ANALISIS SENTIMEN MENGENAI VAKSIN SINOVAC MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DAN K-NEAREST NEIGHBOR (KNN)

Anna Baita [1] **, Yoga Pristyanto** [2] **, Nuri Cahyono** [3]

[1,3] *Informatika Universitas AMIKOM Yogyakarta*
[2] *Sistem Informasi Universitas AMIKOM Yogyakarta*

email : anna@amikom.ac.id[1)], yoga.pristyanto@amikom.ac.id[2)], nuricahyono@amikom.ac.id[2)]

Abstraksi

Pandemi COVID-19 yang bermula di Wuhan, Tiongkok, saat ini menjadi pandemi yang terjadi di berbagai negara di seluruh dunia. Upaya vaksinasi dilakukan untuk mengurangi tingkat penyebaran dari virus COVID-19. Pemberian vaksin memberikan dampak yang berbeda-beda, sehingga menimbulkan berbagai opini terhadap pemberian vaksin ini. Sentimen analisis dapat digunakan untuk mengalisa opini masyarakat terhadap pemberian vaksin ini. Dalam penelitian ini menggunakan algoritma SVM dan KNN untuk melakukan analisa mengenai sentimen masyarakat terhadap pemberian vaksin ini. Adapun opini di dapatkan dari aplikasi twitter dengan keyword sinovac. Dataset yang digunakan merupakan cuitan dalam bahasa Inggris. Proses pelabelan teks dilakukan secara otomatis menggunakan textblob. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa yang lebih baik jika dibandingkan dengan algoritma KNN. Akurasi algoritma SVM sebesar 0.7, sedangkan akurasi algoritma KNN sebesar 0.56.

***Kata Kunci* :**
Sentimen, SVM, KNN, textblob, sinovac

Abstract

The COVID-19 pandemic, which started in Wuhan, China, has now become a pandemic that is occurring in many countries around the world. Vaccination efforts are carried out to prevent the spread of the COVID-19 virus. Giving vaccines has different impacts, giving rise to various opinions on the administration of this vaccine. Sentiment analysis can be used to analyze public opinion on the administration of this vaccine. In this study, the SVM and KNN algorithms were used to analyze public sentiment regarding the administration of this vaccine. This opinion is obtained from the twitter application using sinovac keyword. The dataset which used is tweets in English. The labeling process is done automatically using textblob. The results obtained in this study indicate that the SVM algorithm has better accuracy when compared to the KNN algorithm. The accuracy of the SVM algorithm is 0.7, while the accuracy of the KNN algorithm is 0.56.

***Keywords* :**
Sentiment, SVM, KNN, textblob, sinovac

1. Pendahuluan

COVID-19 merupakan penyakit yang disebabkan oleh jenis corona virus. Wabah virus corona berawal di Wuhan, Tiongkok, pada bulan Desember 2019. Namun saat ini COVID menjadi pandemi yang melanda berbagai negara di seluruh dunia[1].

Salah satu cara yang dilakukan untuk mengurangi penyebaran COVID-19 adalah dengan vaksinasi. Vaksinasi merupakan pemberian vaksin (*antigen*) yang digunakan untuk merangsang sistem imun di dalam tubuh.

Pemberian vaksin memberikan dampak yang berbeda-beda, sehingga menimbulkan berbagai opini terhadap pemberian vaksin ini. Ada yang beropini positif, ada yang beropini negatif dan ada yang beropini netral. Pemberian vaksin itu hangat diperbicangkan diberbagai sosial media, tak terkecuali *Twitter*. Dari berbagai macam tanggapan *netizen* di *twitter* dapat dianalisis sentimennya sehingga dapat dikelompokkan jenis opininya menjadi sentimen positif, negatif ataupun netral.

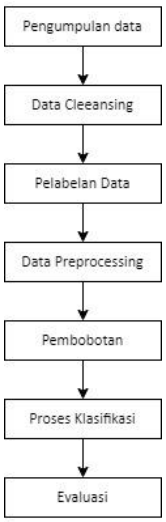
Telah ada penelitian mengenai sentimen analisis dengan topic covid-19 menggunakan metode klasifikasi seperti *KNN*[7], *Naive Bayes*[2][3][5][7], *SVM*[3][7], *LSTM-RNN*[6], dan *Backpropagation*[8] . Menurut penelitian F.S. Pamungkas, Algoritma SVM memiliki nilai akurasi yang lebih baik jika dibandingkan dengan algoritma Naive Bayes maupun KNN dalam sentimen Analisis. Dalam penelitian tersebut menggunakan data *tweet* berbahasa Indonesia. Menurut penelitian [4][9],

Algoritma KNN memiliki performa yang baik dalam analisa sentimen berbahasa Indonesia.

