Jurnal Sistim Informasi dan Teknologi
https://jsisfotek.org/index.php
2022 Vol. 4 No. 2 Hal: 64-69 e-ISSN: 2686-3154

Klasifikasi Tweet di Twitter dengan Menggunakan Metode K-Nearest Neighbor

Ahmad Fauzi Rahman[1]✉
[1]Independent Researcher
ahmadfauzirahman99@gmail.com

Abstract

Twitter is a social media and also a microblog that allows users to send and read short messages of no more than 280 characters as tweets. Tweet users are currently not limited in terms of age, information, and communication. The lack of information on tweets on Twitter results in a lot of information that is difficult to know the facts or intentions of the tweet. This can have negative impacts such as fraud, leading public opinion to negative. One of the topics that has been discussed in Indonesian society since March 2020. Coranavirus Disease 2019 (COVID-19) is a disease of the new type of coronavirus virus (SARS-CoV-2) that has taken the world by storm and the WHO has declared it a pandemic. As of March 10, 2022, 5.85 million Covid cases have been confirmed out of a total of 453 million cases in the world. This Covid-19 case can be a place for certain parties to disseminate information to the wider community. Tweet data obtained from social networks based on queries in Indonesian, this study aims to determine the pros or cons of the public or society against Covid-19. To find out that a tweet is counter can be done by looking at the existing tweets one by one, but this takes a long time and effort because of the large number of tweets. In this study using the K-Nearest Neighbor (KNN) method using as many as 1200 tweet data. Based on the research that has been done, it can be said that public sentiment in 2020 regarding COVID-19 tends to be negative, followed by positive and neutral opinions. But in 2021, as time goes by, public opinion tends to follow neutral.

Keywords: Sentiment Analysis, Classification, Twitter, Data Mining, K-Nearest Neighbor.

Abstrak

Twitter merupakan salah satu jejaring social dan juga mikroblog yang memungkinkan penggunanya berkirim dan membaca pesan singkat yang tidak lebih dari 280 karakter sebagai tweet. Pengguna tweet saat ini tidak dibatasi dari segi umur, informasi, dan komunikasi. Minimnya sebuah informasi pada tweet yang ada ditwitter mengakibatkan banyak informasi yang sulit untuk diketahui fakta atau maksud dari tweet tersebut. Hal ini dapat menimbulkan dampak negatif seperti penipuan, penggiringan opini publik terhadap hal yang negatif. Salah satu topik yang menjadi perbincangan di masyarakat Indonesia sejak Maret 2020 lalu. Coranavirus Disease 2019 (COVID-19) adalah penyakit dari virus coronavirus jenis baru (SARS-CoV-2) yang menggemparkan seluruh dunia dan WHO menetapkan sebagai sebuah pandemic. Hingga 10 Maret 2022 sudah dikonfirmasikan sejumlah 5,85 juta kasus Covid dari total 453 juta kasus di dunia. Kasus Covid-19 ini dapat menjadi sebuah tempat bagi pihak tertentu melakukan penyebaran informasi kepada masyarakat yang luas. Data tweet yang diperoleh dari jejaring sosial berdasarkan query dalam bahasa indonesia, penelitian ini bertujuan menentukan pro atau kontra publik atau masyarakat terhadap Covid-19. Untuk mengetahui bahwa suatu tweet adalah kontra dapat dilakukan dengan melihat satu persatu tweet yang ada, namun hal ini membutuhkan waktu dan tenaga yang lama karena jumlah tweet yang banyak. Pada penelitian ini menggunakan Metode K-Nearest Neighbor (KNN) menggunakan sebanyak 1200 data tweet. Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa sentiment masyarakat pada tahun 2020 mengenai COVID-19 cenderung negatif, kemudian diikuti dengan opini positif dan juga netral. Namun pada tahun 2021, seiring berjalannya waktu, opini masyarakat cenderung kearah netral.

Kata kunci: Analisis Sentimen, Klasifikasi, Twitter, Data Mining, K-Nearest Neighbor.

