SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK V - Surabaya, 26 April 2025
Fakultas Teknik Elektro dan Teknologi Informasi, Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2025.7046

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Pengaruh Penggunaan Emoji Pada Tingkat Akurasi Sentimen Di Twitter Menggunakan Metode Support Vector Machine

Tio Dharmawan, Virli Galuh Kinanti, dan Achmad Maududie

Fakultas Ilmu Komputer, Universitas Jember
e-mail: tio.pssi@unej.ac.id

ABSTRACT –

Opinions and preferences expressed on social media and microblogging services are very important for sentiment analysis. A Support Vector Machine (SVM) is a learning system that uses a hypothetical space in the form of a linear function in a high dimensional feature space and applies a learning bias derived from statistical learning theory. The accuracy results obtained by the Support Vector Machine method from the first topic, namely booster vaccines as a homecoming requirement, were 65% for text only and 69% for text containing emoji. The accuracy results for the second discussion topic, namely demonstrations against Jokowi for 3 periods, were 79% for text only and 82% for text containing emoji. As for the third topic regarding the scarcity of cooking oil and rising fuel prices, the accuracy obtained is 74% for text only and 76% for text containing emojis.

Keywords: sentiment analysis, emoji characters, social media, SVM

ABSTRAK –

Opini dan preferensi yang diungkapkan di media sosial dan layanan microblogging sangat penting untuk analisis sentimen. *Support Vector Machine* (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi linier dalam ruang fitur berdimensi tinggi dan menerapkan bias pembelajaran yang berasal dari teori pembelajaran statistik. Penelitian tentang pengaruh penggunaan emoji terhadap tingkat akurasi sentimen di Twitter dengan metode Support Vector Machine dilakukan dengan membandingkan data training dan testing 80:20 untuk tiga topik bahasan. Hasil akurasi yang diperoleh dengan metode *Support Vector Machine* dari topik pertama yaitu vaksin booster sebagai persyaratan mudik didapatkan 65% untuk teks saja dan 69% untuk teks yang berisi emoji. Hasil akurasi untuk topik pembahasan kedua yaitu demonstrasi menentang Jokowi selama 3 periode, adalah 79% untuk teks saja dan 82% untuk teks yang berisi emoji. Sedangkan untuk topik ketiga mengenai kelangkaan minyak goreng dan kenaikan harga BBM, akurasi yang didapat adalah 74% untuk teks saja dan 76% untuk teks yang berisi emoji.

Kata kunci: analisis sentimen, karakter emoji, media sosial, SVM

PENDAHULUAN

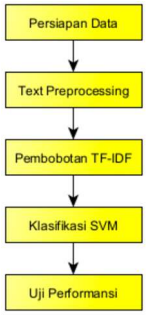
Kementerian Komunikasi dan Informatika (Kemenkominfo) mengungkapkan pengguna internet di Indonesia saat ini mencapai 63 juta orang. Dari angka tersebut, 95% menggunakan internet untuk mengakses jejaring sosial. Indonesia menduduki peringkat ke-6 dengan jumlah pengguna twitter 15,7 juta [1]. Twitter adalah situs mikroblog yang memungkinkan pengguna untuk mengirim pesan teks dengan jumlah karakter maksimum melalui SMS, pesan instan, surat elektronik [2]. Setiap tweet ditampilkan dalam bentuk pengirim dengan foto profil, alamat balasan pesan dengan tanda "@" dan tanggal.

Karakter emoji dapat meningkatkan ekspresi pesan teks [3]. Emoji memainkan peran penting dalam konten tweet yang dikirim pengguna di Twitter. Shigetaka Kurita mengembangkan emoji pada tahun 1990 dengan tujuan awal untuk mempercantik pesan. Ideogram dan smiley memungkinkan pengguna untuk lebih mudah mengekspresikan emosi mereka melalui teks. Opini dan preferensi yang diungkapkan di media sosial dan layanan microblogging sangat penting untuk analisis sentimen dan pengumpulan opini. Analisis sentimen atau lebih sering disebut opinion mining adalah bidang penelitian text mining yang bertujuan untuk memastikan opini atau subjektivitas masyarakat umum mengenai subjek, peristiwa, atau masalah. Salah satu analisis sentimen yang dapat dilakukan adalah analisis sentimen terhadap suatu isu yang terdapat di Twitter [4].

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi linier dalam ruang fitur berdimensi tinggi dan menerapkan bias pembelajaran yang diturunkan dari teori pembelajaran statistic [5]. SVM bertujuan untuk menemukan hyperplane “terbaik” yang berperan penting sebagai garis batas dua kelas. Penelitian ini ingin melihat bagaimana pengaruh penggunaan emoji terhadap tingkat akurasi sentimen di Twitter dengan metode *Support Vector Machine* (SVM).