Berdasarkan ulasan tersebut di atas, maka penulis mengusulkan penggunaan algoritma SVM dan KNN dalam sentimen analisis mengenai vaksin *sinovac*, dengan data *tweet* berbahasa Inggris.

2. Metode Penelitian

Obyek yang digunakan dalam penelitian ini berupa cuitan *netizen* yang diperoleh dari *twitter*. Cuitan ini di *crawling* menggunakan *library* dari *python* yang bernama *tweepy*. Adapun metode penelitian digambarkan oleh diagram berikut ini:



Gambar 2.1 Diagram Metode Penelitian

1. Pengumpulan Data

Pengumpulan dengan *tools tweepy*, menggunakan *keyword*: **sinovac** tanpa *retweet*. Data *tweet* merupakan cuitan berbahasa inggris. Dari hasil *crawling* didapatkan 2105 data. Semua data *tweet* dilakukan proses *cleansing*, yakni penghilangan *hastag*, *user* dan *hyperlink*.

2. Pelabelan data

Proses pelabelan dilakukan secara otomatis menggunakan *textblob*. *Text blob* akan menghitung nilai *polarity* dan *subjectivity*. *Polarity* merupakan fungsi untuk melihat kecondongan sentimen sebuah teks. Sedangkan *subjectivity* merupakan fungsi untuk melihat *value* dari sebuah teks. Jenis *value* dari teks ini bisa berupa ebuah opini atau fakta. Semakin tinggi *subjectivity* sebuah teks, maka teks tersebut dapat dikatakan sebagai sebuah opini, sedangkan semakin tinggi *polarity* maka menandakan sebuah emosi yang positif. Dari nilai polarity, maka akan didapatkan 3 kelas, yakni positif, negatif dan netral.

3. *Data Preprocessing*

Preprocessing dilakukan dengan beberapa tahap antara lain: *casefolding*, *tokenizing*, *stopword removal* dan *stemming*. Proses *casefolding* dilakukan dengan mengubah teks menjadi *lowercase*, menghilangkan karakter angka, dan menghilangkan beberapa karakter yang tidak diperlukan seperti ":, _,=,+" dll. Proses t*okenizing* dilakukan untuk mengubah sebuah kalimat menjadi kata/*token*/*term*. Proses *stopword removal* merupakan proses untuk menghilangkan kata sambung seperti: *and*, *with*, *the*, dll. *Stemming* merupakan proses mengubah sebuah kata menjadi kata dasar. Misal: *programming* diubah menjadi *program*

4. Pembobotan

Tahap berikutnya adalah menghitung bobot dari setiap *token*/*term* berdasarkan frekuensi kemunculan term tersebut dalam document menggunakan metode TF-IDF.

5. Proses Klasifikasi

Proses Klasifikasi dilakukan dengan menggunakan algoritma SVM dan KNN. Sebelum proses klasifikasi, data dibagi menjadi 2 bagian, yakni: d*ata training* dan *data testing*, dengan komposisi pembagian *data testing* sebesar 20% dari total *dataset* yang ada. Model SVM dan KNN dibangun menggunakan *data training*, kemudian dievaluasi menggunakan *data testing*.

a. Algoritma SVM

SVM merupakan algoritma klasifikasi yang dilakukan dengan menentukan *hyperplane*. *Hyperplane* yang bagus akan berada tepat ditengah-tengah kedua kelas, sehingga memiliki jarak yang paling jauh ke data-data terluar di kedua kelas[10]

b. Algoritma KNN

Algoritma KNN merupakan algoritma klasifikasi yang mengelompokkan data baru berdasarkan jarak data baru ke beberapa data atau tetangga terdekat[11].

6. Evaluasi

Evaluasi dilakukan dengan menghitung nilai *confusion matrix*. *Confusion matrix* akan menghitung nilai *accuracy*, *precision*, *recall*, dan *F1-Score*. Hasil perhitungan akan divalidasi menggunakan *K-Fold Cross Validation*.

