Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3569

Proyeksi Data dan Analisis Sentimen Penggunaan Vaksin di Kabupaten Indragiri Hulu Berbasis *Machine Learning*

[1]Ahmad Rizky Andriawan, [2]Mustakim

[1]Puzzle Research Data Technology (Predatech), Faculty of Science and Technology, Sultan Syarif Kasim State Islamic University Riau, Pekanbaru, Indonesia
[2]Department of Information System, Faculty of Science and Technology, Sultan Syarif Kasim State Islamic University Riau, Pekanbaru, Indonesia
Email: [1]11950314477@students.uin-suska.ac.id, [2]mustakim@uin-suska.ac.id
Email Penulis Korespondensi: mustakim@uin-suska.ac.id
Submitted: **01/12/2021**; Accepted: **22/12/2021**; Published: **31/12/2021**

Abstrak–Sentimen Analisis merupakan riset yang diproses menggunakan komputer yang berasal dari opini dan emosi yang diwujudkan dalam bentuk teks. Metode yang digunakan dalam menganalisis sentimen menggunakan *Text Mining* digunakan untuk menambang teks agar mendapatkan pemahaman terhadap suatu aspek penting. Teks Mining dilakukan pada media sosial *Twitter*. Penelitian ini menganalisis sentimen masyarakat terkait vaksinasi. Tahap *Preprocessing* berisi *Crawling Data, Cleaning, Filtering, Stemming, TF-IDF* dan Pelabelan. Hasil dari Pelabelan dan penghitungan persentase mendapatkan hasil Sentimen Positif dengan persentase sebesar 29.17%, Sentimen Negatif dengan Persentase 55.09 %, dan Netral dengan Persentase 15.74%. Terlihat bahwa masyarakat Indonesia masih banyak melontarkan komentar Negatif yang berkaitan dengan vaksinasi. Hasil penghitungan K-NN dengan K-9 menghasilkan akurasi sebesar 84.53%.

Kata Kunci: Data Mining; K-Nearest Neighbor (K-NN); Text Mining; Twitter

Abstract–Sentiment analysis is research that processed used computer that comes from opinion and emotion wich realized in shape of text. The method that used to analyze sentiment is using Text Mining used for mining the text in order to get comprehension about the important aspect. Text Mining done at social media Twitter. This research analyze public sentiment related with vaccination. Step of Preprocessing contains Crawling Data, Cleaning, Filtering, Stemming, TF-IDF, and labeling. Result from labeling and percentage calculation get percentage that Positive Sentiment is 29.17%, Negative Sentiment is 55.09% percentage, and Neutral is 15.74% percentage. Could be seen that Indonesian public still given many Negative Comments related with vaccination. Result of K-NN calculation with K-9 generate the accuracy is 84.53%.

Keywords: Data Mining; K-Nearest Neighbor (K-NN); Text Mining; Twitter

1. PENDAHULUAN

Corona Viruses Desease 2019 (Covid-19) merupakan virus yang muncul pada Desember 2019 di Wuhan, Cina. Virus tersebut kemudian menyebar ke seluruh china. Dengan meningkatnya kasus positif Covid-19 di Cina saat itu, pemerintah Cina menerapkan aturan *Lock Down* yakni tidak beraktivitas di luar seperti biasanya yang mengakibatkan terhambatnya mobilitas mereka. Namun, hal itu juga bertujuan untuk memutus penyebaran Covid-19 [1]. Tidak membutuhkan waktu yang lama, virus Covid-19 menyebar ke seluruh negara termasuk Indonesia.

Di Indonesia sendiri virus Covid-19 pertama kali dikonfirmasi pada bulan Maret 2020, Dengan jumlah kasus positif 2 orang. Namun, tidak disangka Covid-19 ini begitu cepat menyebar ke seluruh penjuru Indonesia. Dikarenakan di Indonesia belum menerapkan Protokol Kesehatan yang ketat sebagai bentuk pencegahan terhadap Covid-19. Dikutip dari laman covid19.go.id, kasus terkonfirmasi positif Covid-19 di indonesia per tanggal 28 September 2021 mencapai 4.211.460 kasus. Sedangkan sebanyak 4.031.099 kasus terkonfirmasi sembuh[2]. Hal ini menandakan virus Covid-19 sangat mudah menular, namun berpeluang tinggi untuk sembuh. Dibuktikan dengan data yang ada saat ini. Penyebaran Covid-19 di Indonesia sudah tersebar di berbagai daerah tidak terkecuali di Provinsi Riau. Dilansir dari website corona.riau.go.id, di provinsi Riau total konfirmasi Covid-19 per tanggal 28 September 2021 sebanyak 127.258 kasus. Sedangkan di Kabupaten Indragiri Hulu kasus Covid-19 per tanggal 28 September 2021 terkonfirmasi sebanyak 6233 kasus [3].

