Jurnal Ilmiah Wahana Pendidikan, Oktober 2023, 9 (20), 59-65
DOI: https://doi.org/10.5281/zenodo.8398895
p-ISSN: 2622-8327 e-ISSN: 2089-5364
Accredited by Directorate General of Strengthening for Research and Development
Available online at https://jurnal.peneliti.net/index.php/JIWP

Analisis Sentimen Data Twitter Topik Ekonomi Dan Industri Dengan Metode Naive Bayes Dan Random Forest

Aji Susanto[1], Iskandar Agung Dzulkarnain[2]

[1,2] Program Studi Statistika, Fakultas Matematika Dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret

Received: 15 September 2023
Revised : 22 September 2023
Accepted: 29 September 2023

Abstract

Twitter has become a valuable source of information, and sentiment analysis can provide insights into public views and attitudes towards economic and industrial issues. This research aims to develop and compare the performance of two widely used classification methods, Naive Bayes and Random Forest, for sentiment analysis on Twitter data related to the economy and industry. By addressing the existing knowledge gap in sentiment analysis using Naive Bayes and Random Forest, this study provides a clear framework that empowers companies to efficiently process and leverage Twitter data, yielding valuable decision-making insights in the realm of economy and industry. A total of 11,833 data were divided into 70% training data and 30% testing data then classified using Naive Bayes, and Random Forest algorithms. The calculation results show positive sentiment of 28,52%, negative sentiment of 31,44%, and neutral sentiment of 40,04%. The comparison of the two algorithms obtained using Naïve Bayes gets the highest accuracy of 71,89%.

Keywords: *twitter, sentiment analysis, naive bayes, random forest, economic issues, industrial issues*

(*) Corresponding Author: ajisusanto@student.uns.ac.id

How to Cite: Susanto A, & Dzulkarnain I A. (2023). Analisis Sentimen Data Twitter Topik Ekonomi Dan Industri Dengan Metode Naive Bayes Dan Random Forest. https://doi.org/10.5281/zenodo.8398895

PENDAHULUAN

Latar Belakang

Era digital yang terus berkembang saat ini telah membuat media sosial menjadi sumber informasi yang penting dalam berbagai aspek kehidupan, termasuk ekonomi dan industri. Twitter, sebagai *platform* media sosial yang populer, memberikan ruang bagi pengguna untuk berbagi pendapat, pemikiran, dan perasaan terkait topik tertentu, termasuk ekonomi dan industri. Dengan jutaan cuitan yang dibagikan setiap hari, Twitter menjadi sumber data yang berharga untuk menganalisis sentimen publik.

Analisis sentimen adalah metode yang digunakan untuk mengidentifikasi, memperoleh, dan mengevaluasi sentimen, opini, atau emosi yang terkandung dalam teks. Dalam konteks ekonomi dan industri, analisis sentimen dapat membantu memahami pandangan dan sikap publik terhadap isu-isu, perusahaan, produk, atau tren tertentu. Informasi ini sangat berharga bagi pengambil keputusan, termasuk pemerintah, dalam mengantisipasi perubahan atau merancang strategi yang lebih efektif.

Penelitian ini bertujuan untuk mengembangkan dan membandingkan kinerja dua metode klasifikasi umum dalam analisis sentimen, yaitu Metode Naive Bayes dan Random Forest. Metode Naive Bayes adalah metode statistik yang menggunakan teorema Bayes dan mengasumsikan independensi kondisional

antara fitur-fitur dalam data. Sementara itu, Random Forest adalah metode klasifikasi *ensemble* yang terdiri dari kumpulan pohon keputusan acak (*decision tree*). Kedua metode ini telah terbukti efektif dalam analisis sentimen pada berbagai konteks dan dapat digunakan untuk mengklasifikasikan sentimen teks data Twitter terkait ekonomi dan industri.

Meskipun analisis sentimen terhadap data Twitter telah dilakukan dalam berbagai topik sebelumnya, penelitian tentang sentimen terkait ekonomi dan industri dengan menggunakan metode Naive Bayes dan Random Forest masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan pengetahuan ini dan memberikan wawasan berharga bagi pengambil keputusan di bidang ekonomi dan industri. Dalam konteks magang atau praktik kerja, penelitian ini diinisiasi untuk mengatasi tantangan yang dihadapi perusahaan dalam mengumpulkan dan menganalisis data Twitter terkait ekonomi dan industri.

