Jurnal Ilmiah Wahana Pendidikan, September 2023, 9 (18), 244-250
DOI: https://doi.org/10.5281/zenodo.8312289
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

Penentuan Metode Terbaik Analisis Sentimen Pada Twitter Pelanggaran Hukum (Korupsi Dan Pajak)

Faradilla Izzah Rahmadani[1], Rizki Rahmawati[2]

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

Received: 18 Agustus 2023
Revised : 29 Agustus 2023
Accepted: 31 Agustus 2023

Abstract

A lot of topics are discussed on social media. One of the topics discussed about Twitter is violations of law in Indonesia such as corruption and taxes. Sentiment analysis is the process of extracting, understanding, and processing unstructured data to obtain sentiment information found in opinion sentences. This study was conducted to analyze public opinion towards the violation of law on the social media Twitter uses sentiment analysis. A total of 14,446 data were divided into 80% training data and 20% testing data then classified using Naive Bayes, Random Forest, and Support Vector Machine (SVM) algorithms. The calculation results show positive sentiment of 16,3%, negative sentiment of 52,3%, and neutral sentiment of 31,4%. The comparison of the three algorithms obtained using SVM algorithm gets the highest accuracy of 72%.

Keywords: *sentiment analysis, violation of law, Naive Bayes, Random Forest, Support Vector Machine*

(*) Corresponding Author: faradilla21@student.uns.ac.id

How to Cite: Rahmadani F I, & Rahmawati R. (2023). Penentuan Metode Terbaik Analisis Sentimen Pada Twitter Pelanggaran Hukum (Korupsi Dan Pajak). https://doi.org/10.5281/zenodo.8312289

LATAR BELAKANG

Perkembangan era digital membuat teknologi informasi sangat penting. Salah satu hal yang menjadi tren adalah adanya teknologi *Big Data*. *Big Data* dapat menyimpan informasi dalam bentuk teks yang jumlahnya sangat besar. Informasi yang berbentuk teks adalah sebuah informasi yang diambil dari sebuah teks seperti buku, literatur, maupun media sosial disebut juga *text mining*.

Twitter merupakan salah satu aplikasi yang memungkinkan penggunanya untuk berbagi gambar, teks, dan video, melalui media ini sesama pengguna dapat saling berinteraksi untuk mengeluarkan pendapat maupun opininya. Twitter adalah salah satu media sosial yang banyak digunakan oleh masyarakat Indonesia dan menjadi negara peringkat lima pengguna Twitter di dunia dengan jumlah 24 juta pengguna pada awal 2023 (DataIndonesia.id, 2023).

Tweet di Twitter dapat mewakili opini dari masyarakat sesuai dengan kondisi yang sedang terjadi. Opini dari masyarakat dapat berupa ulasan suatu produk, layanan, tokoh politik, maupun isu politik. Isu yang sedang ramai diperbincangkan belakangan ini dimulai dari kasus penganiayaan yang dilakukan oleh seorang remaja yang merupakan anak dari seorang pegawai pajak. Kasus ini sangat ramai diperbincangkan dengan berbagai kritik, saran, dan ujaran kebencian dari masyarakat. Isu lain mengenai KPK tidak pernah berhenti menjadi buah bibir masyarakat, apalagi ketika Ricky Ham Pagawak seorang koruptor tertangkap KPK pada 19 Februari 2023.

Cuitan mengenai "Pajak" dan "KPK" sangat banyak disampaikan oleh pengguna Twitter, baik berupa saran, kritik, pujian, maupun hinaan, maka dapat digunakan untuk mencari informasi dari polaritas positif, netral dan negatif pada opini-opini yang disampaikan. Pengambilan informasi tentu saja membutuhkan metode analisis yang tepat agar menghasilkan informasi yang dapat berguna untuk orang yang membutuhkan (Widowati & Sadikin, 2020). Metode yang tepat untuk menganalisis opini-opini di Twitter adalah analisis sentimen.

Analisis sentimen merupakan salah satu metode *text mining* untuk mengklasifikasi objek ke dalam tiga kategori sentimen positif, netral, dan negatif. Analisis sentimen adalah studi yang menganalisis pendapat, evaluasi, sikap, dan emosi orang dari bahasa tertulis (Liu, 2012). Pendekatan *machine learning* dapat digunakan untuk menganalisis data. Dalam penelitian ini analisis sentimen digunakan untuk melihat pendapat atau kecenderungan opini terhadap isu yang berkembang di masyarakat. Analisis sentimen pada penelitian ini menggunakan perbandingan klasifikasi algoritma *Naive Bayes*, *Support Vector Machine*, dan *Random Forest*.