3. Hasil dan Pembahasan

Implementasi sistem dan Analisa hasil dilakukan dengan menggunakan bahasa pemrograman *python*. Adapun implementasi setiap tahapannya adalah sebagai berikut:

1. Pengumpulan Data

Untuk dapat menggunakan *tweepy*, maka kita perlu memiliki *API Key* dan *Access Token* terlebih dahulu. Data yang telah didapatkan kemudian disimpan dalam format.csv. Data mentah hasil *crawling* ditunjukkan oleh gambar 3.1 berikut ini:

	date	text	user	sentiment
0	2022-01-14 21:42:51	b"@jeffspolitics @KyleKulinski Sure, vaccines ...	NikolaMK	NaN
1	2022-01-14 21:25:47	b'Moderna vaccines the best \xe2\x80\x94 and S...	Alejandro Valerio	NaN
2	2022-01-14 21:21:51	b'@sailorrooscout Thanks but as I see, it\xe2\...	10	NaN
3	2022-01-14 20:12:44	b'Inactivated vaccines might protect against s...	JARI HERRANEN	NaN
4	2022-01-14 19:56:20	b"For most of 2020, North Korea rejected offe...	christopher white	NaN

Gambar 3.1 Data Mentah

Setelah itu dilakukan proses *cleansing* dengan menghilangakan *hastag*, *user* dan *hyperlink*. Hasil proses *cleansing* ditunjukkan oleh gambarl 3.2 berikut ini:

	date	text	user	sentiment
0	2022-01-14 21:42:51	Sure vaccines are better just keep jabbing eve...	NikolaMK	NaN
1	2022-01-14 21:25:47	Moderna vaccines the best and Sinovac least ef...	Alejandro Valerio	NaN
2	2022-01-14 21:21:51	Thanks but as I see it s just about efficiency...	10	NaN
3	2022-01-14 20:12:44	Inactivated vaccines might protect against sev...	JARI HERRANEN	NaN
4	2022-01-14 19:56:20	For most of 2020 North Korea rejected offers o...	christopher white	NaN

Gambarl 3.2 Hasil *Cleansing*

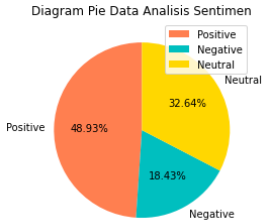
2. Pelabelan data

Pelabelan data dihitung dengan menggunakan *textblob* untuk mendapatkan nilai *polarity* dan subjectifity. Nilai *polarity* <0, akan dilabeli sebagai sentiment *negative*, untuk *polarity* =0 akan dilabeli sebagai sentimen netral dan untuk *polarity*>0 akan dilabeli sebagai sentimen positif. Adapun hasil pelabelan ditunjukkan oleh gambar 3.3 berikut ini.

text	user	sentiment	TextBlob_Polarity	TextBlob_Subjectivity
Sure vaccines are better just keep jabbing eve...	NikolaMK	Positive	0.333333	0.546296
Moderna vaccines the best and Sinovac least ef...	Alejandro Valerio	Positive	0.433333	0.500000
Thanks but as I see it s just about efficiency...	10	Positive	0.200000	0.200000
Inactivated vaccines might protect against sev...	JARI HERRANEN	Positive	0.100000	1.000000
For most of 2020 North Korea rejected offers o...	christopher white	Positive	0.300000	0.250000

Gambarl 3.3 Hasil Pelabelan

Komposisi hasil pelabelan, didapatkan data sebagai berikut:



Gambar 3.4 Diagram Pie Hasil pelabelan

3. Data Preprocessing

Data preprocessing dilakukan dengan beberapa tahap antara lain:

a. *Casefolding*

Hasil proses *casefolding* ditunjukkan oleh tabel berikut ini gambar 3.5 berikut ini:

	date	text
0	2022-01-14 21:42:51	sure vaccines are better just keep jabbing eve...
1	2022-01-14 21:25:47	moderna vaccines the best and sinovac least ef...
2	2022-01-14 21:21:51	thanks but as i see it s just about efficiency...
3	2022-01-14 20:12:44	inactivated vaccines might protect against sev...
4	2022-01-14 19:56:20	for most of north korea rejected offers of th...

Gambar 3.5 Hasil *Casefolding*

b. *Tokenizing*

Proses *tokenizing* digunakan untuk membagi kalimat kedalam *token*/kata/*term*. Adapun hasil proses *tokenizing* ditunjukkan oleh gambar 3.6 berikut ini:

	date	text
0	2022-01-14 21:42:51	[sure, vaccines, are, better, just, keep, jabb...
1	2022-01-14 21:25:47	[moderna, vaccines, the, best, and, sinovac, l...
2	2022-01-14 21:21:51	[thanks, but, as, i, see, it, s, just, about, ...
3	2022-01-14 20:12:44	[inactivated, vaccines, might, protect, agains...
4	2022-01-14 19:56:20	[for, most, of, north, korea, rejected, offers...