Dengan banyaknya kasus Covid-19 yang semakin parah, para peneliti melakukan pengembangan vaksin yang ditujukan sebagai antisipasi dan perlawanan terhadap Covid-19. Namun, meyakinkan masyarakat indonesia agar segera vaksin juga bukan perkara yang mudah. Hal ini disebabkan adanya keraguan, kesalahan informasi, dan berita bohong mengenai vaksinasi yang beredar luas di kalangan masyarakat [4]. Namun, dengan begitu tidak sedikit pula masyarakat yang melakukan vaksinasi atas dasar kesadaran diri sendiri.

Sentimen analisis merupakan riset yang diproses menggunakan komputer yang berasal dari opini dan emosi yang diwujudkan dalam bentuk teks [5]. Ini merupakan cara agar mendapatkan pendapat publik dengan mudah tanpa harus survei secara manual yang menghabiskan banyak biaya [5]. Salah satu metode yang digunakan dalam menganalisis sentimen menggunakan *Text Mining*. *Text Mining* digunakan untuk menambang teks agar mendapatkan pemahaman terhadap suatu aspek penting [6]. *Text Mining* dapat dilakukan pada media sosial seperti Instagram dan Twitter. Implementasi *Text Mining* Twitter sangat populer saat ini. Seperti yang telah dilakukan oleh Bunoro tahun 2017 [7] melakukan penelitian tentang sentimen masyarakat mengenai calon Gubernur DKI

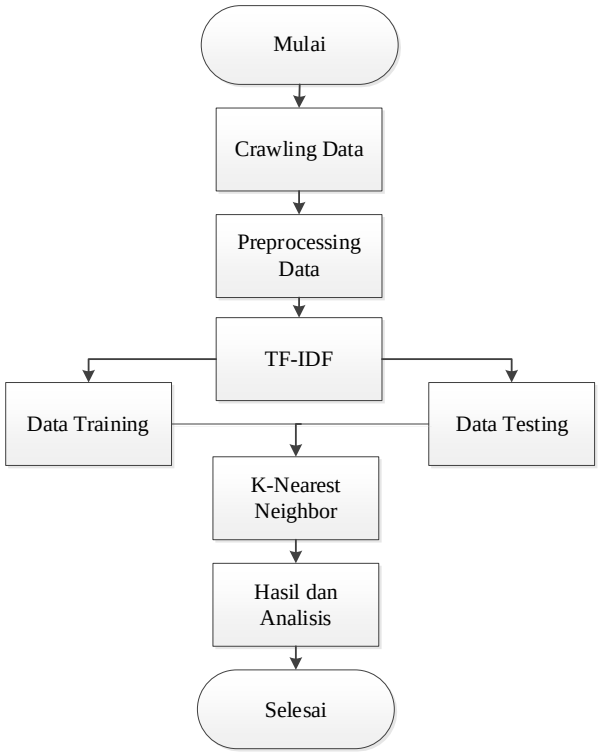
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3569

Jakarta saat itu. Menggunakan metode Klasifikasi *Naive Bayes Classifier* (NBC) yang menghasilkan tingkat akurasi sebesar 95% [7].

