ANALISA SENTIMEN RESPON MASYARAKAT TERHADAP COVID 19 BERBASIS LINEAR SUPPORT VECTOR MACHINE

Puri Sulistiyawati[1]**,Viry Puspaning Ramadhan**[2]**, Farrikh Alzami***[3]**, Ricardus Anggi P**[4]**, Rama Aria Megantara**[5]**, Dwi Puji Prabowo**[6]

[1,3,4,5,6]*Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang*

Gedung H Lantai 2, Kampus 1 Jl. Imam Bonjol No 207, Semarang , Jawa Tengah

[2]*Sistem Informasi,Fakultas Teknologi Informasi, Universitas Merdeka Malang*

Jl. Terusan Raya Dieng 62-64 Malang, Jawa Timur

E-mail*: virypuspa@gmail.com[2]

Abstrak

Twitter banyak digunakan oleh masyarakat Indonesia dalam mengutarakan pendapat pribadi. Dikarenakan pengguna media social berkisar di segala umur, maka diperlukan sebuah metode untuk mengklasifikasikan bahwa suatu tulisan twitter yang ditulis termasuk dalam kategori positif atau negatid. Tujuan klasifikasi ini digunakan sebagai batu pijakan dalam memberikan masukan kepada pemerintah tentang keaadaan informasi di masyarakat. Penelitian ini, menggunakan Sastrawi untuk proses stemming tulisan twitter, kemudian melalui proses pembersihan karakter, bigram dan Linear Support Vector Machine digunakan untuk tahap klasifikasi, serta 3 fold cross validation, menghasilkan performa dengan recall sebesar 99.13%, precision sebesar 99.14% dan akurasi sebesar 99.13 %. Dengan performa yang cukup menggembirakan, dapat disimpulkan bahwa kombinasi sastrawi, pembersihan karakter, bigram dan linear support vector machine dapat digunakan untuk mendeteksi sentiment masyarakat terhadap covid 19.

Kata Kunci**:** Sentiment Analysis, Bigram, Linear Support Vector Machine, Sastrawi

I. PENDAHULUAN

COVID-19 merupakan penyakit yang awal mulanya dikenal sebagai penyakit Coronavirus 2019 yang telah dinyatakan sebagai sebuah pandemi oleh Organisasi Kesehatan Dunia (WHO) pada tanggal 11 Maret 2020 lalu. Selama ini belum pernah terjadi sebuah tekanan yang sangat meningkat di setiap negara dimana hal tersebut membuat setiap negara terdesak untuk mengendalikan populasi dengan jumlah kasus yang ada serta memanfaatkan sumber daya yang tersedia (Chakraborty et al., 2020). COVID-19 sekarang sudah menjadi sumber depresi, stres, serta kecemasan yang menimbulkan penyakit mental bagi seseorang yang bisa dikarenakan oleh banyaknya informasi yang menyesatkan yang diposting di media sosial. Kesehatan mental bisa juga dipengaruhi oleh tersebarnya informasi palsu secara cepat di media sosial.

Banyaknya pengguna aktif di jejaring sosial berarti pada saat yang sama, komentar negative yang dihasilkan memperlihatkan kepanikan dan kurangnya literasi terhadap penulis komentar. Maka dari itu, diperlukan sebuah upaya peninjauan komentar negative dari masyarakat terhadap kebijakan penanganan Covid-19.

Analisis sentimen bertujuan mengidentifikasi sikap atau suasana hati orang melalui pemrosesan bahasa alami, analisis teks, dan linguistik komputasi. Dalam beberapa tahun terakhir, pembelajaran mesin telah menjadi alat yang sangat kuat untuk mengklasifikasikan sentimen. Dari literatur, analisis sentiment sudah dipakai untuk deteksi cyberbullying(Zhao et al., 2016), deteksi bahasa kasar (Park & Fung, 2017), ulasan film (Jefferson et al., 2017) dan identifikasi cyberhate (Burnap & Williams, 2015).

Menggunakan dataset twitter sebagai data latih dan uji, penelitian ini menghasilkan analisis sentiment analysis menggunakan sastrawi sebagai stemmer, bigram sebagai fitur ekstraksi dan linear support vector machine sebagai sebagai model pembelajaran.

