Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 1. August 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Analisis Sentimen Opini Berbahasa Indonesia Pada Sosial Media Menggunakan TF-IDF dan Support Vector Machine

Putu Ayu Novia Aryanti[1],
Ida Bagus Made Mahendra[2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana,
Bali, Indonesia
[1]niputunovia@gmail.com
[2]ibm.mahendra@unud.ac.id

Abstract

Dengan berkembangnya media sosial, telah menjadi forum yang memberikan kebebasan berekspresi bagi semua individu. Beragam pendapat disampaikan, mulai dari yang positif, netral, hingga yang negatif. Penelitian ini akan mengklasfikasikan opini pada sosial media ke dalam tiga kategori, yaitu positif, netral, dan negatif, dengan *Support Vector Machine* (SVM) yang dikombinasikan dengan pemanfaatan rekayasa fitur TF-IDF. terdapat beberapa tahapan yang dilakukan yaitu pengumpulan data yang bersumber dari data publik IndoNLU, *text cleaning* dan *data pre-processing*, rekayasa fitur TF-IDF, *modeling* SVM dengan tiga jenis *kernel*, yaitu linear, rbf, dan sigmoid serta evaluasi model dengan menghitung nilai presisi, *recall*, *f-1 score*, akurasi. Dilihat berdasarkan pengujian beberapa jenis kernel yang menghasilkan akurasi paling tinggi adalah kernel RBF dengan tingkat akurasi 88%. Analisis sentimen yang dilakukan pada 1260 teks data uji dengan menggunakan SVM dan *kernel* RBF, menghasilkan klasifikasi positif sebanyak 741 teks sebesar 59%, negatif sebanyak 422 teks sebesar 33%, dan netral sebanyak 97 sebesar 8%. Ini menunjukkan sentimen positif yang paling mendominasi dibandingkan sentimen negatif dan sentimen netral. Sentimen positif lebih banyak membahas mengenai ulasan makanan atau restoran dan kebijakan pemerintah. Sementara sentimen negatif bersifat lebih variatif dan sentimen netral hanya berbagi berita tanpa berkomentar.

Keywords: Opini, Analisis Sentimen*, Support Vector Machine, TF-IDF, Natural Language Processing*

1. Pendahuluan

Sosial media kini menjadi salah satu jenis teknologi informasi yang berkembang cukup pesat. Pengertian dari sosial media adalah perangkat atau alat komunikasi dan kolaborasi yang memungkinkan terjadinya interaksi antara berbagai individu di berbagai belahan dunia yang sebelumnya mustahil untuk dilakukan. [1] Melihat perkembangan sosial media yang semakin meningkat, peranan sosial media kini menjadi salah satu wadah yang memberikan kebebasan bagi setiap individu yang menggunakannya. Berbagai opini atau gagasan pikiran dikemukakan dalam berbagai platform sosial media, mulai dari yang bersifat positif, netral, bahkan negatif.

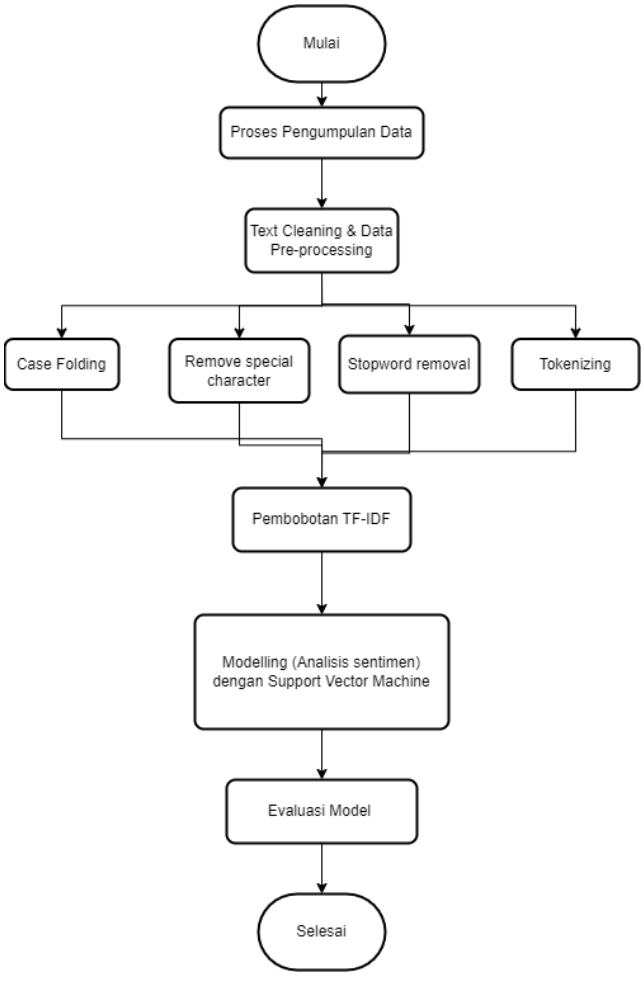
Dengan munculnya berbagai opini tersebut, pendekatan informatika terkait pemrosesan bahasa alami atau lebih dikenal dengan *Natural Language Processing* (NLP) akan terus diperlukan. Penerapan NLP telah dilakukan dalam berbagai bidang, seperti bisnis, pendidikan, keuangan, perawatan kesehatan, dan layanan pelanggan. Salah satu penerapan NLP adalah *sentiment analysis* atau analisis sentimen. *Sentiment analysis* merupakan proses mengkategorikan sekumpulan teks atau dokumen ke dalam kategori tertentu. Terdapat banyak metode yang bisa digunakan menyelesaikan permasalahan klasifikasi data tekstual, diantaranya *Logistic Regression, Naïve Bayes*, *K-Nearest*