Gambar 3.6 Hasil *Tokenizing*

c. *Stopword Removal*

Proses *stopword removal* merupakan proses penghapusan kata hubung seperti *the, for, are, of* dll. Hasil proses *stopword removal* ditunjukkan oleh gambar 3.7 berikut ini.

	date	text
0	2022-01-14 21:42:51	[sure, vaccines, better, keep, jabbing, every,...
1	2022-01-14 21:25:47	[moderna, vaccines, best, sinovac, least, effe...
2	2022-01-14 21:21:51	[thanks, see, efficiency, data, hku, study, se...
3	2022-01-14 20:12:44	[inactivated, vaccines, might, protect, severe...
4	2022-01-14 19:56:20	[north, korea, rejected, offers, sinovac, astr...

Gambar 3.7 Hasil *Stopword Removal*

d. *Stemming*

Proses *stemming* dilakukan dengan menggunakan algoritma porter. Algoritma ini digunakan untuk mengubah *token* menjadi sebuah kata dasar. Hasil dari proses *stemming* ditunjukkan oleh Gambar 3.8 berikut ini:

	date	text
0	2022-01-14 21:42:51	sure vaccin better keep jab everi month pfizer...
1	2022-01-14 21:25:47	moderna vaccin best sinovac least effect stop ...
2	2022-01-14 21:21:51	thank see effici data hku studi see hospit sev...
3	2022-01-14 20:12:44	inactiv vaccin might protect sever diseas inac...
4	2022-01-14 19:56:20	north korea reject offer sinovac astrozeneca v...

Gambar 3.8 Hasil *Stemming*

4. Pembobotan

Proses pembobotan kata digunakan dengan algoritma TF-IDF. Bobot sebuah kata akan dinilai berdasarkan frekuensi kemunculan kata tersebut dalam sebuah kalimat ataupun dalam sebuah dokumen.

Hasil Pembobotan TF-IDF dengan nilai bobot terbesar kata ditunjukkan oleh gambar 3.9 berikut ini:

	term	rank
792	sinovac	168.691291
928	vaccin	113.341408
99	booster	83.878241
627	pfizer	82.817672
183	covid	63.682095

Gambar 3.9 *Top Rank-TF-IDF*

5. Proses Klasifikasi dan Evaluasi Hasil

Proses Klasifikasi dilakukan dengan algoritma SVM dan KNN. Pembuatan model klasifikasi dilakukan dengan library *sckitlearn*. *Data training* yang digunakan sebesar 80%, dan d*ata testing* yang digunakan untuk evaluasi sebesar 20 %.

A. Penerapan Algoritma SVM

Hasil Klasifikasi dengan menggunakan algoritma SVM ditunjukkan oleh *confusion matrix* berikut ini:

	precision	recall	f1-score	support
Negative	0.80	0.39	0.53	61
Neutral	0.73	0.68	0.70	145
Positive	0.72	0.86	0.78	215
accuracy			0.73	421
macro avg	0.75	0.64	0.67	421
weighted avg	0.73	0.73	0.72	421

Gambar 3.10 *Confusion matrix* hasil SVM

Dari tabel tersebut, didapatkan hasil akurasi sebesar 73%. Hasil pengujian kemudian dievaluasi menggunakan *K-Fold Cross Validation*, dengan jumlah *fold* sebesar 10 *fold*. Dari hasil validasi didapatkan rata-rata akurasi sebesar: 0.70

Pengujian juga dilakukan dengan memakai beberapa jenis kernel, antara lain sebagai berikut:

Tabel 3.1 Pengujian Ragam *Kernel* SVM

kernel	accuracy	Rata-rata K-fold(10 Fold)
Linear	0.73	0.70
Polynomial	0.60	0.57
RBF	0.65	0.66

Dari tabel tersebut dapat disimpulkan bahwa hasil terbaik ditunjukkan oleh model SVM yang menggunakan *kernel linear*, dengan rata-rata akurasi sebesar: 0.70.

B. Penerapan Algoritma KNN

Hasil Klasifikasi dengan menggunakan algoritma KNN (*n=7*) ditunjukkan oleh *confusion matrix* berikut ini:

	precision	recall	f1-score	support
Negative	0.60	0.34	0.44	61
Neutral	0.53	0.66	0.59	145
Positive	0.67	0.63	0.65	215
accuracy			0.60	421
macro avg	0.60	0.55	0.56	421
weighted avg	0.61	0.60	0.60	421

Gambar 3.11 *Confusion Matrix* KNN

Dari tabel tersebut, dapat diketahui bahwa nilai akurasi yang didapatkan adalah sebesar 0.60. Adapun setelah divalidasi menggunakan *k-fold* validation (10 *fold*), maka didapatkan rata-rata akurasi sebesar 0.56

Tabel 3.2 Pengujian Ragam nilai *n* pada KNN

Jumlah *n*	*accuracy*	Rata-rata *K-fold*(10 *Fold*)
3	0.60	0.55
5	0.56	0.55
7	0.60	0.56

Dari tabel 3.2 tersebut dapat disimpulkan bahwa hasil terbaik ditunjukkan oleh model KNN yang menggunakan *n* sebanyak 7, dengan rata-rata akurasi sebesar: 0.56.

