

Sentiment Analysis Covid-19 Vaccination on Twitter Social Media Using Naïve Bayes Method

Analisis Sentimen Masyarakat Terhadap Tindakan Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes

Dihin Muriyatmoko¹, Triana Harmini², Maulana Kemal Ardiansyah³
{dihin@unida.gontor.ac.id¹, triana@unida.gontor.ac.id², maulanakemal@mhs.unida.gontor.ac.id³}

Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Darussalam Gontor Ponorogo

Abstract. *The government regulations regarding the implementation of vaccinations to tackle the COVID-19 pandemic. The regulation was issued by the Minister of Health Number ten of 2021. This program raises pros and cons so that it requires feedback for evaluation. Feedback can be obtained from opinions and stories that users convey through social media such as Twitter. This study aims to develop a model to determine public sentiment towards Covid-19 vaccination in three topics, namely the vaccination program, the effect of vaccination and the Covid-19 vaccine. The classification method used in this research is Bernoulli Naïve Bayes and Logistic Regression. The results of the comparison of the two methods show that Bernoulli Naïve Bayes gets better accuracy results. The number of tweet messages processed from Twitter is 5877. The model was tested to read public sentiment on Twitter from 7 September to 21 September 2021. The model concluded that public opinion regarding the vaccination program and the effect of vaccination tended to be positive. And opinions regarding the Covid-19 vaccine topic tend to be neutral. For further research, it can be developed by adding datasets.*

Keywords - covid-19 vaccination; logistic regression method; naïve bayes method; sentiment analysis; social media

Abstrak. *Pemerintah membuat peraturan tentang pelaksanaan vaksinasi guna menanggulangi pandemi covid 19. Peraturan tersebut dikeluarkan oleh Menteri Kesehatan Nomor sepuluh tahun 2021. Program ini menimbulkan pro dan kontra sehingga membutuhkan umpan balik untuk evaluasi. Umpan balik bisa diperoleh dari pendapat dan cerita yang pengguna sampaikan melalui Media sosial seperti Twitter. Penelitian ini bertujuan untuk mengembangkan model dalam mengetahui sentimen publik terhadap Vaksinasi Covid-19 di tiga topik yaitu program vaksinasi, efek vaksinasi dan vaksin Covid-19. Metode klasifikasi yang digunakan dalam penelitian ini adalah Bernoulli Naïve Bayes dan Logistic Regression. Hasil komparasi kedua metode menunjukkan bahwa Bernoulli Naïve Bayes mendapatkan hasil akurasi lebih baik. Jumlah pesan tweet yang diolah dari Twitter adalah 5877. Model diuji untuk membaca sentimen public di twitter sejak tanggal 7 September sampai dengan 21 September 2021. Model menghasilkan kesimpulan bahwa opini publik terkait program vaksinasi dan efek vaksinasi cenderung positif. Dan opini terkait topik vaksin Covid-19 cenderung netral. Untuk penelitian selanjutnya bisa dikembangkan dengan penambahan dataset.*

Kata Kunci – vaksinasi covid-19; metode logistic regression; metode naïve bayes; analisis sentimen; sosial media

I. PENDAHULUAN

Jumlah kasus COVID-19 Hingga 21 September 2021 di dunia terdata mencapai 229.734.000 kasus. Pada awal kemunculannya kasus tersebut paling banyak terdapat di Cina, namun seiring berjalannya waktu kasus tersebut mulai menyebar ke negara-negara lain. Hingga saat ini kasus paling banyak terdapat di Amerika Serikat dengan 43.054.107 kasus, diikuti oleh India dan Brasil dengan masing masing puluhan juta kasus. Sedangkan di negara kita ini terkonfirmasi pada tanggal tersebut mencapai 4.192.695 kasus dengan tingkat kematian yang cukup tinggi pada angka 140.634 jiwa[1]. Hingga kini pandemi COVID-19 belum juga berakhir, sehingga pemerintah mengeluarkan kebijakan vaksinasi terhadap seluruh masyarakat di Indonesia. Peraturan tersebut dikeluarkan oleh Menteri Kesehatan Nomor sepuluh tahun 2021 guna menanggulangi pandemi COVID-19 [2].