Banyak perusahaan menghadapi kesulitan dalam menganalisis data Twitter secara efektif karena volume yang besar dan sifatnya yang tidak terstruktur. Oleh karena itu, melalui penelitian ini, kami berupaya untuk mengimplementasikan metode analisis sentimen yang dapat membantu perusahaan dalam memproses dan memanfaatkan data Twitter terkait ekonomi dan industri secara lebih efisien. Hasil penelitian ini diharapkan dapat memberikan manfaat langsung bagi perusahaan tertentu, dengan memberikan kerangka kerja yang jelas dalam menerapkan metode analisis sentimen dan memanfaatkan data Twitter dalam konteks ekonomi dan industri. Selain itu, penelitian ini juga dapat memberikan kontribusi pada pemahaman tentang tren dan pola sentimen publik terkait topik ekonomi dan industri di kalangan praktisi dan pengambil keputusan di sebuah perusahaan.

METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini berasal dari Twitter, sebuah *platform* media sosial yang populer di kalangan pengguna internet. Dalam konteks penelitian ini, kami mengumpulkan data yang terkait dengan ekonomi dan industri dengan menggunakan kata kunci "resesi". Penting untuk dicatat bahwa dalam penggunaan data dari Twitter, kami menghormati privasi pengguna dan mengikuti aturan dan kebijakan *platform* tersebut. Kami tidak menggunakan informasi pribadi pengguna dalam penelitian ini dan hanya mengambil data publik yang tersedia secara bebas. Berikut ini merupakan langkah-langkah dalam melakukan analisis sentimen, meliputi:

1. *Crawling* Data
2. *Labeling* Data
3. *Preprocessing* Data
4. Ekstraksi Fitur
5. Pemodelan dan Pelatihan
6. Evaluasi Model
7. Prediksi dan Analisis

2.1 *Crawling* Data

Crawling data adalah proses mengambil informasi dari situs web atau sumber online lainnya. Ini melibatkan navigasi otomatis melalui halaman web, mengunduh data, dan menyimpannya untuk analisis atau penggunaan lebih lanjut.

Crawler, juga dikenal *sebagai web spider* atau *web rob*ot, digunakan untuk menjelajahi situs web secara sistematis, mengikuti tautan, dan mengambil data yang relevan. Adapun [roses pengambilan data penelitian ini diperoleh secara langsung dari sosial media Twitter melalui metode crawling dengan kata kunci "resesi" menggunakan TwitterAPI pada bulan Februari sampai April 2023. Pencarian dengan memanfaatkan fitur advanced search dan didapatkan sebanyak 11.833 data.

2.2 *Labeling* Data

Data dari Twitter diamati dan diberikan label positif, negatif, atau netral. Perlu pemahaman lebih untuk menganalisa setiap tweet atau teks yang ada dalam data, dan kemudian menentukan sentimen yang tepat sesuai konteksnya. Dalam pengerjaan tahap ini ada beberapa daftar *rules* dan polaritasnya untuk mempermudah dalam pemberian label. Contohnya seperti "jika ada data yang sekilas menyatakan ajakan kebersamaan atau dukungan, tetapi disertai dengan ekspektasi yang berlawanan dan di mana si penulis pun terbukti secara logika tidak akan melakukan ajakan tersebut, kita akan padankan data tersebut dengan polaritas negatif dengan konteks menyindir". Selain itu, ada juga panduan konteks kalimat tanya, *adjectives*, *how to differ context*, dan lain-lain. Pemberian label dalam penelitian ini juga dilakukan secara manual dan membutuhkan waktu satu bulan untuk memberi label pada 11833 data tweet. Gambar 1 menunjukkan contoh data Twitter yang telah diberi label.