Penelitian lain menggunakan bahasa *Python* yang dilakukan oleh Fauziyyah tahun 2020 [8] dari pencarian terhadap kata kunci Covid-19 menghasilkan Netral sebesar 58,94% dan 55,10% untuk kata kunci *Coronavirus*. Artinya, masyarakat masih dalam batas Netral dalam beropini mengenai Covid-19 [8]. Dalam prosesnya mesin pembelajaran yang digunakan sangat beragam, salah satunya adalah algoritma K-Nearest Neighbor (K-NN), yang pernah digunakan oleh Mustakim tahun 2016 dalam menganalisis prediksi predikat prestasi Mahasiswa Sistem Informasi Universitas Islam Negeri Sultan Syarif Kasim dengan menggunakan data Mahasiswa Program Studi Sistem Informasi angkatan 2014/2015 sebagai data testing dengan jumlah 50 data, dan sebagai data training dari data angkatan 2012/2013 dengan jumlah 165 data yang menghasilkan pengujian akurasi sebesar 82% [9]. Pada penelitian ini dilakukan analisis sentimen masyarakat terkait Vaksinasi di Provinsi Riau, yang bertujuan untuk mengetahui besar atau kecilnya antusiasme masyarakat terhadap vaksinasi dikarenakan banyaknya keraguan dan berita bohong dikalangan masyarakat terhadap vaksinasi. Sehingga, menimbulkan kontroversi di masyarakat. Penelitian ini ditujukan untuk masyarakat luas sebagai evaluasi terhadap adanya vaksinasi.

2. METODOLOGI PENELITIAN

Metodologi pada penelitian ini terdapat proses utama yaitu tahap pengumpulan data melalui metode *Crawling Data*, pengolahan data, penghitungan algoritma, dan hasil serta analisis. Adapun metode penelitian yang dilakukan terdapat pada Gambar 1.



Gambar 1. Metodologi Penelitian

Pada penelitian ini data yang digunakan merupakan hasil dari proses *Crawling Data* melalui jejaring sosial Twitter dengan hasil pencarian atribut waktu, *id*, *username*, dan *text* yang berisi nama pengguna dan isi *tweet*. Kemudian dilakukan tahap *Preprocessing* berupa tahap cleaning, filtering, hingga tranformasi data. Transformasi yang dilakukan adalah dengan merubah data yang awalnya teks menjadi angka dengan metode TF-IDF. Dilanjutkan dengan proses penghitungan akurasi menggunakan Algoritma K-NN dengan menerapkan percobaan K-3 sampai dengan K-9. Tahap terakhir, hasil dan analisis berdasarkan hasil penghitungan menggunakan Algoritma.

2.1 *Data Mining dan Text Mining*

Data Mining adalah proses mengekstrak dan mendapatkan pengetahuan menggunakan suatu teknik statistik hingga *Machine Learning* yang berasal dari berbagai *Database* [9][10]. *Data Mining* adalah proses dalam menghasilkan informasi yang penting dengan cara menggali data [11]. *Data Mining* merupakan suatu teknik untuk mendapatkan pengetahuan dan pola yang berasar dari data dalam ukuran besar [12]. *Text Mining* merupakan metode yang digunakan untuk mengklasifikasikan data dari sekumpulan data dalam bentuk teks dengan jumlah besar dan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3569

merupakan turunan dari *Data Mining* [13]. *Text Mining* digunakan untuk menggali informasi dalam bentuk teks pada sebuah dokumen untuk mendapatkan informasi yang berkualitas [14]. Dalam skala besar, *Text Mining* digunakan untuk mencari dan mengolah data berupa teks yang tidak beraturan hingga menghasilkan informasi.

2.2 K-Nearest Neighbor (K-NN)

K-NN adalah salah satu algoritma strategi dengan kalkulasi *Supervised Learning* dan sering digunakan sebagai proses dalam klasifikasi data [15]. Perhitungan K-NN merupakan prosedur belajar dengan mencari kumpulan K item dalam informasi terdekat dengan item dalam informasi baru [9]. K-NN merupakan algoritma yang menghitung bobot dari suatu item baru berdasarkan item sebelumnya [11], sesuai dengan persamaan 1.

$$(x,y)=\sqrt{\Sigma}(xk-yk)2nk-1 \quad (1)$$

2.3 Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF merupakan metode yang digunakan untuk mengukur serta memberikan bobot pada data yang digunakan serta memilah kata yang sering digunakan dan menghilangkan kata yang tidak diperlukan serta mengkonversi kedalam bentuk angka [16]. Persamaan untuk menghitung TF-IDF, dapat dilihat pada persamaan 2.

$$W_{d,t} = TF_{d,t} * IDF_t \quad (2)$$

2.4 *Euclidean Distance*

Euclidean Distance merupakan metode untuk menghitung kedekatan jarak dari dua objek yang berdekatan [17]. *Euclidean Distance* merupakan sebuah distribusi yang baik untuk membuat antara perbandingan nilai yang berbeda [18]. Persamaan untuk *Euclidean Distance* dapat dilihat pada persamaan 3.