Dengan adanya kebijakan yang dikeluarkan oleh pemerintah maka munculah berbagai tanggapan masyarakat, tanggapan tersebut diungkapkan dalam berbagai media sosial. Salah satu media yang digunakan oleh kebanyakan masyarakat adalah media sosial *Twitter*. *Twitter* merupakan salah satu media sosial yang gemar digunakan oleh masyarakat dalam beropini. Setiap individu dalam masyarakat memiliki tanggapan dan opini yang berbeda beda. Oleh karena itu media sosial *Twitter* menjadi peluag sumber data yang besar dalam kasus analisis sentimen terhadap masyarakat. Cerminan perilaku seseorang bisa terlihat dari opini yang ia ungkapkan. Perilaku juga dapat mempengaruhi seseorang dalam beropini. Sentiment analisis merupakan bidang studi yang mempelajari mengenai analisa terhadap opini, sikap, emosi dan penilaian kepada seorang atau pelayanan publik maupun peristiwa yang sedang menimpa masyarakat itu sendiri [3].

Beberapa penelitian terkait analisis sentimen telah dilakukan, salah satunya analisis sentimen publik pada program kesehatan masyarakat. Penelitian tersebut di laksanakan pada tanggal 3 Agustus 2018 hingga 9 Agustus 2018. Beberapa kata kunci terkait program kesehatan digunakan dalam penelitian ini. Data yang berhasil dikumpulkan berjumlah 1000 data berasal dari sosial media *twitter*. Hasil yang didapatkan dalam penelitian ini adalah tingginya tren sentimen negatif yang terdapat pada kata kunci stunting kemudian gizi buruk dan disusul oleh imunisasi. Metode yang digunakan *leksikon dengan tingkat presentase akurasi sebesar 68% hingga 77%*. Tetapi pada penelitian ini tidak langsung diambil kesimpulan bahwasanya program yang memiliki sentimen negatif tinggi, benar benar program itu buruk.[4].

Pada penelitian lain membandingkan antara dua metode klasifikasi yaitu metode *K-Nearest Neighbor (KNN)* dengan salah satu metode klasifikasi yang sering digunakan yaitu *Naïve Bayes*. Data yang digunakan sebanyak 1098 data, opini tersebut diambil dari media sosial *twitter*. Pada penelitian ini berfokus pada perbandingan antara kedua metode yang digunakan dan juga kecenderungan sentimen masyarakat terhadap COVID-19. Hasil dari penelitian tersebut adalah lebih besarnya masyarakat yang bersentimen positif terhadap COVID-19 dengan 122 opini lebih banyak. Pada penelitian ini mendapatkan hasil bahwasanya metode *Naïve Bayes* mendapatkan akurasi lebih besar di banding dengan KNN. Tingkat akurasi yang didapatkan pada metode *Naïve Bayes* sebesar 63.21%, dan KNN sebesar 58.10%. Maka pada penelitian ini metode yang di gunakan untuk mengklasifikasi adalah metode *Naïve Bayes* karena mendapatkan akurasi lebih tinggi.[5].

Pada penelitian terkait analisis sentimen terhadap tindakan vaksinasi dalam upaya mengurangi jumlah kasus terinfeksi Covid-19. Pada penelitian tersebut menghasilkan akurasi sebesar 85,59% dengan menggunakan metode *Naïve Bayes* dan sedangkan metode SVM mendapatkan akurasi 84,41%. Pada penelitian ini mendapatkan hasil bahwasanya metode *Naïve Bayes* mendapatkan akurasi lebih besar di banding SVM pada kasus vaksinasi dengan data yang cenderung positif. Data yang digunakan perkata kunci yaitu kata kunci vaksin *sinovac* 253 data dan vaksin merahputih 253 data.[6]

Dalam penelitian ini akan membuat model yang dapat melihat Tren sentiment pengguna twitter terhadap tiga kata kunci yaitu program vaksinasi, efek vaksin dan vaksin covid 19 pada tanggal 7 September 2021 sampai dengan 21 September 2021, dengan pengelompokan opini negatif, positif dan netral. Pada penelitian ini berfokus pada perbandingan antara kedua metode yang digunakan yaitu metode klasifikasi *Logistic Regression* dan *Naïve Bayes*. Serta mengetahui kecenderungan sentimen masyarakat pada tanggal tersebut terhadap tiga kata kunci pada media sosial *twitter*.