1. Pendahuluan

Twitter adalah micro-blogging dan layanan jejaring sosial waktu nyata yang telah mendapatkan popularitas luas sejak dekade terakhir. Dengan lebih dari 313 juta pengguna aktif dan lebih dari 500 juta posting per hari. Twitter adalah salah satu cara untuk tetap terhubung secara sosial dengan teman, keluarga, dan kolega lewat media online. Ini memungkinkan penggunanya untuk mengekspresikan pandangan mereka dan berinteraksi dengan orang lain dengan bantuan posting yang biasa disebut tweet [1]. Twitter merupakan salah satu platform tempat masyarakat menuangkan isi pikiran, permasalahan yang terjadi dan saling mendebat satu sama lain [2].

Berita tentang virus menyebar di seluruh situs media sosial. Akibatnya, outlet media sosial ini mengalami dan menghadirkan pandangan, pendapat, dan emosi yang berbeda selama berbagai insiden terkait wabah [3]. Jumlah tweet harian pada platform twitter mencapai 500 juta tweet opini. Berita soal COVID-19

Diterima: 16-03-2022 | Revisi: 20-04-2022 | Diterbitkan: 30-06-2022 | doi: 10.37034/jsisfotek.v4i2.125

menjadi salah satu topik yang menjadi perbincangan di masyarakat Indonesia sejak Maret 2020 lalu. Coranavirus Disease 2019 (COVID-19) adalah penyakit dari virus coronavirus jenis baru (SARS-CoV-2) yang menggemparkan seluruh dunia dan WHO menetapkan sebagai sebuah pandemic [4]. Hingga 10 Maret 2022 sudah dikonfirmasikan sejumlah 5,85 juta kasus Covid dari total 453 juta kasus di dunia[5]. Kasus Covid-19 ini dapat menjadi sebuah tempat bagi pihak tertentu melakukan penyebaran informasi kepada masyarakat yang luas.

Sejak COVID-19 ini menjadi pandemic diseluruh dunia banyak opini masyarakat yang bersifat pro dan kontra terhadap Covid-19 [6]. Tweet yang masyarakat buat mempunyai peranan yang penting, karena bisa dijadikan dataset untuk proses data mining. Data mining merupakan salah satu metode analisis untuk mendapatkan informasi yang bersumber dari data dengan jumlah yang sangat besar [7]. Salah satu yang bisa dilakukan dengan data penelitian yang disediakan yaitu klasifikasi sentiment terhadap Covid-19 dan memprediksi keakuratan algoritma K-Nearest Neighbor, karena setiap tweet masyarakat pasti memiliki pendekatan pro, kontra ataupun netral [8].

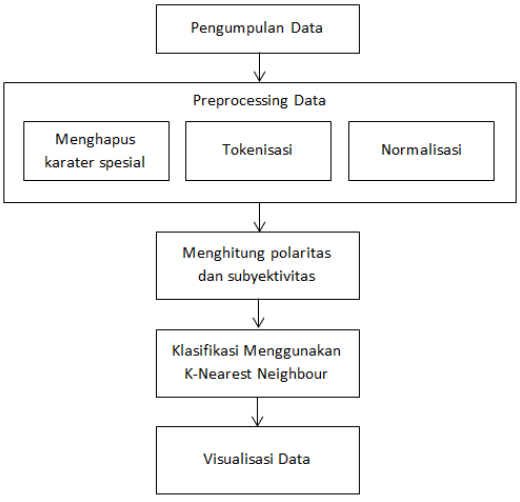
Analisis sentimen memberikan cara yang efektif untuk mengukur opini publik tentang topik apapun yang sedang menjadi isu saat ini [9]. Sentiment analisis atau disebut juga opinion mining merupakan sebuah bidang ilmu yang menganalisa pendapat, sentiment, evaluasi, penilaian, sikap dan emosi pengguna atau orang pada satu layanan, produk, peristiwa, dan topik [10]. Sentimen bertujuan untuk mengetahui apakah opini masyarakat dalam bentuk sebuah tweet yang dikirimkan oleh pengguna memiliki makna pro dan kontra. Dengan melakukan sentimen analisis ini bisa dijadikan referensi untuk pengembangan suatu produk atau layanan [11].

Salah satu teknik dalam data mining adalah klasifikasi, klasifikasi adalah tugas untuk memprediksi label kategori yang tidak diketahui sebelumnya untuk membedakan antara satu objek dengan objek lainnya berdasarkan atribut atau fitur [12]. Teknik klasifikasi yang paling populer adalah K-Nearest Neighbor (KNN) karena sangat sederhana dan mudah [13]. Namun ada juga algoritma klasifikasi K-Nearest Neighbor (MKNN) yang merupakan algoritma turunan dari KNN. Penelitian tentang perbandingan algoritma KNN dan MKNN telah dilakukan dan mendapat hasil nilai akurasi MKNN lebih tinggi dari KNN [14].