$$d_{(x,y)} = \sqrt{(\sum^{n}_{i=1}(x_i - y_i)^2)} \quad (3)$$

2.5 Vaksinasi

Vaksinasi merupakan puncak dan solusi tersendiri akibat tersebarnya Covid-19 ke penjuru dunia tak terkecuali Indonesia. Namun, di balik itu, tidak sedikit yang meragukan tentang keamanan dan dampak pasca vaksinasi [19]. Salah satu tempat membagikan pendapat tersebut terdapat pada media sosial Twitter.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Teknik pengumpulan data pada penelitian ini menggunakan teknkik *Crawling Data* yang merupakan proses untuk mendapatkan data dari *Internet* yang ditujukan untuk penelitian [20]. Dalam melakukan *Crawling Data*, penulis mencari atribut berupa waktu, *id*, *username*, dan *text* yang berisi nama pengguna dan isi *tweet* sebanyak 10567 record. Kemudian menyisakan atribut text dengan jumlah *Record* sebanyak 10000 *record*. Hasil *Crawling data* dapat dilihat pada tabel 1.

Tabel 1. Hasil *Crawling Data*

No	text
0	b’rt @republikaonline: sehingga perlu disiapka…
1	b’rt @anggita_lung: masyarakat diharapkan teru…
2	b'rt @ariestarico2: mengajak segenap masyaraka...
3	b'rt @abdulazizlatte: juru bicara @satgascovid...
...	...
10564	b'rt @polres_demak: siaga wali\nvaksinasi mala...
10565	b'rt @4ntiho4x: selain dipengaruhi oleh penera...
10566	b'akselerasi vaksinasi serentak seluruh indone...
10567	b'kominfo deteksi ribuan hoaks covid-19, vaksi...

3.2 Transformasi data

Transformasi data dilakukan dengan merubah data yang awalnya berbentuk teks menjadi angka sesuai penghitungan bobot menggunakan metode TF-IDF. Hasil tranformasi data dapat dilihat pada tabel 2.

Tabel 2. Hasil Transformasi Data

	afrika	akselerasi	aman	amp	anak	ayo	babinsa	capai	cegah	cepat
0	0	0	0	0	0.937841	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3569

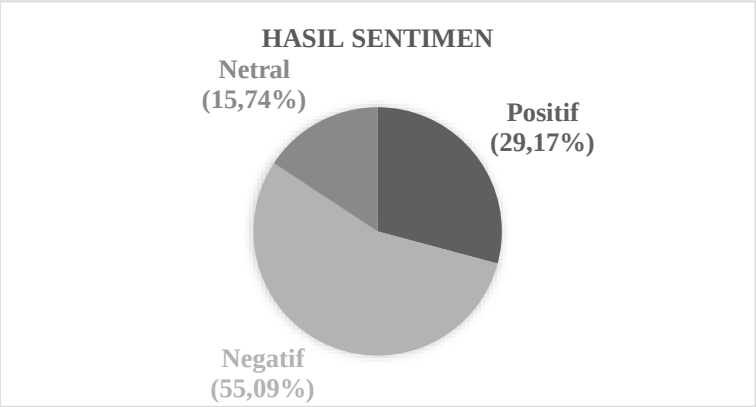
	afrika	akselerasi	aman	amp	anak	...	ayo	babinsa	capai	cegah	cepat
...	...	...	...	...	...	...	...	...	...	...	...
9996	0	0	0	0	0	0	0	0	0	0	0
9997	0	0	0	0	0	0	0	0	0	0	0
9998	0	0	0	0	0	0	0	0	0	0	0
9999	0	0	0	0	0	0	0	0	0	0	0.417794

3.3 Proses Algoritma K-NN

Semua data yang akan dihitung menggunakan algoritma K-NN berdasarkan data yang telah ditransformasi pada tabel 2. Atribut berdasarkan semua kata yang muncul dalam *tweet*. Penghitungan mendapatkan 2 hasil yakni berdasarkan persentase pembobotan pada kata yang bermakna negatif, netral, dan positif, serta hasil akurasi yang dihasilkan dari penghitungan Algoritma K-NN dengan K-3, K-5, K-7, dan K-9.