II. METODE

Penelitian ini memiliki tujuan utama yaitu membangun model yang dapat di gunakan untuk mengklasifikasi sentimen pada opini masyarakat di *twitter* dengan perbandingan hasil akurasi antara metode *Naïve Bayes* dan *Logistic Regression*. Model terbaik akan digunakan untuk menganalisa sentimen pengguna *twitter* terhadap tiga kata kunci yaitu program vaksinasi, efek vaksinasi dan vaksin covid 19 pada tanggal 7 September 2021 sampai dengan 21 September 2021.

Tabel 1. Jumlah data per kata kunci

No	Kata Kunci	Jumlah Data
1	Vaksin covid 19	5547
2	Program vaksinasi	4326
3	Efek vaksinasi	4916

A. Pengumpulan data

Pada penelitian ini memanfaatkan *twitter* API dalam memperoleh data, dikarenakan data yang akan diambil adalah data dari media sosial *twitter*. Penelitian ini juga mengambil komentar dari video mengenai vaksinasi yang tersedia di media sosial *YouTube*. Data yang digunakan untuk membuat model diambil pada tanggal 25 Juni sampai 13 Agustus 2021, data berjumlah 5877. Kemudian data yang akan dianalisa diambil dari *twitter* pada tanggal 7 September sampai dengan 21 September 2021, menggunakan 3 kata kunci yang berjumlah 14.789 dengan rincian pada Tabel 1.

Tabel 2. Jumlah data per kategori kelas

No	Nama Kelas	Jumlah Data
1	Negatif	1932
2	Positif	1939
3	Netral	2006

B. Pelabelan data

Pelabelan data yang akan dibuat sebagai model dilakukan secara manual terhadap 5877 data dalam tiga kategori sentimen negatif, positif dan netral dengan jumlah data pada Tabel 2. Data yang akan masuk kedalam kategori negatif yaitu data tweet yang mengandung pernyataan kontroversi terhadap vaksinasi. Data netral adalah data tweet yang hanya berisi informasi atau berita. Data positif adalah data tweet yang menyatakan dukungan terhadap vaksinasi. Setelah data diberi label secara manual maka data harus melalui proses validasi ahli pada penelitian ini data divalidasi oleh guru Bahasa Indonesia.

C. Pre-processing

Pada tahap ini data mentah diubah menjadi data yang mudah dimengerti oleh mesin pembelajaran. Pada tahap ini juga memecahkan berbagai jenis masalah termasuk data *noise*, data yang tidak memiliki *value* dan data yang tidak konsisten, dengan tujuan membuat data lebih terstruktur. Adapun langkah-langkah data *pre-proccesing* [7].

- a) *Case Folding*, bertujuan membuat teks menjadi konsisten dan terstruktur dengan merubah teks yang ada menjadi huruf kecil, menghilangkan *username*, *hashtag*, angka, tanda baca dan *Uniform Resource Locator* (URL)
- b) *Tokenizing* adalah tahapan dimana proses pemisahan data yang berupa text menjadi perkata maupun diubah menjadi perkalimat agar bisa masuk ke tahapan Filtering, Pada penelitian ini menggunakan pemisahan perpotongan kata.
- c) *Filtering* adalah tahapan yang ditujukan untuk mengambil kata kata yang memiliki nilai dari hasil proses *tokenizing*. Algoritma *stoplist* digunakan untuk menghilangkan kata kata yang tidak meemiliki makna didalamnya. Sedangkan algoritma *wordlist* digunakan untuk menyimpan kata kata yang meemiliki makna didalamnya. *Stopword* merupakan sebutan bagi kata kata yang tidak memiliki makna dan biasanya muncul dalam jumlah besar. Contoh *stopword* yang ada dalam bahasa Indonesia pada *library* Sastrawi adalah “tapi”, “juga”, “dan”, “dari”, “ini”, “akan”, “oleh”, dll.
- d) *Stemming*, untuk mengkonversi kata yang telah melalui proses *filtering* ke dalam kata dasar yang tidak memiliki imbuhan. Misalnya pada kata “melakukan”, “berlaku”, “dilakukan” akan dikonversi menjadi kata “laku”. Dalam melakukan proses pada tahapan stemming pada kata berbahasa Indonesia kita bisa menggunakan *library* yang ada *library* tersebut bernama Sastrawi.