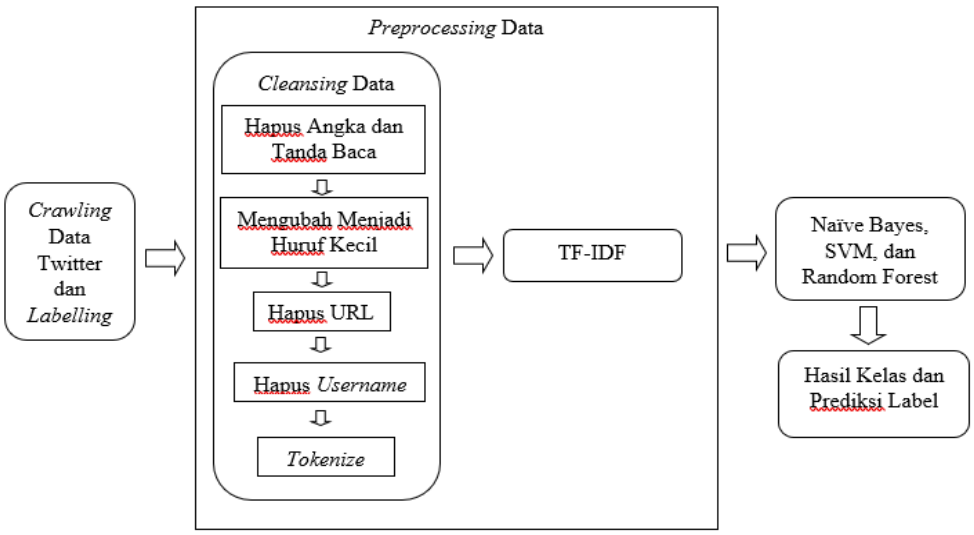
Pratiwi dkk (2021) melakukan analisis sentimen pada *review* produk kecantikan menggunakan metode *Support Vector Machine*. Ternyata metode tersebut belum sepenuhnya relevan untuk memprediksi kelas sentimen yang sesuai berdasarkan *rating* atau bintang pada *review* produk kecantikan. Hasil analisis diperoleh nilai akurasi sebesar 87% menggunakan dataset 80% data *training* dan 20% data *testing*.

Penelitian oleh Wicaksono, dkk (2022) mengklasifikasikan *review* produk Female Daily menggunakan metode *Random Forest, Support Vector Machine* (SVM), dan K-*Nearest Neighbour* (KNN) untuk mencari akurasi tertinggi dan F1-*score* Dalam penelitian tersebut didapatkan bahwa algoritma SVM mendapatkan nilai akurasi terbaik sebesar 67.10%.

Sentimen terhadap kata kunci "pajak" dan "KPK" pada media sosial Twitter akan digunakan untuk dapat menganalisis opini masyarakat melalui cuitan yang disampaikan. Kemudian diklasifikasikan ke dalam polaritas positif, netral, atau negatif menggunakan tiga metode, yaitu *Naive Bayes*, *Support Vector Machine*, dan *Random Forest*. Hasil yang diperoleh akan dibandingkan untuk mengetahui metode mana yang memberikan hasil terbaik dalam *accuracy, precision* dan *recall*.

METODOLOGI PENELITIAN

Dalam menganalisis sentimen dan untuk mengetahui nilai akurasi, dilakukan beberapa tahapan untuk memperoleh hasil terbaik. Tahapan-tahapan yang dilakukan yaitu, pengumpulan data, *labelling* data, *preprocessing*, analisis sentimen dan evaluasi model untuk mendapatkan *accuracy, precision*, dan *recall*. Gambar 1 menunjukkan tahapan penelitian analisis sentimen.



Gambar 1. *Flowchart* Penelitian Analisis Sentimen

***Crawling* Data**

Data diperoleh secara langsung dari sosial media Twitter melalui metode *crawling* dengan kata kunci "pajak" dan "KPK" menggunakan TwitterAPI pada bulan Februari sampai April 2023. Pencarian dengan memanfaatkan fitur *advanced search* dan didapatkan sebanyak 14.446 data.

***Labelling* Data**

Tahap berikutnya adalah *labelling* data, pemberian label dilakukan untuk membagi data ke dalam 3 kelas sentimen. Label yang digunakan dalam sentimen ini yaitu, "positif", "netral", dan "negatif". Kriteria yang digunakan dalam melabeli data adalah jika data mengandung kata positif, kalimat yang dituliskan berupa dukungan dan pujian kepada pajak dan KPK akan dikategorikan dalam sentimen positif. Jika data mengandung kata netral dan berupa pemberitaan fakta dan pertanyaan akan dikategorikan dalam sentimen netral. Sedangkan jika data mengandung kata negatif dan berupa olokan, sindiran, atau kekecewaan akan dikategorikan dalam sentimen negatif. Pada Tabel 1 adalah dataset yang telah diberi label.