Penerapan metode K-Nearest Neighbor bisa diterapkan di berbagai macam bidang. Salah satunya di dunia pendidikan yang biasa dikenal dengan Edutional Data Mining (EDM). EDM telah banyak digunakan dalam dunia pendidikan, salah satunya adalah untuk memprediksi prestasi akademik siswa [15]. Selain itu, penelitian mengenai identifikasi dan klasifikasi tweet spam menggunakan Data Mining di Twiiter sudah dilakukan sebelumnya. Penelitian tersebut membahas tentang sentiment masyarakat terhadap putusan pemerintah untuk melakukan lockdown di India untuk meminimalisir angka kasus COVID-19 yang terjadi. Pada penelitian tersebut didapat hasil mayoritas masyarakat menyambut baik kebijakan pemerintah [16]. Untuk itu, penelitian ini akan melakukan klasifikasi Tweet pada Twitter dengan menggunakan metode K-Nearest Neighbor.

2. Metodologi Penelitian

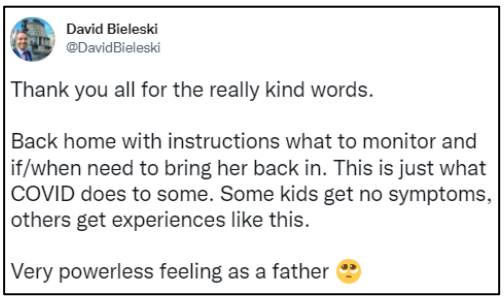
Metodologi penelitian merupakan sebuah proses pelaksanaan penelitian yang terdiri dari langkah-langkah dan juga menerapkan prinsip metode ilmiah. Penelitian memerlukan kerangka kerja agar hasil dan tujuan penelitian dapat dicapai secara maksimal. Pada penelitian ini yang menjadi tujuan capaian hasil penelitian ialah klasifikasi tweet menggunakan algoritma K-Nearest Neighbour. Data diambil dengan kata kunci "COVID-19" pada tahun 2020-2021 secara acak dan didapatkan sebanyak 1200 tweets populasi. Data ini perlu diproses terlebih dahulu karena mengandung karakter khusus, hyperlink, retweet, emoji, dan stiker. data tweet pertama diambil dari Twitter dengan pencarian standar Twitter menggunakan Library Tweepy, dan data yang diambil disimpan dalam format file csv. Data ini perlu diproses terlebih dahulu karena mengandung karakter khusus, hyperlink, retweet, emoji, dan stiker. Pemrosesan bahasa alami digunakan untuk memproses data sebelumnya dan agar bisa diterapkan algoritma supervised klasifikasi yaitu KNN [17]. Setelah menghapus karakter khusus, data diberi token. Untuk proses selanjutnya dilakukan normalisasi kemudian data diolah menggunakan NLP. Selanjutnya klasifikasi menggunakan algoritma KNN dan data hasil klasifikasi akan divisualisasi agar informasi lebih mudah dipahami. Adapun kerangka kerja penelitian dapat dilihat pada Gambar 1.



Gambar 1. Kerangka Kerja Penelitian

2.1. Pengumpulan Data

Pada tahap ini merupakan salah satu tahapan pengambilan data. Data yang digunakan yaitu 1200 data yang nantinya dibagi sebagai data latih dan data uji. Pengumpulan data latih dan data uji didapatkan dari tweet yang terbaru. Pembagian data latih dan data uji berdasarkan komentar yang diambil yaitu 70% data latih : 30% data uji, 80% data latih : 20% data uji, dan 90% data latih : 10% data uji. Pada tahapan ini data diambil menggunakan octoparse. Data diambil dengan kata kunci "COVID-19" pada tahun 2020-2021 secara acak dan didapatkan sebanyak 1200 tweets populasi. Data ini perlu diproses terlebih dahulu karena mengandung karakter khusus, hyperlink, retweet, emoji, dan stiker. Adapun salah satu contoh tweet opini positif dan negatif dari www.twitter.com bisa dilihat pada Gambar 2, 3 dan 4.