II. METODOLOGI PENELITIAN

1. Metodologi Penelitian

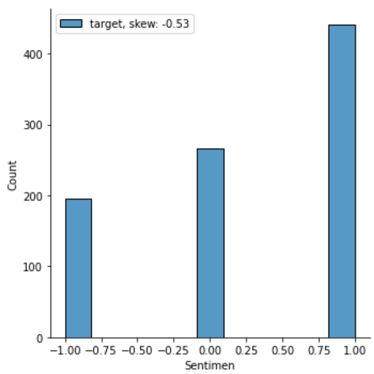
Alur penelitian merupakan proses dari pembuatan suatu system, dapat dilihat pada gambar berikut:

Gambar 1 Alur Penelitian

Tahapan pada Gambar 1 dapat dijelaskan pada sub-bab sebagai berikut:

a. **Mengambil dataset dari Twitter**

Dataset yang kami ambil, menggunakan dataset dari https://github.com/yahdiindrawan/covid19-sentiment-dataset. Disini, kami hanya mengambil label sentiment saja dimana terdapat 3 class, dengan jumlah positif sebanyak 441 tweet, netral sebanyak 266 tweet dan negative sebanyak 195 tweet seperti terlihat pada gambar 2.



Gambar 2 Sebaran Jumlah Class pada dataset Sentimen Covid 19

b. **Proses Penataan Dataset dan Pembersihan Element yang tidak digunakan**

Setelah data diambil dari twitter, tahap berikutnya adalah ditata dataset. Pada mulanya dataset yang diambil mempunyai karakter tambahan seperti '$', '?', '!', '#', '\'. Tujuan peneliti membersihkan karakter supaya data dapat diolah menggunakan Stemmer

c. **Stemming Menggunakan Sastrawi**

Stemming merupakan proses pemetaan untuk mendapatkan kata dasar dari berbagai bentuk. Proses ini dilakukan dengan menghilangkan awalan, akhiran, sisipan, atau kombinasi awalan dan akhiran yang terdapat disetiap kata. Disini peneliti menggunakan Sastrawi sebagai tool stemming dikarenakan Sastrawi cocok digunakan untuk teks Bahasa Indonesia. Contoh stemming dapat dideskripsikan sebagai berikut:

Tabel 1. Contoh Stemming menggunakan Sastrawi

Teks Awal	Sebaiknya hanya yang berkepentingan yang berkomentar tentang covid
Teks setelah Stemming	baik hanya yang penting yang komentar tentang covid

d. **Pembersihan Stop Words**

Setelah dataset destemming, tahap berikutnya adalah pembersihan stop words menggunakan library dari Sastrawi. Contoh pembersihan dapat dideskripsikan sebagai berikut:

Tabel 2. Contoh Pembersihan Stop Words Menggunakan Sastrawi

Teks Stemming	baik hanya yang penting yang komentar tentang covid
Teks setelah Stemming	baik penting komentar tentang covid

e. **Ekstrak Fitur Menggunakan bigram**

Pada pemrosesan dokumen, skema representasi teks biasanya menggunakan vector space model (VSM) yang sering digunakan untuk pembobotan kata (word-weighting). Hasil yang diterima dari VSM merupakan dokumen yang relevan. VSM yang digunakan disini menggunakan kata kunci atau frase, yang secara umum dikenal sebagai unigram, bigram, trigram dan n-gram(Figueiredo et al., 2011; Lee et al., 2012; Xie et al., 2017). Untuk lebih mudahnya, N-gram merupakan urutan dari kata-kata N. misalkan terdapat sebuah kalimat: "baik penting komentar tentang covid", maka jika dibuat n-gram didapatkan sebagai berikut:

Kata dasar	baik penting komentar tentang covid
Unigram	'baik', 'penting', 'komentar', 'tentang', 'covid'
Bigram	'baik penting', 'penting komentar', 'komentar tentang', 'tentang covid'
trigram	'baik penting komentar' 'penting komentar tentang', 'komentar tentang covid'

metode N-gram membuat keputusan dengan membandingkan nilai ini dengan ratio similaritas, yang didefinisikan sebagai ratio identic N-gram dibandingkan dengan jumlah total N-grams. Rasio similaritas dapat dihitung dengan cara (Gencosman et al., 2014):

ISBN : 978-623-6602-57-7

$$ratio\ similaritas = \frac{\delta}{\min(\alpha, \beta)} \qquad (1)$$

Dimana:

δ: jumlah n-gram yang identic

α: jumlah n-gram untuk $kata_A$

β: jumlah n-gram untuk $kata_B$

Disini $kata_A$ merupakan kata pertama dan $kata_B$ merupakan kata kedua yang digunakan sebagai pembanding karakter n-grams.