Tabel 1. Dataset yang Telah Diberi Label

Tweet	Sentimen
@6an9_Said KPK sudah lemah syahwatnya... #KPKImpotenKorupsiNgetren #KPKImpotenKorupsiNgetren	negatif
Maju terus KPK #TegakkanHukum #KPK-RI #UndangUndang https://t.co/VEtPa5KQ9T	positif
@EllyKoro Saya yakin KPK bukan orang yang bodoh yang tidak bisa membuktikan , hanya karena alasan politik memang KPK kelihatan	negatif

bodoh .	
KPK lemah koruptor tertawa bahagia Bukan tanpa sebab kini lembaga antirasuah ini tak lagi sebagai lembaga yang independen dan terbebas dari pengaruh kekuasaan. #KPKImpotenKorupsiNgetren #KPKImpotenKorupsiNgetren	negatif
@6an9_Said KPK sudah lemah syahwatnya... #KPKImpotenKorupsiNgetren #KPKImpotenKorupsiNgetren	netral

***Pre-processing* Data**

Data yang diambil dari Twitter merupakan data yang tidak terstruktur. Oleh karena itu, perlu dilakukan *pre-processing*. Isnain, et al (2021) melakukan tahap *pre-processing* meliputi beberapa proses: menghapus url, *case folding, tokenizing, filtering,* dan *stemming*.

a. Menghapus url

Saat dilakukan pengumpulan data, ada *tweet* yang masih mengandung alamat url. Proses ini penting dilakukan agar tidak memengaruhi proses analisis sentimen.

b. *Case Folding*

Case folding merupakan proses mengganti semua huruf dalam tweet menjadi huruf kecil (*lowercase*).

c. *Tokenizing*

Proses ini menguraikan kalimat sehingga menjadi kata-kata yang terpisah.

d. *Filtering*

Kata yang tidak memiliki makna (*stopword*) dibuang, misalnya dan, di, ke, dari, dan lain-lain.

e. *Stemming*

Proses mengubah kata berimbuhan dengan menjadi kata dasarnya. Contoh: “membaca” menjadi “baca”.

TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode yang digunakan dalam pengolahan teks dan pemodelan informasi untuk mengevaluasi pentingnya sebuah kata dalam sebuah dokumen atau korpus teks. Metode ini memperhitungkan frekuensi kata di dalam dokumen dan keberadaannya di seluruh koleksi dokumen. Dengan menggunakan metode TF-IDF, kata-kata yang memiliki nilai TF-IDF tinggi cenderung lebih penting dalam konteks dokumen atau korpus yang sedang dianalisis.

Klasifikasi

Data yang telah diberi label kemudian dibagi menjadi data training 80% dan data testing 20%. Setelah itu dilakukan klasifikasi menggunakan algoritma *naive bayes, random forest*, dan *support vector machine* (SVM). Algoritma *naive bayes* adalah algoritma klasifikasi yang berdasarkan teorema Bayes dengan asumsi sederhana (naif) bahwa semua fitur dalam dataset adalah independen satu sama lain. *Random forest* merupakan salah satu metode klasifikasi dengan mengembangkan metode pohon keputusan. Dalam random forest, sekumpulan pohon keputusan dibangun dan hasil prediksi dari setiap pohon digabungkan untuk menghasilkan

prediksi akhir. Metode SVM bertujuan untuk menentukan *hyperplane* paling optimum (Suyanto, 2018). *Hyperplane* adalah suatu subspace yang memiliki dimensi satu kurang dari dimensi ruang fitur. Dalam kasus paling sederhana, yaitu pemisahan dua kelas dalam ruang dua dimensi, hyperplane merupakan garis lurus yang membagi ruang menjadi dua bagian. Hasil perhitungan ketiga metode tersebut akan dilihat mana yang paling baik untuk klasifikasi.

HASIL DAN PEMBAHASAN

Data hasil preprocessing kemudian dikelompokkan berdasarkan sentimen. Tabel 2 di bawah menunjukkan sebaran persentase jumlah tweet berdasarkan sentimen.

