Website : http://pilar.unmermadiun.ac.id/index.php/pilarteknologi

Analisis Sentimen Komentar Youtube Mengenai Vaksin Covid-19 Menggunakan Support Vector Machine

Ardhi Wijayanto[1], Afrima Dhia Defara[2]

[1,2]*Program Studi Informatika, Universitas Sebelas Maret, Jln. Ir. Sutami no. 36 A, Surakarta, 57126*
E-mail: [1]ardhi.wijayanto@staff.uns.ac.id, [2]afrima47@student.uns.ac.id

Abstract— **Vaccination is one of the policies taken by the Indonesia government to control the Covid-19 pandemic. Vaccination is expected to form community immunity to prevent the spread of Covid-19. This policy taken by the government has generated various responses from the public in the form of opinions posted on social media platforms, one of which is on the Youtube. In order to better understand public sentiment towards this vaccination policy, it is necessary to analyze the opinions given. The Support Vector Machine (SVM) algorithm used in this study to classify public sentiment on vaccination policy. The data sources used are comments posted on Youtube videos regarding Covid-19 vaccinations. Three scenarios were used in the test with variations in the ratio of the testing data and training data of 90:10, 80:20, and 70:30 with 10 tests on each variation of the data ratio. Based on the tests carried out, the results obtained accuracy of 86%, 85%, and 84%.**

Keywords—: **Covid-19; sentiment analysis; Support Vector Machine; vaccine.**

I. PENDAHULUAN

Youtube tidak hanya menjadi platform bagi para pengguna yang menjadi *content creator* namun juga sebagai tempat bagi pengguna lain saling bertukar pikiran melalui kolom komentar. Selain fitur komentar, Youtube juga menyediakan fitur rekomendasi yang membantu pengguna untuk menemukan video yang sesuai dengan minat pengguna. Sebagai contoh, Youtube menggunakan halaman depan untuk menayangkan video yang direkomendasikan dan mendapatkan highlight mengenai peristiwa tertentu yang dicari banyak orang (Zhou, Khemmarat, Gao, Wan, & Zhang, 2016)

Salah satu video yang banyak dibicarakan saat ini yaitu berita mengenai perkembangan vaksinasi Covid-19 yang ada di Indonesia. Pada akhir bulan November 2021, Indonesia telah mencapai *vaccine rate* sebesar 34% dari seluruh penduduk di Indonesia (Ritchie et al, 2020). Jumlah tersebut menunjukkan kurang lebih sebanyak 95 juta penduduk dari 233 juta penduduk telah melakukan vaksinasi. Sebagai informasi, saat ini terdapat 9 jenis vaksin yang telah disetujui oleh BPOM diantaranya adalah Sinovac, Bio Farma, Astra Zeneca, Sinopharm, Moderna, Pfizer, Sputnik V, Janssen, dan Convidecia. Namun kebijakan yang diambil pemerintah tidak selalu dapat diterima oleh masyarakat, hal ini juga terjadi pada awal penerapan kebijakan vaksinasi ini. Tidak semua elemen masyarakat menyetujui atau mendukung berjalannya proses vaksinasi di Indonesia. Bahkan tidak sedikit yang berpendapat negatif mengenai vaksin Covid-19.