METODE

Metode penelitian merupakan langkah serta prosedur yang hendak dicoba untuk menggapai tujuan serta memperoleh jawaban atas kasus dalam penelitian. Langkah serta prosedur ini adalah perwujudan dari kerangka pikir penelitian. Langkah-langkah tersebutdapat dilihat pada Gambar 1 di bawah.



Gambar 1. Tahapan penelitian

Pengumpulan Data

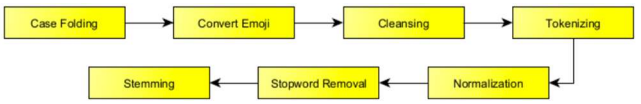
Pada tahap awal pengumpulan data *tweet*, pengambilan data dilakukan dengan metode *crawling* dari media sosial twitter menggunakan *python*. Sebelum proses *crawling* diperlukan *API Key* dari twitter sebagai autentikasi di *python*. Twitter API Key sendiri memiliki kunci konsumen, rahasia konsumen, kunci akses, dan rahasia akses. Pada tahap pendataan dilakukan dengan menggores halaman pencarian Twitter. Ada tiga topik pembahasan yang terkumpul yaitu vaksin booster sebagai syarat mudik, demonstrasi menentang Jokowi selama 3 periode, dan kelangkaan minyak goreng dan naiknya harga bahan bakar.

Pada tahap ini data yang diperoleh kemudian dikelompokkan menurut tiga topik yang berbeda. Selanjutnya, dari ketiga topik tersebut, datanya dipisahkan antara teks saja dan teks yang mengandung emoji. Proses pelabelan data dilakukan secara manual.Pelabelan manual dibantu oleh mahasiswa jurusan Pendidikan Bahasa dan Sastra Indonesia di Universitas Jember. Pemilihan Mahasiswa Jurusan Pendidikan Bahasa dan Sastra Indonesia sebagai validator data yang telah diberi label manual karena telah sesuai dengan peminatan Pendidikan dan Pengembangan sehingga dapat memberikan penilaian yang tepat terhadap tweet yang layak konsumsi pada konten tweet. Proses ini menghasilkan tiga topik yang masing-masing berisi sentimen positif dan negatif tanpa emoji, serta sentimen positif dan negatif untuk teks dan emoji.

Preprocessing

Text preprocessing merupakan tahap awal pada analisis di mana data teks yang akan di preprocessing ini mencakup tahapan pengambilan data tweets dari twitter hanya berupa data mentah yang belum dilakukan preprocessing guna mendapatkan data bersih agar memudahkan ke tahap berikutnya [6]. Pada tahap preprocessing ini terdiri dari beberapa tahapan untuk menjadikan kalimat pada tweets menjadi bahasa yang baku, karena tidak sepenuhnya tweets tersebut hasil crawling menggunakan kata baku. Proses ini berguna menghilangkan beberapa bagian kalimat yang tidak berguna. Preprocessing dilakukan dengan menggunakan bantuan

library bahasa pemrograman python. Ada 7 tahapan proses untuk mendapatkan hasil yang maksimal. Langkah-langkah preprocessing dapat dilihat pada Gambar 2 di bawah.



Gambar 2. Tahapan preprocessing

Tahap *text preprocessing* yang pertama kali dilakukan adalah *case folding.* Pada tahap ini semua huruf dalam dokumen menjadi huruf kecil, hanya huruf 'a' sampai dengan 'z' [7]. Tahap selanjutnya adalah *convert emoji.* Pada tahap ini setiap *emoji* akan dikonversikan ke dalam string yang sesuai. Mengubah *emoji* menjadi teks dengan *python* dapat dilakukan menggunakan modul *emoji*. Selanjutnya akan dilakukan proses *cleansing, Cleansing* adalah proses untuk membersihkan dokumen dan kata-kata yang tidak diperlukan. Pada tahap ini dilakukan penghapusan username, URL (*Uniform Resource Locator*), dan "RT" (tanda *retweets*) [7]. Karena *username*, URL, dan "RT" tidak memiliki pengaruh apapun terhadap nilai sentimen, maka ketiga komponen di atas akan dibuang.

Setelah data melalui proses *cleansing*, selanjutnya akan melalui proses *tokenizing*. *Tokenizing* adalah proses pemotongan sebuah dokumen menjadi bagian-bagian, yang disebut dengan token [8]. Setelah melalui proses tokenizing kita bisa mendapatkan jumlah kemunculan setiap token nya. Setelah data melalui proses tokenizing, selanjutnya akan melalui proses normalization. *Normalization* adalah suatu proses mengubah kata yang tidak baku ke dalam kata baku atau formal. Contohnya seperti "abis", "adlh", "akikah", "akyu", "asek", "asoy", dan lain-lain.

