata-kata tersebut muncul dan seberapa pentingnya kata-kata tersebut dalam konteks dokumen yang sedang dianalisis.

b) Invers Doucument Frequency (IDF)

Rumus IDF dihitung dengan menggunakan document frequency (df), yaitu jumlah dokumen yang mengandung kata tersebut. IDF dapat dihitung dengan rumus logaritma dari jumlah total dokumen dibagi dengan document frequency dari kata tersebut. Penambahan 1 pada pembilang rumus IDF dilakukan untuk menghindari pembagian dengan 0 jika document frequency tidak ditemukan pada seluruh koleksi dokumen.

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

Proses TF-IDF pada penelitian ini menggabungkan nilai TF dan IDF untuk menghitung bobot akhir suatu kata dalam konteks dokumen. Bobot akhir suatu kata dihitung dengan mengalikan nilai TF dengan nilai IDF. Dengan menggunakan metode TF-IDF, kita dapat mengidentifikasi kata-kata yang memiliki kontribusi penting dalam analisis teks dan memberikan representasi yang lebih akurat dalam bentuk matriks atau model yang digunakan. Adapun proses TF-IDF pada penelitian ini sebagai berikut :

 ISSN 2614-8277

ExampleSet (Filter Example Range (2))

Open in Turbo Prep Auto Model

Row No.	word	in documents	total
1	covid	448	702
2	https	284	471
3	positif	102	122
4	netral	96	96
5	pandemi	42	43
6	jakarta	18	33
7	positive	32	33
8	indonesia	20	32
9	police	29	31
10	orang	26	30
11	vaksinasi	30	30
12	omicron	22	28
13	vaksin	27	28
14	februari	27	27
15	vaccination	26	26

ExampleSet (500 examples,0 special attributes,3 regular attributes)

Gambar 4. Hasil TF-IDF

d) Hasil Sentimen dan Visualisasi WordCloud

Hasil Analisis Sentimen pada gambar dan gambar serta telah diberikan pelabelan ke dalam kelas sentiment positif dan negatif.

ExampleSet (Generate Attributes)

Open in Turbo Prep Auto Model

Row No.	word	in documents	total	sentimen
1	covid	488	846	positif
2	https	423	718	positif
3	positif	127	156	positif
4	netral	137	137	positif
5	positive	51	53	negatif
6	case	41	41	negatif
7	police	38	40	negatif
8	indonesia	21	36	negatif
9	jakarta	20	35	negatif
10	orang	32	35	negatif
11	omicron	21	34	negatif
12	yustisi	16	32	negatif
13	vaksinasi	30	31	negatif
14	data	17	30	negatif
15	pandemi	28	29	negatif

ExampleSet (20 examples,0 special attributes,4 regular attributes)

pada Gambar 5, dalam Wordcloud tersebut, kata-kata yang lebih besar menunjukkan bahwa kata-kata tersebut sering digunakan oleh pengguna Twitter dalam ulasan mereka tentang COVID-19 di Indonesia. Oleh karena itu, Wordcloud dapat memberikan gambaran yang lebih jelas tentang topik dan kata-kata positif yang sering digunakan oleh pengguna Twitter dalam memberikan ulasan tentang COVID-19 di Indonesia.

Dengan menggunakan visualisasi Wordcloud, kita dapat dengan mudah mengidentifikasi topik yang sering dibahas oleh pengguna Twitter dalam ulasan mereka tentang COVID-19 di Indonesia. Beberapa topik yang sering dibahas adalah COVID.

Dalam kesimpulannya, visualisasi Wordcloud dapat memberikan gambaran yang lebih jelas tentang topik dan kata-kata positif yang sering digunakan oleh pengguna Twitter dalam memberikan ulasan tentang COVID-19 di Indonesia. Wordcloud menjadi alat visual yang berguna untuk menangkap esensi dari ulasan pengguna dan mengidentifikasi kata-kata kunci.

Gambar 5. Wordcloud TF-IDF

4. KESIMPULAN

Kesimpulan dari penelitian adalah bahwa penelitian ini menggunakan metode algoritma TF-IDF untuk melakukan analisis sentimen terhadap tweet-tweet yang berkaitan dengan COVID-19 di Indonesia. Data tweet diambil dari Twitter dan diproses menggunakan RapidMiner. Tahapan analisis sentimen meliputi pengumpulan data, preprocessing dataset, dan analisis sentimen menggunakan algoritma TF-IDF. Hasil penelitian ini dapat memberikan informasi berharga bagi pemerintah dan masyarakat untuk memahami pandangan dan perasaan masyarakat terkait pandemi COVID-19. Proses analisis teks

tersebut juga penting untuk mendapatkan representasi yang lebih akurat dalam bentuk matriks atau model yang digunakan, dan hasil analisis dapat ditampilkan dalam bentuk visualisasi WordCloud.

5. REFERENSI

449-Article Text-1300-1-10-20220424. (n.d.).

Abdillah, W. F., Premana, A., & Herdian Bhakti, R. M. (2021). Analysis Of Covid-19 Handling Sentiment With Support Vector Machine: Lexicon Evaluation And Feature Extraction Methods. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *3*(2), 160–170.

Analisis Sentimen Terhadap Pembelajaran Daring dengan Algoritma Naïve Bayes Classifier. (n.d.).

Apriliani, D., Susanto, A., Fikri Hidayattullah, M., & Wiro Sasmito, G. (2023). *Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors*. *8*(1).

document. (n.d.).

Hafizh Mahendra, M., Triantoro Murdiansyah, D., & Muslim Lhaksmana, K. (2023). *Dike : Jurnal Ilmu Multidisiplin Analisis Sentimen Tweet COVID-19 Menggunakan Metode K-Nearest Neighbors dengan Ekstraksi Fitur TF-IDF dan CountVectorizer*.

Najiyah, I., & Haryanti, I. (2021). 100~111 Naskah diterima 02 Februari 2021; direvisi 18 Februari 2021. *JURNAL RESPONSIF*, *3*(1). http://ejurnal.ars.ac.id/index.php/jti

Pendidikan, J., & Konseling, D. (n.d.). *Analisis Sentimen Tempat Wisata Di Jakarta Pasca Covid-19 Dengan Algoritma Naïve Bayes Universitas Pahlawan Tuanku Tambusai*. *4*.

Silalahi, N., & Ginting, G. L. (2022). Analisa Sentimen Masyarakat Dalam Penggunaan Vaksin Sinovac Dengan Menerapkan Algoritma Term Frequence – Inverse Document Frequence (TF-IDF) dan Metode Deskripsi. *Journal of Information System Research (JOSH)*, *3*(3), 206–217. https://doi.org/10.47065/josh.v3i3.1441

Yuyun, Nurul Hidayah, & Supriadi Sahibu. (2021). Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 820–826. https://doi.org/10.29207/resti.v5i4.3146