No	Clean Text	Polaritas
1	Apakah sebentar lagi resesi ;(((	netral
2	@xiaoyuejiu Aah so happy, Aurel nngga takut ada resesi keknya soalnya bakal cpt ketrima dimanapun	positif
3	Bagaimana cara bertahan hidup di era resesi? @ChatGPTBot	netral
4	bunga bank tapi riba hiyaa	positif
5	@sbyfess Terus blackberry terus path gak mbok anggep resesi ?	negatif
6	Kolaborasi tetap efektif untuk tumbuhkan bisnis hadapi ancaman resesi https://t.co/y9NJg9TPG7	netral
7	Beneran 2023 resesi teh (bagi aku)	negatif
8	@sbyfess Ga jirr,,	negatif
9	2023 beneran resesi nggak sih?	positif
10	@wikidpr #Kom1 #IndonesiaKetuaASEAN Dave @fraksigolkar #Jabar8: Ini mendorong terhadap resesi	netral
11	@sbyfess Basi, omongannya selalu resesi terus padahal daya beli masyarakat tetep tinggi	negatif
12	@Askrlfess Jika finansial adalah alasan kalian.lihatlah resesi sex orang Jepang	positif
13	Sayuran sekarang anjlok parah resesi beneran kalo begini	negatif
14	Resesi is real ya -rek jd id wes tutup per 15 februari https://t.co/N2VhFhmkWH	positif
15	Nama Sri Mulyani dikabarkan masuk sebagai salah satu kandidat calon Gubernur BI. detikers setuju	netral
16	Dari kemarin sosial media ramai dengan kata â??resesiâ?. Emangnya sebenernya resesi itu apa sih?	netral
17	@Bayu_Ndim_ex Tokopedia lg ngalamin resesi gk sih? Desember kmrn sahamnya anjlok kryawannya	netral
18	Indonesia diprediksi terbebas dari awan gelap ekonomi	netral
19	Pujian Menkeu Sri Mulyani untuk BRI: Saya Selalu Terkesan https://t.co/RuhZy0UaYc	netral
20	Setelah saya pelajari. Inflasi, resesi dan segala kekacauan ekonomi dunia ini di sebabkan oleh riba	netral

Gambar 1. *Labeling* Data Twitter

2.3 *Pre-Processing* Data

Preprocessing pada data umumnya digunakan untuk membersihkan data dan mengoptimalkan kualitasnya. Beberapa langkah yang umum dilakukan dalam pre-processing data adalah sebagai berikut:

- Menghapus tanda baca dan karakter khusus yang tidak relevan.
- Mengubah semua teks menjadi huruf kecil agar tidak ada perbedaan antara huruf besar dan kecil.
- Menghapus kata-kata yang tidak berguna atau *stop words* seperti "dan", "di", "yang", dan sebagainya.

- Mengatasi masalah dengan kata-kata yang salah eja atau singkatan.

2.4 Ekstraksi Fitur

Ekstraksi fitur merupakan cara mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma pembelajaran mesin. Beberapa teknik ekstraksi fitur yang umum digunakan dalam analisis sentimen adalah:

- *Bag of Words* (BoW): Menghitung frekuensi kemunculan kata-kata dalam setiap teks.
- TF-IDF (*Term Frequency-Inverse Document Frequency*): Menghitung bobot kata berdasarkan frekuensi kemunculannya dalam teks dan keseluruhan dokumen.
- *Word Embeddings*: Menggunakan model bahasa seperti *Word2Vec* atau *GloVe* untuk menghasilkan vektor representasi kata-kata.

Pada penelitian ini, ekstraksi fitur yang digunakan adalah TF-IDF Vectorizer. TF-IDF Vectorizer adalah sebuah algoritma yang digunakan dalam pemrosesan teks untuk mengubah dokumen teks menjadi representasi vektor numerik. TF-IDF digunakan untuk mengevaluasi pentingnya kata-kata dalam sebuah dokumen berdasarkan seberapa sering kata-kata tersebut muncul dalam dokumen tersebut (TF), serta seberapa umum kata-kata tersebut muncul dalam seluruh korpus dokumen (IDF).

2.5 Pemodelan dan Pelatihan

Setelah ekstraksi fitur, langkah selanjutnya adalah memodelkan data dan melatihnya menggunakan algoritma pembelajaran mesin yang sesuai. Beberapa algoritma yang umum digunakan dalam analisis sentimen adalah:

- Naive Bayes
- Support Vector Machines (SVM)
- Logistic Regression
- Random Forest
- Neural Networks (misalnya, LSTM atau CNN)

Pada penelitian ini, akan lebih fokus terhadap dua metode, yang pertama yaitu Naïve Bayes, dan Random Forest.