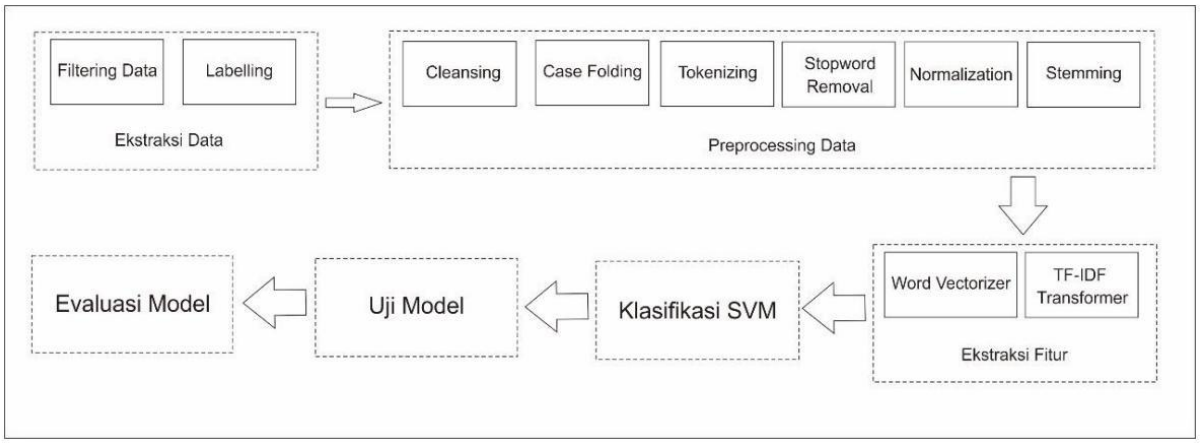
Agar dapat lebih memahami sentimen masyarakat terhadap kebijakan vaksinasi Covid-19, perlu dilakukan analisis terhadap opini atau komentar yang banyak ditulis melalui komentar Youtube. Analisis sentimen pada komentar Youtube dilakukan dengan cara mengekstraksi, mengolah dan memahami data yang mulanya berupa teks yang tidak terstruktur secara otomatis. Analisis ini nantinya digunakan untuk penilaian pendapat atau opini orang terhadap topik tertentu baik itu positif atau negatif (Rozi, 2020). Pada penelitian ini analisis sentimen digunakan untuk menilai pendapat publik terhadap vaksin Covid-19 yang digunakan untuk vaksinasi di Indonesia pada tahun 2021.

Penelitian ini menggunakan metode klasifikasi SVM, dengan pertimbangan metode ini memiliki proses komputasi yang cepat dan jumlah data yang digunakan pada penelitian ini yang tidak banyak sehingga dapat meminimalisir kelemahan SVM yang memerlukan waktu proses

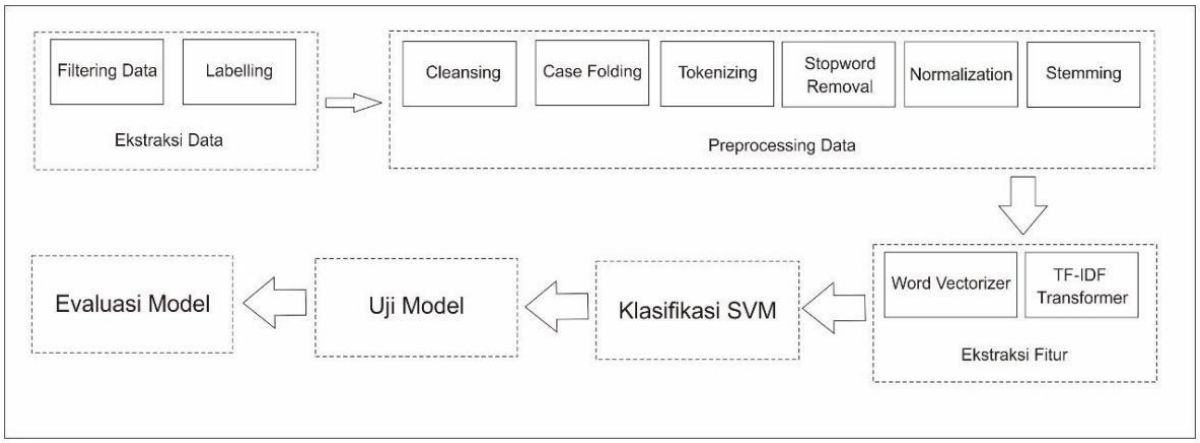
training yang lama. Selain itu juga karena kemampuan SVM untuk meminimalisasi *error* pada *training set*.