Tabel 2. Persentase Jumlah Tweet berdasarkan Sentimen

positif	negatif	netral
16,3%	52,3%	31,4%

Berdasarkan Tabel 2, dapat dikatakan bahwa jumlah tweet dengan sentimen negatif menempati urutan pertama yaitu sebesar 52,3%. Artinya, banyak masyarakat beropini negatif mengenai KPK dan pajak.

Data yang telah diberi label dibagi menjadi data training sebesar 80% dan data testing sebesar 20%. Selanjutnya akan diklasifikasikan ke dalam algoritma *Naive Bayes, Support Vector Machine*, dan *Random Forest*. Dari ketiga algoritma tersebut nantinya akan dibandingkan yang terbaik untuk klasifikasi. Tabel 3 merupakan tabel hasil klasifikasi.

Tabel 3. Hasil Klasifikasi

	Naive Bayes			*Support Vectore Machine*			*Random Forest*		
	precision	*recall*	*f1-score*	*Precision*	*recall*	*f1-score*	*precision*	*recall*	*f1-score*
negatif	0,66	0,98	0,79	0,74	0,85	0,79	0,68	0,94	0,79
netral	0,87	0,49	0,63	0,70	0,63	0,67	0,81	0,54	0,65
positif	0,92	0,21	0,35	0,68	0,49	0,57	0,84	0,34	0,48
accuracy			0,71			0,72			0,72

Berdasarkan Tabel 3, dapat dilihat bahwa nilai *precision* pada polaritas positif menggunakan metode *Naive Bayes* memiliki nilai lebih tinggi dibandingkan metode *Random Forest* dan SVM sebesar 92%, 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 98% 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 79% pada polaritas netral untuk ketiga metode klasifikasi. Data yang tidak seimbang antara opini negatif, netral, dan positif yang mengakibatkan nilai accuracy meningkat, karena data yang diluar negatif jumlahnya lebih banyak daripada data yang diluar positif dan netral. Hasil penelitian yang telah dilakukan dapat diketahui bahwa nilai sentimen yang paling banyak terbentuk dalam menanggapi pajak dan KPK adalah sentimen negatif. Kategori sentimen negatif merupakan kategori opini masyarakat yang negatif dengan adanya kasus-kasus mengenai pajak dan KPK, sehingga dari hasil yang diperoleh pemerintah dapat mengambil kebijakan yang dapat mengurangi terjadinya kasus korupsi dan penyelewengan pajak.

Hasil akurasi terbesar dari ketiga metode tersebut sebesar 72% pada metode SVM dan *Random Forest*, akan tetapi model paling baik yang dapat memprediksi kata-kata dengan hasil terbaik adalah metode *Support Vectore Machine* karena memiliki nilai recall positif yang paling besar.

KESIMPULAN DAN SARAN

Berdasarkan hasil analisis, didapatkan kesimpulan bahwa metode *Support Vectore Machine* mendapatkan nilai akurasi terbesar yaitu 72% dan merupakan model terbaik yang dapat digunakan untuk memprediksi informasi data dengan baik. Pada penelitian ini, polaritas paling tinggi yaitu polaritas negatif, artinya opini masyarakat terhadap pajak dan KPK kurang baik. Sehingga diharapkan kepada pemerintah, pegawai pajak, dan pegawai KPK dapat memperbaiki citranya di hadapan publik.

DAFTAR PUSTAKA

DataIndonesia.id. (2023). Indonesia Masuk Negara Paling Banyak Main Twitter pada Awal 2023. Dikutip dari DataIndonesia.id pada 19 Juni 2023 https://dataindonesia.id/internet/detail/indonesia-masuk-negara-paling-banyak-main-twitter-pada-awal-2023

Isnain, A.R., Marga, N.S., Alita, D. (2021). Sentiment Analysis of Government Policy on Corona Case Using Naive Bayes Algorithm. *Indonesian Journal of Computing and Cybernetics Systems*. 15(1), 55-64.

Liu, B. (2012) Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers, Vermont, Australia.

Pratiwi, R.W., H, S.F., Dairoh, D., Af'idah, D.I., A, Q.R., & F, A.G. (2021). Analisis Sentimen Pada Review Skincare Female Daily Menggunakan

Metode Support Vector Machine (SVM). *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*. 1(1), 40-46.

Suyanto. (2018). *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung.

Wicaksono, M.H. (2022). Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada Review Female Daily. *Skripsi*. Fakultas Informatika Universitas Telkom, Bandung.

Widowati, T.T & Sadikin, M. (2020). Analisis Sentimen Twitter Terhadap Tokoh Publik dengan Algoritma Naive Bayes dan Support Vector Machine. *Jurnal SIMETRIS*. 11(2), 626-636.