Analisis Sentimen Kebijakan Vaksin Covid-19 Menggunakan SVM dan C4.5

Muhammad Siddik Hasibuan, Suhardi
Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara
E-mail: muhammadsiddik@uinsu.ac.id, suhardi@uinsu.ac.id

Abstrak— Analisis sentimen terhadap kebijakan vaksin covid-19 di Indonesia menjadi suatu hal yang perlu untuk diteliti. Opini tersebut dapat di jadikan suatu model penelitian, yaitu menggunakan metode klasifikasi data mining menggunakan algortima SVM dan C4.5. Dataset yang digunakan adalah opini atau sentimen masyarakat yang di posting pada media sosial twitter. Data yang diambil sebanyak 200 data, selanjutnya dilakukan proses pre-prosessing menggunakan metode TD-IDF data menjadi 137 dateset. Proses selanjutnya menyeimbangkan data dengan fungsi SMOTE, hasil dari performance dari algoritma SVM mendapat nilai akurasi 99.46 sedangkan algoritma C4.5 mendapat nilai akurasi 69.02. Dari hasil analisis yang dilakukan algoritma SVM mendapat nilai optimum yang lebih baik dari algoritma C4.5.

Kata Kunci— *analisis sentimen, SVM, C4.5, TD-IDF, SMOTE*

I. PENDAHULUAN

Corona disease 2019 disingkat menjadi Covid-19 atau virus SARS-CoV-2 adalah virus yang dapat menyebar dari partikel cairan kecil yang keluar dari mulut atau hidung manusia. Penyakit ini mengakibatkan pandemi yang sudah dua tahun kita rasakan. Seiring dengan waktu muncul suatu terobosan yaitu adanya vaksin covid-19, vaksin ini dipercaya dapat menurunkan keganasan covid-19, tetapi banyak pula yang meragukan efektifitas vaksin tersebut. Program vaksinasi nasional merupakan kebijakan pemerintah untuk mengurangi bahaya covid-19, oleh sebab itu pemerintah mewajibkan setiap orang yang bersentuhan langsung seperti melakukan perjalanan atau beraktifitas di tempat umum untuk melalukan vaksinansi terlebih dahulu. Kebijakan ini membuat masyarakat beropini melalui media sosial, banyak perdebatan tentang vaksin covid-19 ada yang setuju ataupun tidak setuju vaksin dijadikan syarat untuk melakukan kegiatan seperti tersrebut sebelumnya. Media sosial adalah suatu tempat berbagi informasi yang cepat dan mudah, salah satu media terkenal seperti Twitter. Twitter memiliki karakteristik yang unik yaitu memiliki cara post berita yang dinamakan tweet. Tweet dibatasi jumlah teks yang di muat maksimal 140 karakter [1]. Tweet tersebut dapat dijadikan suatu analisis dari tweet-tweet yang di post oleh masyarakan umum, kemudahan ini juga dibarangi dengan konsep singkroniasi menggunakan API twitter[2]. Kebijakan Respon dari masyarakat terhadap suatu kebijakan dapat dijadikan suatu indikator kepuasa dari kebijakan yang dikeluarkan, respon tersebut dapat dianalisis menjadi suatu data yang dapat mengahsilkan presentase tweet positif, negatif dan netral [3]. Presentase data yang akan digunakan adalah tweet yang di klasifikasi dengan metode Support Vektor Machine (SVM) dan C4.5. Metode SVM dipilih karena dapat mengklasifikasi suatu data (vektor) dengan jarak sebagai penentunya.

Analisis sentimen merupakan bagian dari text mining, text mining memiliki tujuan untuk mengklarifikasi opini ke dalam suatu kelas tertentu. Opini dapat diklasifikasikan ke dalam kelas netral, positif, dan negatif. Contoh penggunakan analisis sentimen dalam perusahaan seperti menganalisis permintaan penjualan, data tersebut dapat digunakan oleh manajemen perusahaan sebagai dasar untuk berbagai proses pengambilan keputusan di perusahaan [4]. Analisis sentimen merupakan pendekatan dengan cara Natural Language Processing (NLP) didalamnya termasuk metode text mining yang berfokus untuk menggali data didalam suatu sumber media informasi [5]. Dalam penelitian ini text di klasifikasi menggunakan SVM dan C4.5.

II. BAHAN DAN METODE

Penelitian ini menggunakan algoritma SVM dan C4.5 menggunakan dataset yang diambil langsun pada flatform twitter dengan pencarian populer "vaksin covid-19" pada tanggal 20 oktober 2021 pukul 13.30 WIB dengan lokasi Indonesia.

1. Pre-processing

Pre-processing menjadi proses yang sanggat penting untuk mendapatkan model yang optimal untuk menghasilkan akurasi yang tinggi. Pada proses ini dilakukan pengumpulan data pada flatform twitter seperti membersihan missing value, membuang tweet ganda dan membuat class atribut. Dataset yang dicari 200 tweet setelah dilakukan proses pre-processing didapat 134 dataset yang siap diolah, adapun dataset dapat dilihat pada tabel 1.