Setelah data melalui proses *normalization*, selanjutnya akan melalui proses *stopword removal*. *Stopword removal* adalah proses penghilangan kata-kata yang tidak berkontribusi banyak pada isi dokumen. Pada tahap *stopword removal* Bahasa Indonesia dapat dibantu dengan menggunakan *library* Sastrawi. Sastrawi merupakan *library* yang dapat digunakan untuk mendapatkan kata dasar dari kata yang kita inputkan serta *library* ini juga mendukung proses *stopword removal* [8]. Tahap terakhir adalah *stemming*. *Stemming* merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. Proses untuk melakukan *stemming* dibantu dengan menggunakan Algoritma Nazief Andriani [9]. Algoritma Nazief Andriani merupakan salah satu algoritma yang cukup populer untuk melakukan stemming kata dalam Bahasa Indonesia [8].

Pembobotan TF-IDF

Pembobotan dilakukan untuk mendapatkan nilai kata yang diekstraksi. Metode yang digunakan adalah TF-IDF (Term Frequency–Inverse Document Frequency). Ini adalah kombinasi dari Term Frequency (TF) dan Inverse document Frequency (IDF) yang digunakan untuk menghitung bobot setiap kata. Metode ini menggabungkan dua konsep untuk menghitung bobot, yaitu Term Frequency (TF) adalah frekuensi kemunculan kata (t) dalam kalimat (d). Document Frequency (DF) adalah jumlah kalimat di mana kata (t) muncul [10].

Pemodelan Menggunakan SVM

Pada tahap ini pembagian data dimulai, dimana data yang sudah preprocessed dan telah melalui proses pembobotan TF-IDF akan dibagi menjadi data latih dan data uji. Data latih akan dijadikan sebagai data yang akan melatih algoritma SVM. Pada tahap ini data latih yang sudah terbentuk dihitung klasifikasi sentimennya menggunakan Support Vector Machine.

Klasifikasi SVM dilakukan dengan pencarian grid menggunakan kernel linier dengan validasi silang k = 10 sehingga dipilih nilai parameter C, derajat, gamma dan kernel yang optimal. Algoritma dalam metode ini menjadikan titik-titik grid yang berjarak sama, lalu menghitung ukuran kesalahan untuk setiap titik-titik parameter tersebut. Titik parameter paling optimal adalah titik yang memiliki nilai ukuran kesalahan yang terkecil [11]. Grid search pada penelitian ini digunakan untuk mencari hyperparameter yang optimum atau terbaik untuk setiap kernel yang berbeda. Pada tahap ini implementasi grid search dibantu oleh library sklearn GridSearch. Hasil dari tahapan ini adalah model klasifikasi. Kemudian model klasifikasi yang telah terbentuk dari data latih diterapkan pada data uji.

Evaluasi

Tahap ini melibatkan analisis data uji dan model klasifikasi yang telah terbentuk dari proses sebelumnya. Hasil yang diperoleh dari pengujian yang dilakukan dengan menggunakan accuracy, precision, recall, dan f1-score [12], kemudian akan dibandingkan dengan model klasifikasi untuk setiap topik dataset yang dikumpulkan. Model pengujian tersebut dapat dibandingkan dengan hasil pada beberapa skenario yang dilakukan [13].

HASIL DAN PEMBAHASAN

Hasil pemodelan pada Tabel 1 menunjukkan perbandingan kinerja antara dua jenis representasi teks: teks biasa dan teks yang disertai dengan emoji. Pengujian dilakukan dengan tujuan untuk melihat pengaruh penggunaan emoji pada data latih terhadap performa model dalam hal beberapa metrik evaluasi.

Dari sisi *accuracy*, model dengan teks biasa menunjukkan nilai 0.65 pada Topik 1, sementara teks dengan emoji sedikit lebih baik dengan nilai 0.69. Pada Topik 2, peningkatan lebih jelas terlihat, dengan model yang menggunakan teks dengan emoji mencapai 0.82 dibandingkan dengan 0.79 pada teks biasa. Hal serupa terjadi pada *precision*, di mana model dengan emoji menunjukkan peningkatan yang signifikan, dari 0.64 menjadi 0.68 pada Topik 1, dan dari 0.75 menjadi 0.82 pada Topik 2.