2.5.1 Naïve Bayes

Naive Bayes adalah algoritma pembelajaran mesin yang berdasarkan pada teorema Bayes dengan asumsi bahwa setiap fitur dalam data adalah independen satu sama lain. Dalam konteks analisis sentimen, Naive Bayes menghitung probabilitas bahwa sebuah dokumen atau teks termasuk ke dalam kategori sentimen tertentu (positif, negatif, atau netral) berdasarkan kemunculan kata-kata dalam teks tersebut. Naive Bayes mengestimasi probabilitas prior dan probabilitas kondisional dari setiap kata atau fitur dalam setiap kategori sentimen. Probabilitas prior adalah probabilitas bahwa sebuah dokumen atau teks termasuk dalam kategori sentimen tertentu tanpa memperhatikan kata-kata yang ada di dalamnya. Probabilitas kondisional adalah probabilitas bahwa sebuah kata muncul dalam teks dengan kategori sentimen tertentu. Sentimen dengan probabilitas tertinggi akan dipilih sebagai prediksi sentimen untuk teks tersebut.

2.5.2 Random Forest

Random Forest adalah sebuah metode *ensemble learning* yang menggunakan sejumlah besar pohon keputusan (*decision tree*). Setiap pohon

keputusan dibangun secara acak dengan menggunakan subset dari data pelatihan dan subset dari fitur yang dipilih secara acak. Selama pelatihan, setiap pohon keputusan akan belajar untuk memprediksi sentimen berdasarkan fitur-fitur dalam data. Setelah model Random Forest dilatih, akan digunakan set pengujian untuk memprediksi sentimen dari teks yang tidak diberi label. Setiap pohon keputusan dalam Random Forest akan memberikan prediksi sentimen, dan hasil akhir akan ditentukan berdasarkan mayoritas suara dari semua pohon keputusan. Sentimen dengan jumlah suara terbanyak akan menjadi prediksi sentimen akhir. Random Forest memiliki beberapa *hyperparameter* yang dapat disesuaikan untuk meningkatkan kinerja model. Beberapa *hyperparameter* yang umumnya diatur adalah jumlah pohon dalam Random Forest, ukuran subset data dan fitur yang digunakan dalam setiap pohon, dan kriteria pemilihan fitur terbaik dalam setiap pemisahan. Dapat juga digunakan metode validasi silang (*cross-validation*) atau teknik tuning lainnya untuk mencari kombinasi *hyperparameter* yang optimal.

HASIL DAN PEMBAHASAN

Data hasil *preprocessing* kemudian dikelompokkan berdasarkan sentimen. Tabel 1 di bawah menunjukkan sebaran persentase jumlah tweet berdasarkan sentimen.

Tabel 1. Persentase Jumlah Tweet berdasarkan Sentimen

positif	negatif	netral
28,52%	31,44%	40,04%

Berdasarkan Tabel 1, dapat dikatakan bahwa jumlah tweet dengan sentimen netral menempati urutan pertama yaitu sebesar 40,04%. Artinya, banyak masyarakat yang beropini netral mengenai ekonomi dan industri di Indonesia.

Data yang telah diberi label dibagi menjadi data training sebesar 70% dan data testing sebesar 30%. Selanjutnya akan diklasifikasikan ke dalam algoritma *Naive Bayes* dan *Random Forest*. Dari kedua algoritma tersebut nantinya akan dibandingkan yang terbaik untuk klasifikasi. Tabel 2 di bawah merupakan tabel hasil dari klasifikasi dua metode tersebut.