D. Pembobotan tf/idf

Dalam proses pembobotan yang disebut TF/IDF (*Term Frequency-Inverse Document Frequency*) didalamnya terdapat proses statistik berupa angka untuk memberi nilai betapa pentingnya sebuah kata bagi kategori tertentu maupun dokumen tertentu dalam korpus. *Term Frequent* (TF) adalah ukuran seberapa sering istilah, t, muncul dalam dokumen, "d". Di sini, di pembilang, n adalah berapa kali istilah "t" muncul dalam dokumen "d". Dengan demikian, setiap dokumen dan istilah akan memiliki nilai TF sendiri[8]. Pada penelitian ini peneliti melihat penelitian sebelumnya [9] yang membahas tentang klasifikasi emosi dengan membandingkan beberapa metode pembobotan antara lain N-Gram, BOW (*Bag-of-Words*) dan TF/IDF dengan kesimpulan pembobotan menggunakan TF/IDF lebih baik, oleh karena itu pada penelitian ini pembobotan dilakukan dengan metode TF-IDF.

E. Klasifikasi

Proses klasifikasi dilakukan menggunakan algoritma *Naïve Bayes* sebagai metode utama. Metode ini dapat memprediksi peluang suatu data terklasifikasi ke kelas tertentu. *Logistic Regression* sebagai metode pembanding yang dapat menjelaskan hubungan antara peubah penjelas dengan respon berupa biner maupun kategorik. Proses klasifikasi dapat dilaksanakan apabila data yang digunakan telah melalui proses pembobotan dan pada penelitian ini menggunakan TF/IDF.

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Gambar 1. 10 K-Fold

F. Validasi

Pada penelitian ini proses validasi model yang dibuat menggunakan metode *k-fold Cross Validation*. Tahapan *K-fold* berfungsi untuk mengacak data dengan membagi data kedalam beberapa bagian sebagai data uji untuk menemukan rata-rata akurasi dari model yang dibuat. Pada penelitian ini *k-fold* akan dilakukan sebanyak sepuluh kali dengan urutan bertingkat seperti Gambar 1 di mana setiap kelas terdiri dari 10% dari keseluruhan data.[10]

III. HASIL DAN PEMBAHASAN

Pengklasifikasian data *tweet* kedalam kelas kelas yang telah ditentukan, dalam tiga pembagian kelas yaitu positif, negatif dan netral. Sebelum mengklasifikasikan data, data harus melewati proses *preprocessing*. *preprocessing* terdiri dari *case folding*, *tokenizing*, *filtering* dan *stemming*. Dalam proses tersebut menjadikan data dapat tersusun dan mudah dimengerti oleh mesin pembelajaran. Setelah keseluruhan data yang didapat terkumpul dan dikelompokkan berdasarkan class yang telah ditentukan. Maka bisa melanjutkan ke tahapan penerapan metode algoritma klasifikasi. Melatih model menggunakan data merupakan langkah terpenting dalam proses pembuatan model yang akan digunakan untuk pengklasifikasian data. Setelah data melalui tahapan *preprocessing* maka data digunakan untuk melatih model menggunakan software *jupyter notebook (anaconda3)* dan *Python 3.8.8*. Model yang telah didapatkan akan digunakan untuk mengklasifikasikan data ke dalam tiga kelas tersebut. Pembagian data latihan dan testing sebesar 0.8 untuk data training dan 0.2 data testing.