Dari penggunaan algoritma SVM ataupun KNN didapatkan hasil yang kurang memuaskan yakni kurang dari 75%. Adapun SVM memiliki performa yang lebih baik jika dibandingkan dengan algoritma KNN dalam penelitian ini.

Hasil akurasi yang cukup rendah ini diduga karna proses pelabelan yang otomatis, serta hanya menggunakan satu jenis platform saja yakni: *textblob*.

4. Kesimpulan

Dari hasil penelitian yang telah dilakukan dapat ditarik kesimpulan sebagai berikut:

1. Opini *netizen* terkait dengan vaksin *sinovac* yang diambil dari aplikasi *twitter* dengan menggunakan keyword **sinovac**, menghasilkan 1030 opini(data *tweet*) bersentimen positif, 388 opini bersentimen negatif, dan 687 opini bersentimen netral.

2. Algoritma SVM dan KNN dapat digunakan untuk mengklasifikasikan data *tweet* dengan rata-rata hasil akurasi sebesar 0.7 (untuk algoritma SVM) dan 0.56 (untuk algoritma KNN)

3. Dalam penelitian ini, algoritma SVM menunjukkan performa yang lebih baik jika dibandingkan dengan algoritma KNN dalam hal akurasi

4. Algoritma SVM memiliki performa yang paling baik ketika dijalankan menggunakan *kernel linear*

5. Algoritma KNN memiliki performa yang paling baik ketika dijalankan menggunakan *n* sejumlah 7

6. Nilai akurasi dari SVM ataupun KNN terbilang rendah, diduga karena proses pelabelan sentimen yang otomatis menggunakan *text blob*, untuk itu dipenelitian selanjutnya dapat dicoba penggunaan pelabelan manual ataupun juga menggunakan *platform* lain seperti *vader* ataupun *spacy*.

Daftar Pustaka

[1] https://www.who.int/indonesia/news/novel-coronavirus/qa/qa-for-public diakses 17 Januari 2022
[2] S. Lestari and S. Saepudin, "Analisis Sentimen Vaksin Sinovac Pada Twitter Menggunakan Algoritma Naive Bayes," SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika), 2021.
[3] B. Laurensz, E.Sediyono, "Analisis Sentimen Masyarakat Terhadap Tindakan Vaksinasi Dalam Upaya Mengatasi Pandemi Covid-19", Jurnal Nasional Teknik Elektro dan Teknologi Informasi,2021, Vol: 10 No:2
[4] S. Ernawati and R. Wati, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel", Jurnal Khatulistiwa Informatika, 2018 Vol: VI No:1
[5] F. Fathonah and A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid-19 Menggunakan Metode Naive Bayes" Jurnal Sains dan Informatika, 2021, Vol: 7 No: 2
[6] R. Chandra, and A. Krishna,"Covid-19 Sentiment Analysis Via Deep Learning During The Rise Of Novel Cases"Plos One,2021.
[7] F. S. Pamungkas and I Kharisudin, "Analisis Sentimen dengan SVM, Naive Bayes, dan KNN untuk Studi Tanggapan Masyarakat Terhadap Pandemi Covid-19 Pada Media Sosial Twitter"
[8] T. Hendrawati and C. P. Yanti, "Analysis of Twitter Users Sentiment Against the Covid-19 Outbreak Using the Backpropagation Method With Adam Optimizer", Journal Of Electrical, Electronics and Informatics,
[9] R. P. Fitrianti, A. Kurniawati, and D. Agusten, "Implementasi Algoritma K-Nearest Neighbor Terhadap Analisis Sentimen Review Restoran Dengan Teks Bahasa Indonesia", Seminar Nasional Aplikasi Teknologi Informasi(SNATi), 2019
[10] Suyanto, " Machine Learning Tingkat Dasar dan Lanjut", Bandung: Informatika, 2018
[11] B. Santosa, "Data Mining: Teknik pemanfaatan data untuk Keperluan Bisnis", Yogyakarta: Graha Ilmu, 2007