Saat ini, dikenal beberapa metode dalam pembuatan feature vector untuk tipe data dokumen, antara lain: Bag of Words (BoW) dan Term Frequency and Inverse Document Frequency (TF-IDF). BoW merupakan sebuah pendekatan algoritma yang menghitung seberapa banyak sebuah kata muncul (frekuensi) pada sebuah dokumen. Kelemahan BoW adalah urutan term dan kelangkaan (rareness) term tidak dipertimbangkan. Sedangkan pada TF-IDF, sebuah kata diberikan sebuah weight berupa TF dan IDF score, bukan berupa frekuensi seperti BoW.

Pada penelitian ini, peneliti menggunakan bigram dan TF-IDF sebagai fitur ekstraksi. Alasan penggunaan bigram adalah peneliti menggunakan pendekatan brute force untuk model pembelajaran dan ditemukan bigram lebih baik sebagai fitur ekstraksi.

f. **Klasifikasi Menggunakan Linear Support Vector Machine**

Setelah data diekstrak menggunakan bigram, tahapan berikutnya adalah menggunakan Linear Support Vector Machine sebagai model pembelajaran. Peneliti memilih Linear Support Vector Machine karena model pembelajaran tersebut cukup cepat untuk mendapatkan hasil. Disini peneliti menggunakan 3 fold cross validation. Kemudian, parameter yang peneliti gunakan adalah sebagai berikut:

Tabel 3. Parameter yang dipakai dalam penelitian

Nama Parameter	Nilai
Ekstraksi Fitur	Bigram
Nilai C pada SVM	10
Kernel	Linear

III. HASIL DAN PEMBAHASAN

Tahap berikutnya adalah menampilkan performa dari model pembelajaran yang peneliti lakukan. Rangkuman performa dapat ditampilkan sebagai berikut:

Tabel 4. Rangkuman Performa

Nama Performa	Nilai
Akurasi	99.13 %
Precision	99.14 %
Recall	99.13 %

Dari tabel 4, dapat disimpulkan, bahwa model pembelajaran yang peneliti usulkan mendapatkan nilai yang cukup memuaskan.

IV. KESIMPULAN

Deteksi sentimen menggunakan sentiment analysis berdasarkan bigram dan linear Support Vector Machine berhasil dikembangkan dengan performa yang cukup memuaskan yaitu 99% untuk akurasi precision dan recall. Tahapan berikutnya adalah penggunaan negation handling untuk mengetahui tingkat kewajaran Analisa suatu text.

V. UCAPAN TERIMA KASIH (Jika ada)

Penelitian ini didanai oleh LPPM Udinus pada Skema Penelitian Penelitian Dasar Perguruan Tinggi.

VI. REFERENSI

Burnap, P., & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, *97*, 106754. https://doi.org/10.1016/j.asoc.2020.106754

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr., W. (2011). Word co-occurrence features for text classification. *Information Systems*, *36*(5), 843–858. https://doi.org/10.1016/j.is.2011.02.002

Gencosman, B. C., Ozmutlu, H. C., & Ozmutlu, S. (2014). Character n-gram application for automatic new topic identification. *Information Processing & Management*, *50*(6), 821–856. https://doi.org/10.1016/j.ipm.2014.06.005

Jefferson, C., Liu, H., & Cocea, M. (2017). Fuzzy approach for sentiment analysis. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015577

Lee, L. H., Isa, D., Choo, W. O., & Chue, W. Y. (2012). High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic. *Expert Systems with Applications*, *39*(1), 1147–1155. https://doi.org/10.1016/j.eswa.2011.07.116

Park, J. H., & Fung, P. (2017). *One-step and Two-step Classification for Abusive Language Detection on Twitter*. 41–45. https://doi.org/10.18653/v1/w17-3006

Xie, F., Wu, X., & Zhu, X. (2017). Efficient sequential pattern mining with wildcards for keyphrase extraction. *Knowledge-Based Systems*, *115*, 27–39. https://doi.org/10.1016/j.knosys.2016.10.011

Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing and Networking - ICDCN '16*, 1–6. https://doi.org/10.1145/2833312.2849567