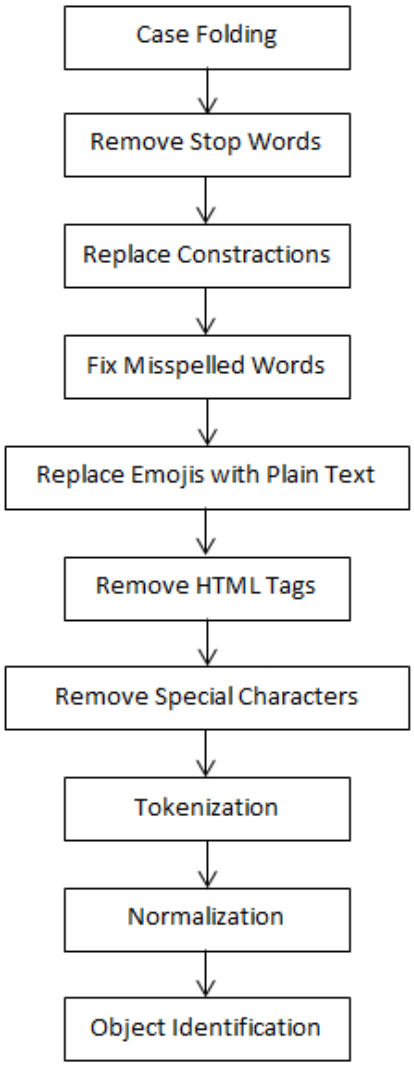
Gambar 2. Opini Positif tentang COVID

Gambar 3. Opini Negatif tentang COVID

Gambar 4. Opini Netral tentang COVID

2.2. Preprocessing dataset

Pada tahapan ini dilakukan preprocessing dataset agar data bisa diolah pada tahapan klasifikasi menggunakan metode KNN. Pada tahap ini tiap data teks akan diubah menjadi huruf kecil kemudian setiap akhir kalimat dengan tanda titik akan dihapus. Daftar akhir kata dalam kalimat akan didefinisikan secara khusus yang digunakan dalam proses ini guna untuk menghindari kata yang salah eja. Salah satu langkah terpenting dalam preprocessing pada tahap ini ialah mengganti emoji dengan ekspresi yang mereka wakili dalam kata atau kalimat biasa misalnya :) dengan teks bahasa Indonesia "tersenyum". Selanjutnya, karakter khusus, URL, dan tag HTML dihapus dari teks. Tahapan akhir dari preprocessing dataset ialah melakukan tokenization dan normalisasi sebelum melakukan identifikasi objek. Bagan alir dari tahapan preprocessing dataset dapat dilihat pada Gambar 5.



Gambar 5. Data Preprocessing

Proses Case Folding adalah proses penyeragaman bentuk dengan mengubah semua huruf menjadi huruf kecil. Hasil *case folding* dapat dilihat pada Tabel 1.

Tabel 1. *Case Folding*

Tweet	*Case Folding*
RT @yusufgunawan: Bagi isoman Covid-19 di wilayah Kota Malang yang membutuhkan bantuan makanan, Dapur Arema memberikan bantuan paket makan…	rt @yusufgunawan: bagi isoman covid-19 di wilayah kota malang yang membutuhkan bantuan makanan, dapur arema memberikan bantuan paket makan…
RT @yoongioppa00: @RayChris91 Sorry ini setau aku adalah surat dari kampus yang dimaksudkan supaya kampus emang ga bertanggung jawab.	rt @yoongioppa00: @raychris91 sorry ini setau aku adalah surat dari kampus yang dimaksudkan supaya kampus emang ga bertanggung jawab.

Selanjutnya proses filtering (stopword) adalah proses pembuangan kata yang kurang penting untuk proses klasifikasi. Untuk kata yang akan di-stopword menggunakan kamus dari KBBI V online dipilih berdasarkan jenis katanya yaitu kata keterangan (adverbia), kata ganti (pronomina), kata seru (interjeksi), kata depan(preposisi), dan kata hubung (konjungsi). Proses filtering dapat dilihat pada Tabel 2.