Neighbor (KNN), *Support Vector Machine* (SVM), dan lainnya. Beberapa metode tersebut telah digunakan pada penelitian – penelitian klasifikasi sebelumnya. Penelitian yang menganalisis ulasan pengguna pada aplikasi *Google Meet* menggunakan metode SVM dan *Logistic Regression*, dengan akurasi terbaik menggunakan metode SVM kernel Linear sebesar 87,02% dibandingkan *Logistic Regress*ion dengan akurasi sebesar 85,17%. [2] Penelitian lainnya menganalisis sentimen pada review aplikasi Grab menggunakan SVM memperoleh tingkat akurasi sebesar 85,54%. [3] Dan terdapat juga penelitian yang membandingkan metode SVM, *Naïve Bayes*, dan *Logistic Regression* pada analisis sentimen aplikasi tokopedia, dengan kinerja klasifikasi terbaik diperoleh dari metode klasifikasi SVM. [4]

Berdasarkan penelitian yang telah dilakukan sebelumnya, pada penelitian ini akan dilakukan analisis sentimen terhadap opini pada sosial media menggunakan *Support Vector Machine* (SVM) yang dikombinasikan dengan pemanfaatan *feature engineering* TF-IDF. Penggunaan metode SVM dalam penelitian ini untuk mengklasifikasikan opini ke dalam tiga kategori, yaitu positif, netral, dan negatif.

2. Metode Penelitian

Dalam penelitian ini terdapat beberapa tahapan yang akan dilakukan untuk dapat menganalisis sentimen opini berbagasa Indonesia pada sosial media. Tahapan tersebut dapat digambarkan sebagai berikut.



Gambar 1. Tahapan Analisis Sentimen

p-ISSN: 2301-5373
e-ISSN: 2654-5101

2.1 Proses Pengumpulan Data

Terdapat berbagai teknik proses pengumpulan data, seperti menggunakan data publik yang tersedia di berbagai platform daring, pengumpulkan data melalui web atau internet (*web scraping*), dan penggabungan beberapa data tertentu (*data augmentation*). Dataset pada penelitian ini berasal dari data publik, yang bersumber dari repository GitHub IndoNLU. IndoNLU (*Indonesian Natural Language Understanding*) merupakan kumpulan sumber data yang digunakan dalam melatih, mengevaluasi, dan menganalisis sistem pemrosesan bahasa alami untuk Bahasa Indonesia. IndoNLU dikembangkan oleh beberapa penggemar NLP Indonesia dari berbagai institusi, diantaranya Gojek, Institut Teknologi Bandung (ITB), HKUST, Universitas Multimedia Nusantara, Prosa.ai, dan Universitas Indonesia (UI). IndoNLU mencakup 12 tugas, yang dibagi dalam empat kategori. Salah satu kategori membahas data yang berhubungan dengan analisis sentimen atau *sentiment analysis*, yaitu SmSA (*sentence-level sentiment analysis dataset*). SmSA berisi sekumpulan komentar dan ulasan dalam bahasa Indonesia yang diperoleh dari beberapa *platform online* ataupun sosial media. Dataset ini terdiri dari 12.260 teks komentar, ulasan, dan opini yang telah dikategorikan dalam tiga kemungkinan, yaitu positif, netral, dan negatif. [5]