II. METODE PENELITIAN

Penelitian ini dibagi menjadi beberapa tahap yang ditunjukkan pada



Gambar 1.



Gambar 1. Alur langkah penelitian

Langkah pertama pada penelitian ini adalah estraksi data dari komentar Youtube menggunakan teknik *crawling* menggunakan R Studio yang kemudian dikumpulkan dalam satu file. Kemudian dilakukan *labelling* dan *filtering* data yang telah terkumpul untuk menentukan sentimen yang terdapat pada masing masing opini yang telah difilter. Langkah selanjutnya adalah melakukan *preprocessing* yang memiliki tujuan untuk mengubah data yang kurang terstruktur menjadi lebih terstruktur. Tahap *preprocessing* dibagi menjadi 6 tahap, yakni :

1. *Cleansing*
   Cleansing bertujuan untuk membersihkan data mentah dari tanda baca seperti titik (.) dan koma (,) dan tanda baca lainnya. Selain itu proses ini juga akan menghilangkan angka di dalamnya.
2. *Case folding*
   Proses selanjutnya adalah *case folding* yakni mengubah data komentar tersebut menjadi *lower-case* atau penulisan dalam huruf kecil.
3. *Tokenization*
   Kemudian dilakukan *tokenization* untuk proses identifikasi kata yang ditandai dengan adanya spasi atau karakter spesial.
4. *Stopword removal*

Setelah itu, dilakukan proses *stopword removal* untuk menghapus kata-kata yang kurang memiliki makna sesuai dengan konteks dari studi kasus yang digunakan, seperti saya, kamu, kita, dan lain – lain.

5. *Normalization*

   Normalization dilakukan untuk mengubah kata yang bermakna namun ditulis dalam bentuk kurang baku sehingga menyulitkan untuk proses *stemming*.

6. *Stemming*

   Kemudian terakhir adalah proses *stemming*, proses ini bertujuan untuk mengembalikan kata yang memilki imbuhan menjadi bentuk aslinya, sebagai contoh kata berharap akan menjadi kata harap.

Langkah berikutnya adalah melakukan pembobotan untuk ekstraksi fitur menggunakan metode TF-IDF. Pada langkah ini dilakukan pengubahan dari data yang telah dilakukan *preprocessing* menjadi data vektor. Hal tersebut bertujuan agar data tersebut dapat diproses pada saat pembobotan.

Tahap berikutnya adalah memulai klasifikasi dengan metode Support Vector Machine berdasarkan sentimen yang telah ada dalam dokumen tersebut. Pada tahap ini data akan dibagi menjadi 3 jenis rasio data *training* : data *testing*, yaitu 70:30, 80:20, dan 90:10. Pada penelitian ini akan dilakukan 10 kali percobaan secara random dan dihitung rata – ratanya. Klasifikasi tersebut akan memberikan hasil berupa akurasi, *recall*, *precision*, dan F1-Score. Pengujian model dilakukan dengan metode *confusion matrix*.

III. HASIL DAN PEMBAHASAN

Sejumlah tahap dilakukan pada penelitian ini, yang meliputi ekstraksi data, preprocessing data, penghitungan TF-IDF, dan klasifikasi data.

1. **Ekstraksi data**

   Ekstraksi data dilakukan dengan menggunakan R Studio . Data diambil dari 10 video dan berasal dari 6 kanal berbeda. *Labelling* diberikan pada komentar yaitu 0 untuk komentar dengan sentimen negatif dan 1 untuk komentar dengan sentimen positif.

   Dari sumber data tersebut diperoleh komentar sebanyak 1007 komentar, diantaranya sebanyak 50,2% adalah komentar negatif dan 49,8% sisanya adalah komentar positif.