Tabel 1. Dataset

Text	Label
BHnasional Pesakit antivaksin berubah fikiran selepas ahli keluarga dijangkiti COVID-19 https://t.co/8dDZuopLO1 vaksin coronavirus covid19	Netral
Efektivitas vaksin covid19 Pfizer, Moderna, dan AstraZeneca menurun seiring berjalannya waktu. Seperti apa perbedaan penurunannya antara satu vaksin dengan yang lain? https://t.co/XctaaRH0r6 Infografis CNNIndonesia https://t.co/iFLbPKxWEq	Netral
Assaam'mualaikum Selamat pagi Sehat dan berkah UK, Rusia dan Rumania mengalami kasus covid19 yg meningkat, karena masih ada kelompok masyarakat yg menolak utk vaksin. Semoga di Indonesia semua terjangkau vaksinasi.	Netral

DGA Euronews world 20 Okt 2021	
komarkomarkomar: BAHAYA-BAHAYA VAKSIN COVID-19, a thread. Thread ini kumpulan bukti-bukti bahaya/efek samping dari vaksin covid. Hinda…	Netral
kurawa: 57. Saat kejadian penangkapan ini diketahui ada beberapa polisi lain berseragam yg sedang bertugas melakukan pengawalan distrib…	Netral
Pengobatan Kanker Bisa Turunkan Efektivitas Vaksin Covid-19 https://t.co/temPXKMdEy https://t.co/Ac1O3FX5cw	Netral
Tapi TIDAK terlalu parah karena berkat vaksin. Semoga baik baik saja. COVID19 https://t.co/2MnAn83hYg	Netral

2. Natural Language Processing

Pada proses ini dilakukan metode TF-IDF (Term Frequency – Inverse Document Frequency) mengukur kesamaan kata[6] berupa token-token seperti menghapus link dan simbol, lower case, menghilangkan angka dan memecah kata, mengukur kata (min : 4 karakter, max : 25 karakter), filter stopwords indonesia.

Tabel 2. Datases setelah di filter

Word	Attribute Name	Netral	Positif	Negatif
anti	anti	0.0	0.0	5.0
antibodi	antibodi	1.0	0.0	0.0
antivaksin	antivaksin	9.0	0.0	0.0
antivax	antivax	0.0	1.0	0.0
antiviksin	antiviksin	1.0	0.0	0.0
assaam	assaam	1.0	0.0	0.0
astrazeneca	astrazeneca	5.0	1.0	0.0
anti	anti	0.0	0.0	5.0
Dst..				

3. Cross Validation

Metode yang paling biasa digunakan untuk evaluasi kinerja prediktif dari model, model ini mengeneralisasikan data independen, proses pemisahan data training dan data testing dilakukan oleh fungsi K-Fold [7]. K-Fold yang dipakai dalam penelitian ini adalah K=10.

4. Klasifikasi

Klasifikasi adalah salah satu teknik dalam data mining, klasifikasi merupakan pengelompokan data dimana data tersebut memiliki kelas atau label [8]. Algoritma-algoritma klasifikasi dikategorikan kedalam pembelajaran terawasi atau supervised learning. Pembelajaran terawasi adalah setiap dataset yang akan di uji memiliki label atau class yang berfungsi sebagai guru yang mengawasi pembelajaran untuk mencapai tingkat akurasi atau presisi tertentu. Jenis klasifikasi yang dipakai dalam penelitian ini adalah SVM dan C4.5

Metode klasifiaksi lainya selain C4.5 adalah Support vector machine atau SVM. SVM adalah suatu algoritma terawasi sama dengan C4.5 fungsi utamanya adalah untuk klasifikasi dan regresi [8]. SVM mampu untuk menyelesaikan masalah klasifikasi untuk data besar terutama pada permasalahan aplikasi multidomain di lingkungan big data.[9]. SVM dikembangkan untuk mengklasifiaksi problem liniear ataupun non linier. Klasifikasi antar kelas linier tersebut dapat dilihat pada persamaan berikut.

$$f(x) = W^T X + b$$

Ekuvalen :

$$[W^T.X_i + b] \geq 1 \; untuk \; Yi = +1$$
$$[W^T.X_i + b] \leq 1 \; untuk \; Yi = -1$$

Xi = data training
Yi = label dari Xi

Agar mendapatkan hyperlane terbaik yaitu mencari hyperlane yang berada di tengah antara pembatas kelas.

Sedangkan C4.5 digunakan untuk proses mengklasifikasi data bersifat prediktif. C4.5 merupakan pembaharuan dari algortima sebelumnya yaitu algoritma *Iterative Dichotomiser3* (ID3) yang ditemukan oleh Ross Quinlan. Selanjutnya Quinlan mepresentasikan metode C4.5, dimana untuk pemilihan split atribut pada ID3 menggunakan Information Gain maka C4.5 menggunakan Gain Ratio (GR). Proses seleksi atribut terbaik ialah atribut yang memungkinkan mendapatkan ukuran pohon keputusan terkecil. Atau atribut yang dapat memisahkan objek berdasarkan kelas. Secara heuristik, atribut yang diseleksi adalah atribut yang menghasilkan node termurni (*cleanest*). Besar kecilnya kemurnian dinyatakan dengan tingkat pengotor, dan untuk menghitungnya dapat dilakukan dengan menggunakan konsep Entropi, Entropi menyatakan pengotor suatu kumpulan benda.