Tabel 2. *Filtering*

Tweet	Remove Stop Words
RT @yusufgunawan: Bagi isoman Covid-19 di wilayah Kota Malang yang membutuhkan bantuan makanan, Dapur Arema memberikan bantuan paket makan…	rt @yusufgunawan: bagi isoman covid-19 wilayah kota malang membutuhkan bantuan makanan, dapur arema memberikan bantuan paket makan
RT @yoongioppa00: @RayChris91 Sorry ini setau aku adalah surat dari kampus yang dimaksudkan supaya kampus emang ga bertanggung jawab.	rt @yoongioppa00: @raychris91 sorry ini setau aku adalah surat dari kampus dimaksudkan supaya kampus emang ga bertanggung jawab

Selanjutnya proses menghilangkan karakter spesial agar mempermudah proses data pada tahap klasifikasi dapat dilihat pada Tabel 3.

Tabel 3. Menghilangkan karakter spesial

Tweet	Remove Special Character
RT @yusufgunawan: Bagi isoman covid-19 di wilayah kota malang yang membutuhkan bantuan makanan, dapur arema memberikan bantuan paket makan	bagi isoman covid-19 di wilayah kota malang yang membutuhkan bantuan makanan, dapur arema memberikan bantuan paket makan
RT @yoongioppa00: @RayChris91: Sorry ini setau aku adalah surat dari kampus yang dimaksudkan supaya kampus emang ga bertanggung jawab	sorry ini setau aku adalah surat dari kampus yang dimaksudkan supaya kampus emang ga bertanggung jawab jika

Selanjutnya, tokenisasi dan normalisasi merupakan fungsi utama dalam Natural Language Proses untuk pra-pemrosesan teks sebelum klasifikasi.

a. Tokenization mengacu pada pemisahan data teks menjadi unit-unit kecil yang disebut token. Pada penelitian ini, setiap kata diubah menjadi token. Tokenizing yaitu proses memisahkan teks menjadi kata. Hasil tokenizing dapat dilihat pada Tabel 4.

Tabel 4. Hasil *Tokenizing*

Pro	Kontra
bantuan	pedih
dukung	bangke

b. Normalisasi ialah kegiatan mengubah teks yang tidak biasa menjadi bentuk standar. Kadang kala, orang menulis sebuah kata dalam bentuk yang tidak biasa untuk mengekspresikan diri. Teks ini perlu diubah menjadi bentuk dan ejaannya yang benar. Normalisasi adalah proses yang bertujuan untuk mengkonversi kata yang tidak sesuai ejaan. Proses ini melibatkan kamus yang dibuat secara manual yang terdiri dari kamus baku dan tak baku.
c. Identifikasi objek yang berfungsi mengambil masing-masing kolom data dan memeriksa apakah data tersebut kosong. Jika kolom kosong, maka ditetapkan nilai 0 dan yang lainnya ditetapkan nilai 1 kemudian disimpan pada kolom identifikasi yang baru.

2.3. Menghitung Polaritas dan Subyektivitas

Pada dasarnya, analisis sentimen tergantung pada polaritas dan subyektivitas. Subyektivitas mengandung fakta, pendapat dan keinginan. Polaritas mengandung perasaan dan emosi. Untuk menganalisis sentimen [18], polaritas dan subyektivitas teks harus diperhitungkan. Dari data polaritas dan subjektivitas, mean, median, rata-rata minimum, rata-rata maksimum adalah dihitung untuk setiap opini pro, kontra dan netral. Polaritas rata-rata maksimum dihitung per 10 tweet. Persamaan yang digunakan dapat dilihat pada Rumus 1, 2, 3 dan 4.

$$mean, \mathrm{x} = \frac{\sum x}{n} \quad (1)$$

$$median = \frac{n+1}{2} \quad (2)$$

$$average\ minimum = \frac{(n-1)min+max}{n} \quad (3)$$

$$average\ maximum = \frac{min-(n-1)max}{n} \quad (4)$$

2.4. Klasifikasi menggunakan K-Nearest Neighbor

Klasifikasi data dilakukan terhadap skor polaritas. Jika tweet memiliki skor polaritas lebih besar dari nol (Polaritas>0) maka itu adalah tweet positif. Jika skor polaritas kurang dari nol (Polaritas<0) maka tweet tersebut masuk ke kategori negatif. Jika skor polaritas sama dengan nol (Tweet Polarity=0) maka masuk pada kategori netral. Untuk mendapatkan hasil klasifikasi pada penelitian ini digunakan algoritma KNN. Algoritma ini menerapkan kesamaan fitur di mana ia menetapkan titik data berdasarkan seberapa dekat data dengan tetangganya. Algoritma untuk KNN yang ditunjukkan pada Algoritma 1 digunakan untuk klasifikasi data.

```
Algoritma 1. K-Nearest Neighbor
Load dataset
Select the value of K
```

```
Calculate the distance between each data point
using Euclidean distance
Sort data point according to the distance
calculated
Select the top K row
Assign data point on the most frequent class
END
```