2. **Preprocessing data**

   Pada tahap preprocessing, dilakukan langkah seperti yang telah dijelaskan sebelumnya, terdiri dari 6 tahap : *cleansing, case folding, tokenization, stopword removal, normalization,* dan *stemming*. Tabel **1** menunjukkan contoh komentar sebelum dan sesudah melalui tahap *preprocessing* :

Tabel 1. Contoh Preprocessing

Sebelum *preprocessing*	**Setelah *preprocessing***
Saya mau vaksin Nusantara .saya bersyukur Indonesia telah mampu membuat vaksin sendiri dan sangat baik.cintailah produk Indonesia.	[‘mau’, ‘vaksin’, ‘nusantara’, ‘syukur’, ‘indonesia’, ‘mampu’, ‘buat’, ‘vaksin’, ‘sendiri’, ‘sangat’, ‘baik’, ‘cinta’, ‘produk’, ‘indonesia’]

3. **Pembobotan TF-IDF**

   TF-IDF dilakukan untuk menjadikan dataset yang telah dibentuk sebelumnya menjadi bahan input untuk model klasifikasi. Sebagai contoh proses sebagai berikut :

(D1) Saya menunggu banget vaksin Nusantara

(D2) Vaksin itu sangat bermanfaat

(D3) Saya mendukung berjalannya vaksin nusantara

Dengan menggunakan rumus (1) dan (2)

$$W_t = tf_{ij} x Idf \quad (1)$$

$$W_t = tf_{ij} x log\left(\frac{N}{df_t}\right) \quad (2)$$

didapatkan vektor terbobot yang hasilnya ditunjukkan pada Tabel **2**

Tabel 2. Hasil Vector Terbobot

	Tunggu	**Vaksin**	**Nusantara**	**Manfaat**	**Dukung**	**Jalan**
(D1)	1,477	1	1,175	0	0	0
(D2)	0	1	0	1,477	0	0
(D3)	0	1	1,175	0	1,477	1,477

4. Pembagian dan Klasifikasi Data

Data akan dibagi menjadi 3 skenario, setiap skenario menggunakan rasio data training dan data testing yang berbeda yaitu 70:30, 80:20, dan 90:10. Masing – masing skenario akan dilakukan percobaan secara random sebanyak 10 kali . Hasil akhir akan dihitung dari rata – rata 10 kali percobaan tersebut.

Pembagian data dilakukan dengan menggunakan bantuan dari library *sklearn* tepatnya fungsi *train_split_test.* Pembagian data pada masing-masing skenario dapat dilihat pada Tabel **3**.

Tabel 3. Pembagian Data

Skenario	**Rasio**	**Data Training**	**Data Tes**
1	70:30	704	303
2	80:20	805	202
3	90:10	906	101

Hasil rata – rata dari skenario 1 dirangkum pada Tabel **4**.

Tabel 4. Hasil Skenario 1

Uji ke -	**Pembagian data 70 : 30**			
	Akurasi	**Recall**	**F1-Score**	**Presisi**
1	0,86138614	0,90909091	0,85106383	0,8
2	0,85478548	0,90551181	0,83941606	0,78231293
3	0,8679868	0,91549296	0,86666667	0,82278481
4	0,8679868	0,94573643	0,85915493	0,78709677
5	0,84488449	0,83453237	0,83154122	0,82857143
6	0,84818482	0,88372093	0,83211679	0,7862069
7	0,82178218	0,864	0,8	0,74482759
8	0,85148515	0,92647059	0,84848485	0,7826087
9	0,86138614	0,91851852	0,85517241	0,8

10	0,81188119	0,8540146	0,80412371	0,75974026
Rata-Rata	**0,84917**	**0,89570**	**0,83877**	**0,78941**

Dari Tabel **4** diketahui rata – rata akurasi yang didapatkan untuk skenario 1 adalah 84 %, *recall* sebesar 89 %, F1-Score sebesar 83 % , dan presisi 78 %. Kemudian untuk hasil rata – rata yang didapatkan dari skenario 2, ditampilkan pada Tabel **5**.