Untuk *recall*, hasil yang serupa terlihat, dengan model yang menggunakan emoji sedikit lebih unggul, meningkatkan recall dari 0.65 menjadi 0.69 pada Topik 1 dan 0.74 menjadi 0.76 pada Topik 2. Metrik *f1-score*, yang menggabungkan precision dan recall, menunjukkan hasil terbaik pada model dengan emoji, dengan nilai 0.65 dibandingkan dengan 0.60 pada Topik 1, dan 0.77 dibandingkan dengan 0.74 pada Topik 2.

Secara keseluruhan, penggunaan emoji dalam teks memberikan peningkatan pada hampir semua metrik evaluasi, menunjukkan bahwa penambahan emoji dapat berkontribusi positif dalam meningkatkan performa model dalam analisis data latih.

Tabel 1. Perbandingan performa model

	Topik 1		Topik 2		Topik 2	
	Teks	**Teks+Emoji**	**Teks**	**Teks+Emoji**	**Teks**	**Teks+Emoji**
Accuracy	0.65	0.69	0.79	0.82	0.74	0.76
Precision	0.64	0.68	0.75	0.82	0.76	0.76
Recall	0.65	0.69	0.79	0.82	0.74	0.76
F1-score	0.6	0.65	0.74	0.77	0.7	0.74

KESIMPULAN

Kesimpulan dari hasil pemodelan ini adalah bahwa penggunaan emoji dalam teks dapat meningkatkan performa model dalam berbagai metrik evaluasi, seperti accuracy, precision, recall, dan F1-score. Model yang menggunakan teks dengan emoji menunjukkan peningkatan yang signifikan dibandingkan dengan model yang hanya menggunakan teks biasa, terutama pada Topik 2. Peningkatan ini menunjukkan bahwa emoji dapat memperkaya konteks informasi dalam teks, sehingga model dapat lebih akurat dalam mengklasifikasikan dan menilai data latih. Oleh karena itu, penggunaan emoji dalam analisis teks dapat dipertimbangkan sebagai cara yang efektif untuk meningkatkan kinerja model.

DAFTAR PUSTAKA

[1] Kemenkominfo, “Pengguna internet di Indonesia 63 Juta Orang,” 2013.

[2] M. Hadi, *Twitter untuk orang awam*. Palembang: Maxikom, 2010.

[3] M. Shiha and S. Serkan, “The Effects of Emoji in Sentiment Analysis,” *Bahcesehir University*, 2017, doi: 10.17706/IJCEE.2017.9.1.360-369.

[4] M. I. Fikri, T. S. Sabrila, and Y. Azhar, “Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter,” *Smatika Jurnal*, vol. 10, no. 02, pp. 71–76, 2020.

[5] F. S. Jumeilah, “Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian,” *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[6] A. M. Zuhdi, E. Utami, and S. Raharjo, “Analisis Sentimen Twitter terhadap Capres Indonesia 2029 dengan Metode K-NN,” *Jurnal Informa Politeknik Indonusa Surakarta*, vol. 5, pp. 2–7, 2019.

[7] D. Rustiana and N. Rahayu, “Analisis Sentimen Pasar Otomotif Mobil: Tweets Twitter Menggunakan Naïve Bayes,” *Jurnal SIMETRIS*, vol. 8, no. 1, Apr. 2017.

[8] D. Indraloka, D. Smaradahana, and B. Santosa, “Penerapan Text Mining untuk Melakukan Clustering Data Tweets Shopee Indonesia,” *Jurnal Sains Dan Seni ITS*, vol. 6, no. 2, pp. 2337–3520, 2017.

[9] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confix-stripping approach," *ACM Transactions on Asian Language Information Processing*, vol. 6, no. 4, p. 13, 2007, doi: 10.1145/1316457.1316459.

[10] M. M. Sya'bani and R. Umilasari, "Penerapan Metode Cosine Similarity dan Pembobotan TF/IDF pada Sistem Klasifikasi Sinopsis Buku di Perpustakaan Kejaksaan Negeri Jember," 2018.

[11] D. Satriani, L. U. Khasanah, and N. A. Rizki, "Penerapan Metode Grid-Search dalam Menentukan Parameter Model Pertumbuhan Penduduk di Kota Samarinda," in *Prosiding Seminar Nasional Matematika dan Statistika*, May 2019, pp. 65–74.

[12] P. Flach, *Machine Learning: The Art and Science of Algorithms that Makes Sense of Data*. New York: Cambridge University Press, 2012.

[13] V. Sari, F. Firdausi, and Y. Azhar, "Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes," *Edumatic: Jurnal Pendidikan Informatika*, vol. 4, no. 2, pp. 1–9, 2020, doi: 10.29408/edumatic.v4i2.2202.

Halaman ini sengaja dikosongkan