2.2 Text Cleaning & Data Pre-Processing

Text cleaning dan *data pre-processing* merupakan proses menyiapkan dataset sebelum dilakukan pemodelan. Tahapan ini adalah tahapan yang penting dalam analisis sentiment, bertujuan untuk menghapus atau menghilangkan *missing value* (data kosong), data ganda, dan memperbaiki format data yang tidak sesuai. Beberapa tahapan yang dilakukan pada *text cleaning* dan *data pre-processing* diantaranya:

- ***Case folding***, mentransformasikan seluruh data pada teks atau opini pada dataset ke dalam huruf kecil. Hal ini bertujuan agar seluruh data memiliki format yang sama dan dapat memudahkan pemrosesan data saat tahap pemodelan.
- ***Remove special character***, penghapusan simbol (seperti tanda baca, nilai tukar mata) selain huruf dari teks ada dataset.
- ***Stopword removal***, penghapusan kata - kata yang memiliki sedikit makna pada teks, seperti kata penghubung, kata ganti, dan sebagainnya.
- ***Tokenizing***, tahapan pemisahan kata pada dataset berdasarkan *whitespace*.

2.3 Feature Engineering TF-IDF

Feature engineering atau rekayasa fitur dapat diartikan sebagai proses pengkodean data teks ke dalam bentuk data numerik sehingga dapat digunakan dalam pemodelan algoritma *machine learning*. Rekayasa fitur dalam representasi teks dibedakan dalam berbagai kategori dan metode. Terdapat empat kategori dengan beberapa metode, diantaranya: [6]

- ***Basic vectorization approaches***, dengan metode One-Hot Encoding, Bag of Words, Bag of N-Grams, TF-IDF;
- ***Distributed representations***, dengan metode *Words Embedding*;
- ***Universal text representations****,* dengan metode *pre-trained embedding*; dan
- ***Handcrafted features****,* dengan metode *domain-specific knowledge*.

Dari beberapa kategori, *basic vectorization approaches* yang paling sering digunakan dalam analisis sentimen. Dalam penelitian ini, rekayasa fitur TF-ID (*Term Frequency-Inverse Document Frequency*) akan digunakan untuk pembobotan teks. Fitur TF-IDF digunakan untuk menentukan besar kepentingan kata dalam suatu dokumen atau teks. Berdasarkan kutipan dari buku "*Practical Natural Language Processing*", perhitungan pembobotan pada TF-IDF digambarkan dengan analogi: “Jika suatu istilah muncul secara berkala dalam teks pertama namun tidak muncul dalam teks

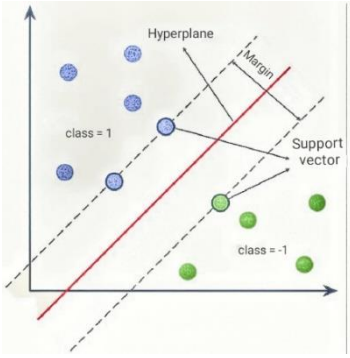
kedua atau teks lainnya, maka istilah tersebut memiliki makna yang penting untuk teks pertama". [7]
Secara matematis metode TF-IDF dapat direprsentasikan dalam dua besaran sebagai berikut.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

- $tf_{i,j}$ merupakan TF (*Term Frequency*), mengukur seberapa sering suatu kata muncul pada teks. Dihitung dengan membagi jumlah kata yang muncul dengan total kata pada teks.
- df_i meruapakan IDF (*Inverse Document Frequency*), mengukur pentingnya istilah di setiap korpus. Berperan untuk mempertimbangkan term yang sangat umum muncul dan jarang dalam teks.

2.4 Modelling dengan Support Vector Machine

Support Vector Machine (SVM) pertama kali diperkenalkan pada tahun 1992 oleh Vapnik sebagai konsep utama di bidang pengenalan pola (memetakan data ke dua atau lebih kelas atau kategori yang telah ditentukan sebelumnya). *Support vector machine* adalah algoritma pembelajaran mesin yang dapat digunakan untuk menyelesaikan masalah klasifikasi, regresi, dan deteksi outlier. [6]
Konsep SVM dengan mudah dapat dijelaskan sebagai upaya menemukan *hyperplane* optimal yang bertindak layaknya pemisah antara dua kelas atau lebih di suatu ruang input. Masalah klasifikasi dapat ditransformasikan dengan mencoba mencari *hyperplane* yang memisahkan dua kelompok. Untuk dapat menemukan *hyperplane* pemisah yang optimal, dilakukan pengukuran margin *hyperplane* tersebut dan menghitung titik maksimalnya. *Margin* adalah jarak antara *hyperplane* dan pola terdekat di setiap kelas. Pola yang paling dekat dengan *hyperplane* disebut *support vector*. Usaha untuk menemukan *hyperplane* ini merupakan inti dari proses pembelajaran di SVM. [6]