Tabel 2. Hasil Klasifikasi

	Naive Bayes			*Random Forest*		
	precision	*recall*	*f1-score*	*precision*	*recall*	*f1-score*
negatif	0,69	0,85	0,75	0,69	0,77	0,73
netral	0,71	0,82	0,77	0,69	0,85	0,76
po	0,8	0	0,	0,8	0	0,

sitif	2	,43	56	0	,44	57
accuracy			0,7189			0,7090

Berdasarkan Tabel 2, dapat dilihat bahwa nilai *precision* pada polaritas positif menggunakan metode *Naive Bayes* memiliki nilai lebih tinggi dibandingkan metode *Random Forest* sebesar 82%, hal ini terjadi karena nilai ketepatan data polaritas positif antara informasi yang diminta oleh pengguna Twitter dengan hasil yang diberikan oleh sistem memberikan nilai *true positive* atau *correct result* yang lebih besar daripada nilai polaritas lainnya. Nilai persentase terbesar dari *recall* sebesar 85% pada polaritas negatif dengan metode *Naive Bayes*, besarnya persentase *recall* terjadi karena tingkat keberhasilan sistem pada data polaritas negatif dalam menemukan kembali sebuah informasi yaitu hanya memiliki sedikit kesalahan saat proses klasifikasi atau *missing result*. Sedangkan nilai persentase *f1-score* terbesar berdasarkan tabel diatas yaitu 77% pada polaritas netral untuk metode Naïve Bayes. Data yang tidak seimbang antara opini negatif, netral, dan positif yang mengakibatkan nilai *accuracy* meningkat, karena data yang berlabel netral jumlahnya lebih banyak daripada data yang berlabel positif dan negatif. Hasil penelitian yang telah dilakukan dapat diketahui bahwa nilai sentimen yang paling banyak terbentuk dalam menanggapi resesi ekonomi adalah sentimen netral. Kategori sentimen netral merupakan kategori opini masyarakat yang netral dengan adanya berita tentang resesi ekonomi sehingga dari hasil yang diperoleh pemerintah dapat mengambil kebijakan yang dapat menyadarkan masyarakat terkait resesi ekonomi yang akan dijumpai negara Indonesia pada masa yang akan datang.

Hasil akurasi terbesar dari kedua metode tersebut berkisar antara 70% - 72%, akan tetapi model paling baik yang dapat memprediksi kata-kata dengan hasil terbaik adalah metode *Naïve Bayes* karena memiliki nilai recall positif yang lebih besar dibandingkan metode Random Forest.

KESIMPULAN DAN SARAN

Kesimpulannya, meskipun tingkat akurasi kedua metode, Naive Bayes dan Random Forest, hampir sama yaitu 0,7189 dan 0,7090, terdapat perbedaan dalam kecepatan pelatihan, interpretasi fitur, dan kinerja pada data tidak seimbang. Naive Bayes cenderung lebih cepat dalam pelatihan dan prediksi, sementara Random Forest dapat memberikan interpretasi fitur yang lebih kompleks. Naive Bayes juga memiliki kecenderungan yang lebih rendah terhadap overfitting, sementara Random Forest dapat memberikan kinerja yang baik pada data tidak seimbang. Pada akhirnya, pemilihan metode tergantung pada kebutuhan spesifik dan karakteristik data yang kita miliki.. Pada penelitian ini, polaritas paling tinggi yaitu polaritas netral, artinya opini masyarakat terhadap resesi ekonomi kebanyakan netral atau tidak merepresentasikan keluh kesah maupun pujian atas kinerja pemerintah. Sehingga diharapkan kepada pemerintah, khususnya dalam bidang ekonomi dan industri agar lebih mengedukasi masyarakat akan bahaya dan

apa yang harus dipersiapkan dalam menghadapi adanya resesi ekonomi di masa yang akan datang.

DAFTAR PUSTAKA

Apriani, R., & Gustian, D. (2019). Analisis Sentimen dengan Naïve Bayes terhadap Komentar Aplikasi Tokopedia. Jurnal Rekayasa Teknologi Nusa Putra, 6(1), 54-62.

Bird, S., Klein, E., & Loper, E. (2020). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional. TEKNO KOMPAK Journal, 15(1), 131-145. P-ISSN: 1412-9663, E-ISSN: 2656-3525.

Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). Pearson.

Manning, C. D., Raghavan, P., & Schütze, H. (2021). Introduction to Information Retrieval. Cambridge University Press.

Pak, A., & Parvez, M. T. (2020). Sentiment Analysis of Twitter Data Using Naive Bayes Classifier. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Shrivastava, A., & Gupta, R. (2019). Comparative study of Random Forest, Gradient Boosting and Support Vector Machine for Classification of IoT attacks. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.