Tabel 5 Hasil Skenario 2

Uji ke -	**Rasio Data 80 : 20**			
	Akurasi	**Recall**	**F1-Score**	**Presisi**
1	0,87128713	0,92391304	0,86734694	0,81730769
2	0,86138614	0,9010989	0,85416667	0,81188119
3	0,90594059	0,93814433	0,90547264	0,875
4	0,84653465	0,93589744	0,82485876	0,73737374
5	0,86633663	0,8988764	0,85561497	0,81632653
6	0,88613861	0,92708333	0,88557214	0,84761905
7	0,81683168	0,90909091	0,79096045	0,7
8	0,84653465	0,90526316	0,84729064	0,7962963
9	0,85643564	0,92857143	0,84324324	0,77227723
10	0,7970297	0,80898876	0,77837838	0,75
Rata-Rata	**0,85544**	**0,90769**	**0,84529**	**0,79240**

Dari Tabel **5** diketahui rata – rata akurasi yang didapatkan untuk skenario 2 adalah 85 %, *recall* sebesar 90 %, F1-Score sebesar 84 %, dan presisi 79 %. Dan untuk hasil rata-rata yang didapatkan dari skenario 3 dirangkum pada Tabel 6.

Tabel 6. Hasil Skenario 3

Uji ke -	**Rasio data 90 : 10**			
	Akurasi	**Recall**	**F1-Score**	**Presisi**
1	0,9108910	0,9230769	0,9142857	0,90566038
2	0,8712871	0,8490566	0,8737864	0,9
3	0,9207920	0,9444444	0,9272727	0,9107142
4	0,8811881	0,9523809	0,8695652	0,8
5	0,8811881	0,8936170	0,875	0,8571428
6	0,8712871	0,9375	0,8737864	0,8181818
7	0,7722772	0,8857142	0,7294117	0,62
8	0,8811881	0,9166666	0,88	0,8461538
9	0,8415841	0,8571428	0,8181818	0,7826087

10	0,8217821	0,88	0,8301886	0,7857142
Rata-Rata	**0,8653465**	**0,9039599**	**0,8591478**	**0,8226176**

Berdasarkan hasil yang tertulis pada Tabel 6 diketahui rata – rata akurasi yang didapatkan untuk skenario 3 adalah sebesar 86 %, *recall* sebesar 90 %, *F1-score* sebesar 82 %, dan presisi 85 %. Selanjutnya hasil yang didapatkan pada skenario 1, 2, dan 3 akan dibandingkan, seperti yang ditampilkan pada Tabel 7.

Tabel 7. Perbandingan Hasil

	Akurasi	**Recall**	**F1-Score**	**Presisi**
Rasio 70:30	0,8491749	0,8957089	0,838774	0,789414
Rasio 80:20	0,8554455	0,9076927	0,845290	0,792408
Rasio 90:10	0,8653465	0,903959	0,859147	0,8226176
Average	0,8563	0,916	0,8473	0,801
Max	0,865346	0,907692	0,859147	0,822617
Min	0,849174	0,895708	0,838774	0,789414

Dari Tabel 7 didapatkan bahwa rata-rata yang didapatkan pada ketiga skenario adalah 85,6%, *recall* sebesar 91,6%, *f1-score* sebesar 84,7%, dan presisi sebesar 80,1 %. Hampir semua aspek performa untuk pembagian data 90 : 10 (skenario 3) lebih besar kecuali untuk nilai *recall*, nilai terbesar berada pada skenario 2 yakni rasio data 80 : 20 sebesar 90,7%, lebih besar dari rasio 90:10 sebesar 90,3%. Dari hasil penelitian juga ditemukan bahwa semakin besar data *training* yang digunakan maka akan semakin kecil persentase kesalahan prediksi yang ada pada data testing, hal ini dirangkum pada Tabel 8.