Gambar 2. Ilustrasi *Hyperplane* Optimal SVM

2.5 Evaluasi Model

Matriks digunakan untuk mengukur performa model *machine learning*. Dalam masalah klasifikasi, hasil perhitungan performa model dapat diringkas dalam *confusion matrix*, yang membagi hasil tes data sampel menjadi empat kategori tergantung *true label* dan *predicted tabel.* [8]

Tabel 1. *Confusion matrix*

	True Label	
	Positive	Negative

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Predicted Label	Positive	TP	FP
	Negative	FN	TN

- *True Positive* (TP), sampel data bernilai positif dan model memprediksi dengan benar.
- *True Negative* (TN), sampel data bernilai negatif dan model memprediksi dengan benar.
- *False Positive* (FP), sampel data bernilai positif tetapi model melakukan prediksi yang salah.
- *False Negative* (FN), sampel data bernilai negatif tetapi model melakukan prediksi yang salah.

Dalam penelitian ini, akan digunakan melihat matriks *precision, recall, F1-score*, dan *accuracy* pada model SVM, yang dapat direpresentasikan sebagai berikut.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F_1 - score = \frac{2TP}{2TP+FP+FN} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

3. Hasil dan Pembahasan

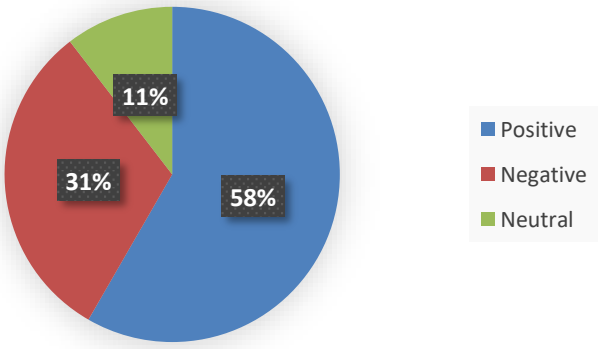
3.1 Hasil Analisis Data

Hasil pengumpulan data IndoNLU berisi 12.260 teks yang dikelompokkan dalam file data *training* dan data *testing*. Setiap teks telah dilabeli atau diklasifikasikan dalam tiga kategori, yaitu positif, netral, dan negatif. Informasi lebih lanjut terkait dataset adalah sebagai berikut.

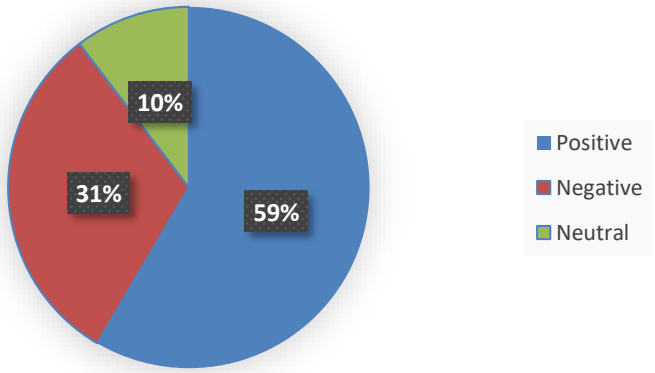
Tabel 2. Informasi jumlah data IndoNLU

	train	test
positive	6416	735
negative	3436	394
neutral	1148	131
Total	**11000**	**1260**

Jika divisualisasikan dalam bentuk diagram pie, perbandingan banyak data teks pada data train dan data test berdasarkan kategori dapat dilihat sebagai berikut.



Gambar 3. Diagram kategori data *training*



Gambar 4. Diagram kategori data *testing*

3.2 Hasil Pengujian *Kernel*

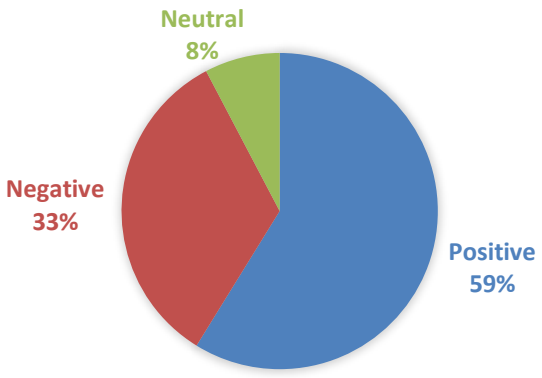
Setelah dilakukan analisis data, data *training* akan melalui tahapan *pre-processing* dan pembobotan TF-IDF yang dijadikan sebagai data pembelajaran algoritma untuk mengklasifikasikan data *testing* dengan metode *Support Vector Machine* (SVM). Proses klasifikasi akan membandingkan tiga jenis kernel SVM, yaitu *kernel* linear, *kernel* RBF, dan *kernel* sigmoid. Penentuan hasil analisis sentimen dilihat dari tingkat akurasi yang paling tinggi.