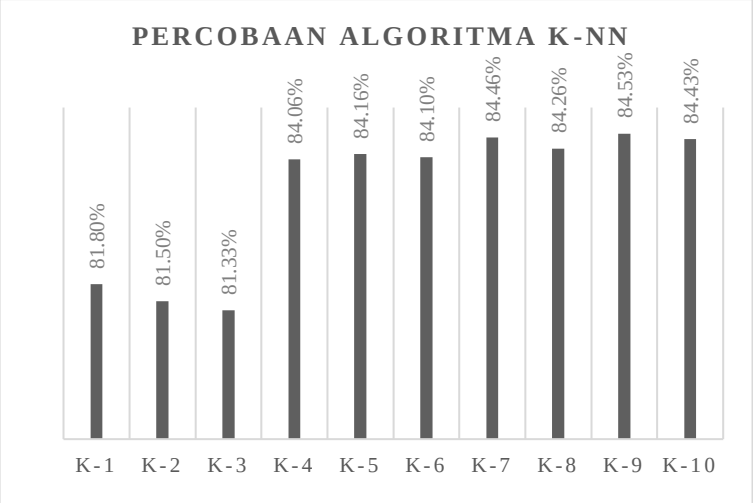
Tabel 3. Hasil penghitungan diurutkan berdasarkan Atribut Negatif, Netral, dan Positif.

	negatif	netral	positif
0	0.0	1.000000	0.000000
1	0.0	1.000000	0.000000
2	0.0	0.714286	0.285714
3	0.0	0.568224	0.431776
...	...	...	...
2996	0.0	1.000000	0.000000
2997	0.0	1.000000	0.000000
2998	0.0	1.000000	0.000000
2999	0.0	0.857143	0.142857



Gambar 2. Perbandingan nilai sentimen positif, negatif, dan netral.

Gambar 2 merupakan visualisasi berdasarkan hasil penghitungan semua *dataset* dengan atribut Positif, Negatif, dan netral berdasarkan Tabel 3. Yang mana tingkat persentase tertinggi pada atribut Negatif dengan persentase 55,09%, atribut Netral dengan persentase terendah sebesar 15,74%, dan atribut Positif dengan persentase 29,17%.



Gambar 3. Grafik Akumulasi Percobaan Algoritma K-NN K-1 sampai dengan K-10

Berdasarkan hasil penghitungan pada Gambar 2, terlihat bahwa opini dari masyarakat indonesia masih banyak mengandung suatu hal negatif, yang mana cukup tinggi dibandingkan dengan sentimen positif dan sentimen netral dengan persentase paling sedikit. hal ini dapat menunjukan bahwa masyarakat indonesia masih banyak yang membuat kata-kata dengan konotasi negatif yang berkaitan dengan vaksinasi. Tentu saja hal tersebut tidak bersifat mutlak. Karena, sistem menghitung berdasarkan kata per kata. Sehingga ada kemungkinan bahwa kata yang dinilai negatif oleh sistem, sebenarnya bermaksud positif.

Kemudian dilakukan penghitungan Algoritma K-NN untuk menghitung akurasi. Percobaan dilakukan dengan menggunakan K-1 sampai dengan K-10, menghasilkan tingkat akurasi terendah pada percobaan K-3 sebesar 81,33%. Sedangkan, akurasi tertinggi pada percobaan K-9 sebesar 84,53%. Hasil tersebut lebih tinggi dibandingkan penelitian sebelumnya yang menghasilkan akurasi sebesar 82%.

4. KESIMPULAN

Berdasarkan penelitian yang dilakukan hasil dari penghitungan persentase berdasarkan label yang diperoleh mendapatkan hasil sentimen Positif dengan persentase sebesar 29.17%, sentimen Negatif dengan persentase 55.09 %, dan Netral dengan persentase 15.74%. Sedangkan dalam penghitungan dengan menggunakan Algoritma K-NN dengan k=7, menghasilkan akurasi sebesar 84.46%.

5. UCAPAN TERIMAKASIH

Ucapan terimakasih kepada Puskesmas Kilan Kec. Batang Cenaku yang telah memberikan kesempatan untuk melakukan penelitian. Keluarga besar *Puzzle Research Data Technology* selalu memberikan masukan, dorongan dan semangat dalam penyelesaian penelitian ini.

REFERENCES

[1] S. Sindi *et al.*, "ANALISIS ALGORITMA K-MEDOIDS CLUSTERING DALAM PENGELOMPOKAN PENYEBARAN COVID-19 DI INDONESIA," *Jurnal Teknologi Informasi*, vol. 4, no. 1, 2020.