$$Entropi\ (S) \sum_{j=1}^{k} -Pj\ Log_2\ Pj$$

Gambar 1. Persamaan Entropi

S = data,
K = jumlah partisi S,
pj = probabilitas Sum(Ya) / Total Kasus.

$$gain\ ratio\ (a) = \frac{gain\ (a)}{split\ (a)}$$

Gambar 2. Persamaan gain ratio

a = atribut,
gain(a) = information gain pada atribut a,
Split(a) = split information pada atribut a

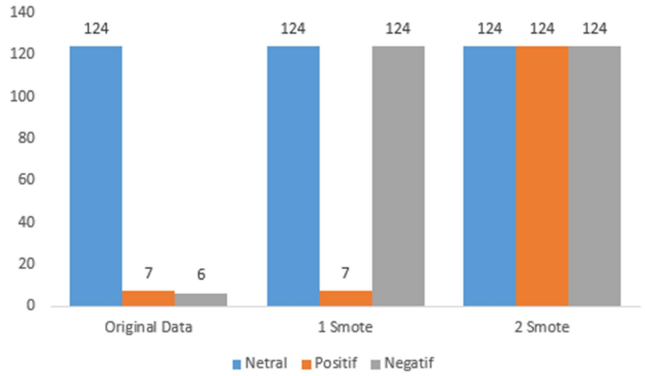
III. HASIL DAN PEMBAHASAN

Pengujian sebanyak 137 data, pengelompokan data sentimen yang telah dikategorikan netral, positif dan negatif. Adapun kelompok data tersebut dilihat pada tabel 3.

Tabel 3. Dataset Sentimen

SENTIMEN	JUMLAH
NETRAL	124
POSITIF	7
NEGATIF	6
	137

Dengan melihat jumlah dataset yang tidak merata antara netral, positif dan negatif maka perlu menggunakan *Synthetic Minority Over Sampling (SMOTE).* SMOTE adalah suatu metode untuk meningkatkan jumlah kelas minoritas, sehingga jumlah sampel simbang[10].



Gambar 3. Proses SMOTE

Performace dari algoritma SVM didapat nilai akurasi 99.46 dapat dilihat pada tabel 4

Tabel 4. Hasil Performance SVM

	true Netral	*true Positif*	*true Negatif*	*class precision*
pred. Netral	124	0	2	98.41%
pred. Positif	0	124	0	100.00%
pred. Negatif	0	0	122	100.00%
class recall	100.00%	100.00%	98.39%	

Performace dari algoritma c4.5 didapat nilai akurasi 69.02 dapat dilihat pada tabel 5

Tabel 5. Hasil Performance c4.5

	true Netral	*true Positif*	*true Negatif*	*class precision*
pred. Netral	85	75	1	52.80%
pred. Positif	39	49	0	55.68%
pred. Negatif	0	0	123	100.00%
class recall	68.55%	39.52%	99.19%	

IV. KESIMPULAN

Hasil dari analisis yang dilakukan dapat disimpulkan algoritma SVM mendapat nilai performance 99.46 sedangkan c.45 mendapat nilai 69.02. Data ini menunjukan algoritma SVM bekerja optimal dibandingkan dengan C4.5 untuk mengklasifikasi sentimen masyarakat mengenai kebijakan vaksin covid-19

DAFTAR PUSTAKA

[1] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," *HP Lab. Tech. Rep.*, 2011.

[2] M. Myslín, S. H. Zhu, W. Chapman, and M. Conway, "Using twitter to examine smoking behavior and perceptions of emerging tobacco products," *J. Med. Internet Res.*, 2013, doi: 10.2196/jmir.2534.

[3] R. Kurniawan And A. Apriliani, "Analisis Sentimen Masyarakat Terhadap Virus Corona Berdasarkan Opini Dari Twitter Berbasis Web Scraper," *J. Instek (Informatika Sains Dan Teknol.*, 2020, Doi: 10.24252/Instek.V5i1.13686.

[4] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, "Corrigendum: Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine," *J. Inf. Syst. Eng. Bus. Intell.*, 2018, doi: 10.20473/jisebi.4.2.169.

[5] N. Munasatya and S. Novianto, "Natural Language Processing untuk Sentimen Analisis Presiden Jokowi Menggunakan Multi Layer Perceptron," *Techno.Com*, 2020, doi: 10.33633/tc.v19i3.3630.

[6] M. Nurjannah, Hamdani, and I. Fitri Astuti, "Penerapan Algoritma Term Frequency-Inverse Document Frequency (Tf-Idf) Untuk Text Mining," *J. Inform. Mulawarman*, 2013.

[7] M. J. Zaki, "OLD---Data Mining and Analysis: Fundamental Concepts and Algorithms," *Personal. Soc. Psychol. Bull.*, 1997.

[8] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning: Methods and Applications to Brain Disorders*, 2019.

[9] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*. 2016.

[10] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*. 2018, doi: 10.1613/jair.1.11192.