Tabel 3. Persentase Hasil Pengujian dengan Kernel Berbeda

Kernel	*Accuracy*	*Weighted Mean Precision*	*Weighted Mean Recall*	*Weighted Mean F-1 Score*
linear	87,22%	87,27%	87,22%	87,19%
RBF	88%	88,06%	87,94%	87,79%
sigmoid	87,46%	87,50%	87,46%	87,43%

Berdasarkan persentase hasil pengujian pada Tabel 3, dapat dilihat jenis kernel yang menghasilkan akurasi paling tinggi adalah kernel RBF dengan tingkat akurasi 88%. Sehingga hasil analisis sentimen yang digunakan berasal dari klasifikasi kernel RBF seperti pada Gambar 5.

3.3 Hasil Analisis Sentimen



Gambar 5. Hasil Analisis Sentimen

Analisis sentimen yang dilakukan pada 1260 tekst data *testing* dengan menggunakan SVM dan *kernel* RBF, menghasilkan klasifikasi positif sebanyak 741

p-ISSN: 2301-5373
e-ISSN: 2654-5101

teks sebesar 59%, negatif sebanyak 422 teks sebesar 33%, dan netral sebanyak 97 sebesar 8%. Sentimen positif lebih banyak membahas mengenai ulasan makanan atau restoran dan beberapa tentang kebijakan pemerintah. Sementara sentiment negatif bersifat variatif, seperti ulasan atau opini tentang pemerintahan, film, restoran, provider, aplikasi dan sebagainya. Serta sentimen netral hanya berbagi berita dari media massa tanpa berkomentar.

4. Kesimpulan

Algoritma Support Vector Machine (SVM) dapat diterapkan dalam mengklasifikasi sentimen opini berbahasa Indonesia pada media sosial dengan tiga kategori, yaitu positif, negatif, dan netral. Hasil pengujian beberapa *kernel* pada SVM menunjukkan bahwa *kernel* RBF yang paling baik digunakan dalam analisis sentiment kali ini. Dengan tingkat akurasi tertinggi yang diperoleh sebesar 88%. Serta hasil klasifikasi menunjukkan sentimen positif yang paling mendominasi dibandingkan sentimen negatif dan sentimen netral dengan persentase, yaitu sentimen positif (59%), sentimen negatif (33%), dan sentimen netral (8%).

Referensi

[1] C. Brogan, Social Media 101: Tactics and Tips to Develop your Bussiness Online, Jhon Wiley & Soins, 2010.

[2] A. Novantika and S. Sugiman, "Analisis Sentimen Ulasan Pengguna Aplikasi Video Conference Google Meet menggunakan Metode SVM dan Logistic Regression.," *PRISMA, Prosiding Seminar Nasional Matematika,* vol. 5, pp. 808-813, 2022.

[3] R. Wahyudi and G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *Jurnal Informatika ,* vol. 8, no. 2, pp. 200-207, 2021.

[4] M. Putri and I. Kharisudin, "Analisis Sentimen Pengguna Aplikasi Marketplace Tokopedia Pada Situs Google Play Menggunakan Metode Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression," *PRISMA, iProsiding Seminar Nasional Matematika,* vol. 5, pp. 759-766, 2022.

[5] B. Wilie, K. Vincentio, G. Winata, Cahyawijaya, Samuel, X. Li and S. Bahar, IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding, Suzhou: Association for Computational Linguistics, 2020.

[6] A. S. Nugroho, A. B. Witarto and D. Handoko, Support Vector Machine – Teori dan Aplikasinya dalam Bioinformatika, 2003.

[7] S. Vajjala, Majumder, Bodhisattwa, Gupta, Anuj, Surana and Harshit, Practical Natural Language Processing, O'Reilly Media, Inc., 2020.

[8] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," *O. Colliot (Ed.), Machine Learning for Brain Disorders, Springer,* 2022.

Halaman ini sengaja dibiarkan kosong