A. Pembuatan model utama menggunakan metode *naïve bayes*

Metode Klasifikasi *BernoulliNB (Bernoulli Naïve Bayes)* digunakan untuk membuat model yang dapat menentukan apakah tweet masuk ke dalam salah satu kelas dari tiga kelas tersebut. Model ini memanfaatkan nilai probabilitas dari setiap tweet untuk masuk kedalam kategori kelas yang telah ditentukan. Data yang telah melewati proses klasifikasi divalidasi menggunakan *k-fold Cross Validation*. Metode ini mendapatkan tingkat akurasi sebesar 0.831 dengan rincian *precision*, *f1-score* dan *recall* pada tabel 3. Untuk nilai rata-rata *k-fold Cross Validation* sebanyak 10 kali pada metode *BernoulliNB*, diperoleh persentase sebesar 0.81 dengan rincian pada Tabel 6.

Tabel 3. Rincian f1-score, precision dan recall

Kelas	F1-Score	Precision	Recall
Negatif	0.84	0.82	0.86
Netral	0.83	0.89	0.83
Positif	0.79	0.78	0.79
AVG	0.83	0.83	0.83

B. Pembuatan model perbandingan menggunakan metode *logistic regression*

Metode Klasifikasi *Logistic Regression* digunakan untuk membuat model yang dapat menentukan apakah tweet masuk ke dalam salah satu kelas dari tiga kelas yang telah ditentukan. Data yang telah melewati proses klasifikasi divalidasi menggunakan *k-fold Cross Validation*. Model ini mendapatkan tingkat akurasi sebesar 0.817 dengan rincian *precision*, *f1-score* dan *recall* pada tabel 4. Untuk nilai rata-rata *k-fold Cross Validation* sebanyak 10 kali pada metode *Logistic Regression*, sebesar 0.82 dengan rincian pada Tabel 6.

Tabel 4. Rincian f1-score, precision dan recall

Kelas	F1-Score	Precision	Recall
Negatif	0.82	0.81	0.83
Netral	0.84	0.82	0.85
Positif	0.79	0.82	0.76
AVG	0.82	0.82	0.82

C. Analisis data menggunakan metode *naïve bayes*

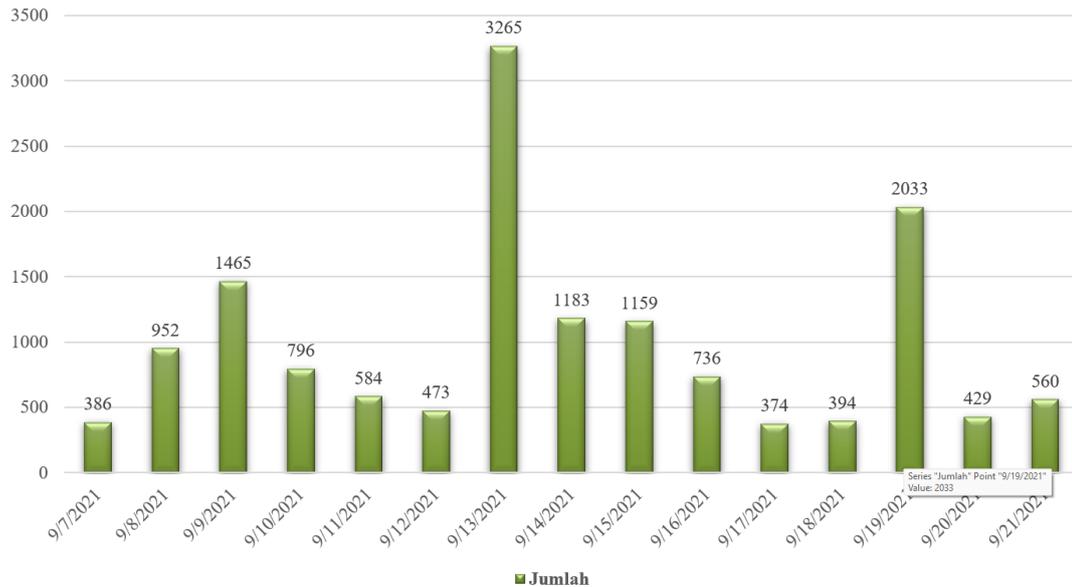
Setelah memperoleh metode *BernoulliNB* yang akurasinya lebih tinggi dibanding *Logistic Regression* maka langkah selanjutnya adalah analisa sentimen data tweet pada tanggal 7 September sampai dengan 21 September 2021, menggunakan 3 kata kunci yang berjumlah 14.789 data riil dengan rincian pada Tabel 1 menggunakan metode algoritma *BernoulliNB*.