Tabel 8. Kesalahan Prediksi

	Rasio 70:30 (data tes sebanyak 303)		**Rasio 80:20 (data tes sebanyak 202)**		**Rasio 90:10 (data tes sebanyak 101)**	
Uji ke-	**False Positive**	**False Negative**	**False Positive**	**False Negative**	**False Positive**	**False Negative**
1	30	12	19	7	5	4
2	32	12	19	9	5	8
3	28	12	13	6	5	3
4	33	7	26	5	10	2
5	24	23	18	9	7	5
6	31	15	16	7	10	3
7	37	17	30	7	19	4
8	35	10	22	9	8	4
9	31	11	23	6	10	6
10	37	20	24	17	12	6

Rata-Rata	**31,8**	**13,9**	**21**	**8,2**	9,1	4,5

Pada tabel 8, rasio data 70 : 30 (skenario 1) memiliki nilai *false positive* dan *false negative* rata – rata sebesar 31,8 dan 13,9 sehingga persentase prediksi yang salah adalah sebesar 15%. Untuk rasio data 80 : 20 memiliki nilai *false positive* dan *false negative* rata – rata sebesar 21 dan 8,2 sehingga persentase prediksi yang salah adalah sebesar 14,4%. Terakhir, untuk rasio data 90 : 10 memiliki nilai *false positive* dan *false negative* rata – rata sebesar 9,1 dan 4,5 sehingga presentase prediksi yang salah adalah sebesar 13,4%. Dapat disimpulkan untuk pembagian rasio data 90:10 memiliki persentase tingkat kesalahan dalam memprediksi yang terkecil.

VI. KESIMPULAN DAN SARAN

a. Kesimpulan

Berdasarkan hasil yang didapatkan pada penelitian ini dapat ditarik sejumlah kesimpulan, antara lain :

1) Algoritma Support Vector Machine yang digunakan pada penelitian ini memberikan akurasi rata-rata sebesar 85,6 % , *recall* sebesar 90,1% , *f1-score* sebesar 84,7% , dan presisi sebesar 80,1%.
2) Pembagian data dengan rasio 90:10 selain memiliki akurasi, presisi, dan *f1-score* paling tinggi yaitu sebesar 86 %, 82,2 % dan 82,2%, dan juga tingkat kesalahan prediksi yang terendah yakni sebesar 13,4%. Namun *recall* tertinggi terdapat pada pembagian data dengan rasio 80 : 20 yaitu sebesar 90,7% dengan selisih 0,4% dari rasio 90 : 10.
3) Dari ketiga skenario pembagian data yang dilakukan menunjukkan perbedaan hasil yang tidak terlalu jauh antara satu dan lainnya namun pembagian data dengan rasio 90:10 didapatkan hasil yang terbaik.

b. Saran

Penelitian ini juga tidak jauh dari kesalahan dan kekurangan, terdapat beberapa hal yang masih dapat dikembangkan atau diperbaiki. Beberapa saran untuk penelitian selanjutnya diantaranya :

1. Menggunakan jenis algoritma klasifikasi lainnya agar dapat dibandingkan dengan hasil uji model yang telah dilakukan.
2. Menambah jumlah data dan banyak percobaan sehingga mendapatkan hasil yang lebih akurat.
3. Menambah jumlah class agar tidak terbatas jenis sentimen positif dan negatif agar jenis emosi lain yang ada pada studi kasus dapat dideteksi.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada rekan-rekan dosen program studi S1 Informatika Universitas Sebelas Maret atas motivasi yang diberikan untuk penyelesaian penelitian ini.

DAFTAR PUSTAKA

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., and Roser, M. (2020). Coronavirus Pandemic (COVID-19) Vaccinations. Retrieved from https://ourworldindata.org/covid-vaccinations.

Rozi, I. F. (2020). Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. *Jurnal EECCIS*, 6(1), 7.

Zhou, R., Khemmarat, S., Gao, L., Wan, J., & Zhang, J. (2016). How YouTube videos are discovered and its impact on video views. *Multimedia Tools and Applications*, 75(10), 6035–6058. https://doi.org/10.1007/s11042-015-3206-0