[2] Satuan Tugas penanganan COVID-19, "Peta Sebaran Covid-19," *Satuan Tugas penanganan COVID-19*, 2021. https://covid19.go.id/peta-sebaran-covid19 (accessed Sep. 28, 2021).

[3] Pemerintah Provinsi Riau, "Data & Statistik," *Pemerintah Provinsi Riau*, 2021. https://corona.riau.go.id/data-statistik (accessed Sep. 28, 2021).

[4] N. P. Astuti, E. G. Z. Nugroho, J. C. Lattu, I. R. Potempu, and D. A. Swandana, "Persepsi Masyarakat terhadap Penerimaan Vaksinasi Covid-19: Literature Review," *Jurnal Keperawatan*, vol. 13, no. 3, pp. 569–580, Jul. 2021, doi: 10.32583/KEPERAWATAN.V13I3.1363.

[5] I. Zulfa and E. Winarko, "Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 11, no. 2, 2017, doi: 10.22146/ijccs.24716.

[6] T. Zhang, B. Li, and N. Hua, "Chinese cultural theme parks: text mining and sentiment analysis," *Journal of Tourism and Cultural Change*, 2021, doi: 10.1080/14766825.2021.1876077.

[7] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," 2017.

[8] A. K. Fauziyyah, "ANALISIS SENTIMEN PANDEMI COVID19 PADA STREAMING TWITTER DENGAN TEXT MINING PYTHON," *Jurnal Ilmiah SINUS*, vol. 18, no. 2, 2020, doi: 10.30646/sinus.v18i2.491.

[9] Mustakim and G. Oktaviani, "Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa," *Jurnal Sains, Teknologi dan Industri*, vol. 13, no. 2, pp. 195–202, 2016, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/sitekin

[10] I. Kamila, U. Khairunnisa, and Mustakim, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 5, no. 1, pp. 119–125, 2019.

[11] R. Harun, K. Chandra Pelangi, and Y. Lasena, "PENERAPAN DATA MINING UNTUK MENENTUKAN POTENSI HUJAN HARIAN DENGAN MENGGUNAKAN ALGORITMA K NEAREST NEIGHBOR (KNN)," Online, 2020. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/misi

[12] A. Ramadhan, Z. Efendi, and Mustakim, "Perbandingan K-Means dan Fuzzy C-Means untuk Pengelompokan Data User Knowledge Modeling," Pekanbaru, May 2017.

[13] T. W. D. Sari, "Penerapan Text Mining Dengan Menggunakan Algoritma TF-IF Untuk Klasifikasi Genre Novel," *Pelita Informatika: Informasi dan Informatika*, vol. 10, no. 1, pp. 29–37, Aug. 2021, Accessed: Dec. 24, 2021. [Online]. Available: http://www.stmik-budidarma.ac.id/ejurnal/index.php/pelita/article/view/3142

[14] Asril Hamdani, Mustakim, and Kamila Insanul, "Klasifikasi Dokumen Tugas Akhir Berbasis Text Mining menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor," 2019.

[15] A. Hidayat, M. Zakiy Fauzi, Syukra Imaduddin, and Mustakim, "Implementasi Algoritma K-Nearest Neighbor dan Probabilistic Neural Network untuk Analisis Opini Masyarakat Terhadap Toko Online di Indonesia," 2019.

[16] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications*, vol. 181, no. 1, 2018, doi: 10.5120/ijca2018917395.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3569

[17] A. M. Habibi and R. R. Santika, "Implementasi Algoritma K-Nearest Neighbor dalam Menentukan Jurusan Menggunakan Metode Euclidean Distance Berbasis Web Pada SMP Setia Gama," vol. 3, pp. 7–14, 2020.
[18] Dmitry Vasilyev and Andrey Rashich, *SEFDM-signals Euclidean Distance Analysis*. IEEE, 2018.
[19] Abidin Yumetri and Iska Zikri Neni, "SOSIALISASI VAKSINASI COVID-19 TERHADAP KERAGUAN MASYARAKAT ATAS EFEK NEGATIF PASCA VAKSIN," 2020.
[20] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, "Data Crawling Otomatis pada Twitter," 2016. doi: 10.21108/indosc.2016.111.