Penyebaran tren data tweet per hari pada tiga kata kunci yaitu program vaksinasi, efek vaksin dan vaksin covid 19 pada tanggal 7 September 2021 sampai dengan 21 September 2021 dengan rincian pada Gambar 2 dan hasil klasifikasi sentimen positif, negatif dan netral perhari pada tanggal tersebut terinci pada Gambar 3.

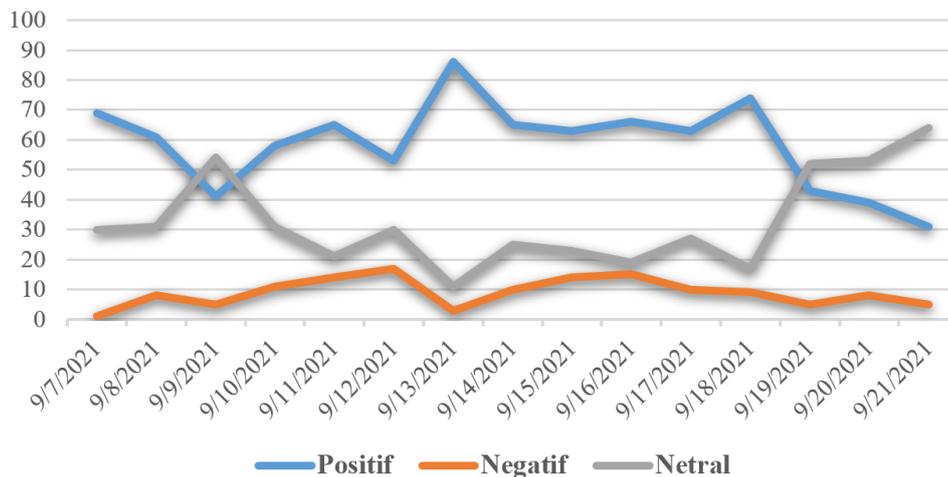
Tabel 5. Tren sentimen data per kata kunci

Kata Kunci	Negatif	Netral	Positif
Vaksin covid 19	6%	79%	15%
Program vaksinasi	1%	18%	81%
Efek vaksinasi	14%	4%	82%
AVG	7%	36%	57%

Untuk persentase sentimen pengguna twitter terhadap tiga kata kunci yaitu program vaksinasi, efek vaksin dan vaksin covid 19 pada tanggal 7 September 2021 sampai dengan 21 September 2021, terdapat pada Tabel 5 dengan rincian persentase didalamnya.



Gambar 2. Tren jumlah data per hari pada tanggal 7 - 21 September 2021.



Gambar 3. Tren sentimen data per hari pada tanggal 7 - 21 September 2021.

D. Hasil analisis sentimen

Untuk persentase sentimen pengguna *twitter* terhadap tiga kata kunci yaitu program vaksinasi, efek vaksin dan vaksin covid 19 pada tanggal 7 sampai dengan 21 September 2021 didominasi oleh sentimen positif dengan persentase 57%. Tren Jumlah data paling tinggi terjadi pada tanggal 13 September yang di barengi dengan tingginya tren sentimen positif. Pada tanggal 19 September tren sentimen netral mulai mengalami kenaikan sampai tanggal 21 September.

Tren sentimen pada kata kunci program vaksinasi dan efek vaksinasi di dominasi oleh sentimen positif. Sedangkan pada kata kunci vaksin covid 19 didominasi dengan tren sentimen netral dengan persentase 79%. Akan tetapi apabila dirata-ratakan antara tiga kata kunci tersebut maka pada sentimen netral mendapatkan persentase sebesar 36%.