Untuk menghitung jarak titik data pada algoritma KNN jarak Euclidean adalah dihitung menggunakan Rumus 5.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(qi - pi)^2} \qquad (5)$$

Algoritma KNN menggunakan polaritas tweet untuk klasifikasi. Ini mengklasifikasikan data menjadi positif negatif dan netral. Hasil klasifikasi disimpan dan dianalisis.

3. Hasil dan Pembahasan

Data *tweet* yang berhasil dikumpulkan berjumlah 1200 data terdiri dari tweet, retweet, nama account. Data diambil dengan kata kunci "COVID-19" pada tahun 2020-2021 secara acak dan didapatkan sebanyak 1200 tweets populasi. Opini positif pada tahun 2020 berjumlah sedangkan opini negatif memiliki jumlah yang lebih banyak yaitu 389 opini dengan netral berjumlah 77 opini. Sedangkan pada tahun 2021 terjadi penurunan pada jumlah opini positif yaitu sebanyak 128 opini dengan opini negatif berjumlah 94 dan netral berjumlah 289 opini. Selanjutnya, Visualisasi tweet yang diproses akan diperlihatkan pada skor polaritas dan subyektivitas tweet. Data skor polaritas dapat dilihat pada Tabel 5.

Tabel 5. Skor Polaritas

Opini	Polaritas			
	Mean	Max	Min	Median
Pro	0.05896	1.0	-1.0	0.0
Kontra	0.12569	1.0	-1.0	0.0
Netral	0.08473	1.0	-1.0	0.0

Sedangkan untuk data skor subyektivitas dapat dilihat pada Tabel 6.

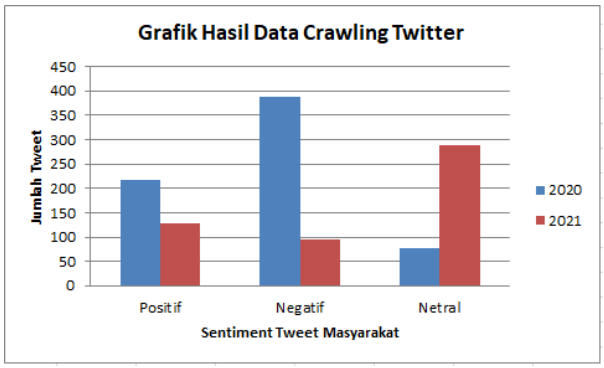
Tabel 6. Skor Polaritas

Opini	Subyektivitas			
	Mean	Max	Min	Median
Pro	0.22567	1.0	0.0	0.215
Kontra	0.29548	1.0	0.0	0.289
Netral	0.23821	1.0	0.0	0.247

Data teks diubah menjadi skor polaritas. Skor ini digunakan dalam proses klasifikasi karena proses klasifikasi KNN tidak dapat memproses data teks. Selain itu, polaritas juga menunjukkan emosi atau sentimen di balik data teks. Dalam sistem yang diusulkan, algoritma supervised klasifikasi KNN diimplementasikan. Algoritma ini mengklasifikasikan skor polaritas menjadi tiga kelas, positif, negatif dan netral. Klasifikasi dilakukan pada data opini di Twitter. Hasil akhir klasifikasi disajikan pada Tabel 7.

Tabel 7. Klasifikasi opini

Tahun	Klasifikasi Opini		
	Positif	Negatif	Netral
2020	18,08	32,41	6,41
2021	10,67	7,83	24,08



Gambar 6. Hasil Data Crawling

Pada Gambar 6 menyajikan grafik bahwa ada sentimen positif, negatif, dan netral dalam data tweet dengan berbagai opini yang berbeda.

3. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa sentiment masyarakat pada tahun 2020 mengenai COVID-19 cenderung negatif, kemudian diikuti dengan opini positif dan juga netral. Namun pada tahun 2021, seiring berjalannya waktu, opini masyarakat cenderung kearah netral. Hal ini dapat diasumsikan bahwa selama setahun masa pandemi, pemikiran masyarakat Indonesia lebih terbuka dan telah memahami bahwa pandemi COVID-19 telah menjadi bagian dari kehidupan yang tidak bisa dihindari. Dalam hal ini, twitter merupakan bagian dari platform tempat masyarakat meekspresikan perasaan dengan berbagai sudut pandang.

Daftar Rujukan

[1]. Chiang, O. (2011, January 19). *Twitter hits nearly 200M accounts, 110M tweets day, focuses on global expansion.* Forbes.http://www.forbes.com/sites/oliverchiang/2011/01/19/twitter-hits-nearly-200m-users-110m-tweets-perday-focuses-on-global-expansion. Psychological Science, 21, 372–374.

[2]. Nemes, L., & Kiss, A. (2020). *Social media sentiment analysis based on COVID-19*. Journal of Information and Telecommunication, 5(1), 1–15. doi:10.1080/24751839.2020.1790793

[3]. Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., … Alaa, M. (2021). *Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review*. Expert Systems with Applications, 167, 114155. doi:10.1016/j.eswa.2020.114155

[4]. Mertz, L. (2021). *CRISPR Tech Behind Super-Sensitive, Smartphone COVID Test*. IEEE Pulse, 12(2), 8–11. doi:10.1109/mpuls.2021.3066716

[5]. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1

[6]. Ma, X., Wang, J., Qin, H., & Wang, Y. (2020). *Public opinion analysis about The coronavirus COVID-19 based on micro-blog data*. 2020 International Conference on Information

Science and Education (ICISE-IE). doi:10.1109/icise51755.2020.00104

[7]. Evadini, S. (2022). Analisis Faktor Risiko Kematian dengan Penyakit Komorbid COVID-19 menggunakan Algoritma ECLAT. Jurnal Informasi Dan Teknologi, 52–57. doi:10.37034/jidt.v4i1.181

[8]. Phand, S. A., & Phand, J. A. (2017). *Twitter sentiment classification using stanford NLP*. 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). doi:10.1109/icisim.2017.8122138

[9]. Ikoro, V., Sharmina, M., Malik, K., & Batista-Navarro, R. (2018). *Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers*. 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/snams.2018.8554619

[10]. Ahmad, T., & Doja, M. N. (2013). *Opinion Mining Using Frequent Pattern Growth Method from Unstructured Text*. 2013 International Symposium on Computational and Business Intelligence. doi:10.1109/iscbi.2013.26

[11]. Vamshi, K. B., Pandey, A. K., & Siva, K. A. P. (2018). *Topic Model Based Opinion Mining and Sentiment Analysis*. 2018 International Conference on Computer Communication and Informatics (ICCCI). doi:10.1109/iccci.2018.8441220

[12]. Larose, D.T. (2015). *Discovering Knowledge in Data An Introduction to Data Mining*. Wiley Interscience, pp. 90-106.

[13]. Parvin, Hamid, Alizadeth, Hoseinali, and M. Behrouz. (2010). *A Modification on K-Nearest Neighborn Classifier*. Global Journal of Computer Science and Technolgy. Vol. 10 No. 14, pp. 37-41.

[14]. Okfalisa, Gazalba, I., Mustakim, & Reza, N. G. I. (2017). *Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification*. 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). doi:10.1109/icitisee.2017.8285514

[15]. Wafi, M., Faruq, U., Supianto, A., A. (2019). *Automatic Feature Selection for Modified K- Nearest Neighbor to Predict Student's Academic Performance*. Proceedings Article published Sep 2019 in 2019 International Conference on Sustainable Information Engineering and Technology (SIET). doi:10.1109/siet48054.2019.8986074

[16]. Barkur, G., Vibha, & Kamath, G. B. (2020). *Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India*. Asian Journal of Psychiatry, 51, 102089. doi:10.1016/j.ajp.2020.102089

[17]. P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in IEEE Access, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759

[18]. Sun, Xiao & He, Jiajin. (2020). A novel approach to generate a large scale of supervised data for short text sentiment analysis. Multimedia Tools and Applications. 79. 10.1007/s11042-018-5748-4