Pada sentimen negatif persentase paling tinggi terdapat pada kata kunci efek vaksin dengan persentase 14%. Apabila dibandingkan dengan kata kunci program vaksinasi yang hanya 1% dan vaksin covid 19 yang hanya 6%. Oleh karena itu apabila di rata-ratakan presentasi *tweet* bersentimen negatif pada tiga kata kunci tersebut hanya 7% pada tanggal 7 September 2021 sampai dengan 21 September 2021.

Tabel 6. Rincian K- Fold

K-Fold	BernoulliNB	LogisticR
1	0.79	0.80
2	0.81	0.78
3	0.81	0.83
4	0.82	0.82
5	0.82	0.82
6	0.82	0.82
7	0.80	0.83
8	0.77	0.83
9	0.81	0.81
10	0.79	0.81
AVG	0.81	0.82

IV. KESIMPULAN

Bedasarkan pengujian menunjukkan bahwa algoritma *BernoulliNB* lebih baik dibandingkan *Logistic Regression* dalam kasus analisis sentimen terhadap vaksinasi Covid-19. Hal tersebut ditunjukkan dengan tingkat akurasi yang didapatkan dengan algoritma *BernoulliNB* sebesar 0.831 dan *Logistic Regression* sebesar 0.82.

Tren Sentimen pada 7 September sampai dengan 21 September 2021 didominasi oleh sentimen positif. Sedangkan sentimen negatif mengalami penurunan pada 13 September dan mengalami kenaikan pada 16 September. Tren sentimen pada kata kunci vaksin Covid-19 didominasi dengan tren sentimen netral dengan persentase 79%, sedangkan pada kata kunci program vaksinasi dan efek vaksinasi di dominasi oleh sentimen positif.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh Program Studi Teknik Informatika Universitas Darussalam Gontor, Ponorogo Indonesia.

REFERENSI

- [1] T. Shelavie, "Update Corona Global 21 September 2021: Indonesia Urutan 32 Penambahan Kasus Terbanyak Dunia," *Tribunnews*, 2021. <https://www.tribunnews.com/corona/2021/09/21/update-corona-global-21-september-2021-indonesia-urutan-32-penambahan-kasus-terbanyak-dunia>.
- [2] Kementerian Kesehatan RI, "Pmk No 10 Tentang Pelaksanaan Vaksinasi Dalam Rangka Penanggulangan Pandemi Corona Virus Disease 2019 (Covid-19)," 2021.
- [3] Bing Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.
- [4] A. Sasmito Aribowo, "Analisis Sentimen Publik pada Program Kesehatan Masyarakat menggunakan Twitter Opinion Mining," in *Seminar Nasional Informatika Medis (SNIMed)*, 2018, vol. 1, no. 1, pp. 17–23, [Online]. Available: <https://journal.uui.ac.id/snimed/article/view/11877>.
- [5] M. Syarifuddin, "Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn," *Inti Nusa Mandiri*, vol. 15, no. 1, pp. 23–28, 2020.
- [6] C.- Pandemic, B. Laurensz, and E. Sedyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," vol. 10, no. 2, pp. 118–123, 2021.
- [7] S. Mujilawati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.
- [8] M. Nurjannah and i. Fitri astuti, "Penerapan Algoritma Term Frequency-Inverse Document Frequency (Tf-Idf) Untuk Text Mining," *j. Inform. MULAWARMAN*, VOL. 8, NO. 3, PP. 110–113, 2013.
- [9] O. V. Putra, F. M. Wasmanson, T. Harmini, and S. N. Utama, "Sundanese Twitter Dataset for Emotion

- Classification,” *CENIM 2020 - Proceeding Int. Conf. Comput. Eng. Network, Intell. Multimed. 2020*, no. Cenin 2020, pp. 391–395, 2020, doi: 10.1109/CENIM51130.2020.9297929.
- [10] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Feb. 2016, no. Cv, pp. 78–83, doi: 10.1109/IACC.2016.25.