



Analisis Sentimen Masyarakat Terhadap Perilaku Korupsi Pejabat Pemerintah Berdasarkan Tweet Menggunakan Naive Bayes Classifier

Abdul Syakir, Firman Noor Hasan*

Fakultas Teknologi Industri dan Informatika, Teknik Informatika, Universitas Muhammadiyah Prof. Dr. Hamka, DKI Jakarta, Indonesia

Email: ¹syakirkiki17@gmail.com, ^{2,*}firman.noorhasan@uhamka.ac.id

Email Penulis Korespondensi: firman.noorhasan@uhamka.ac.id

Abstrak—Perilaku korupsi pejabat pemerintah menjadi masalah yang meresahkan masyarakat dan mengancam integritas sistem pemerintahan. Dalam era digital saat ini, media sosial menjadi sarana penting bagi masyarakat untuk menyuarakan pendapat dan sentimen terhadap isu-isu sosial, termasuk korupsi dari pejabat pemerintah. Penelitian ini bertujuan untuk melakukan analisis sentimen masyarakat terhadap perilaku korupsi dari pejabat pemerintah berdasarkan data tweet di media sosial menggunakan metode Naive Bayes Classifier. Data tweet diambil dari media sosial twitter yang berhubungan dengan kasus-kasus korupsi yang melibatkan pejabat pemerintah dalam periode waktu tertentu. Data tersebut kemudian diproses untuk menghilangkan unsur yang tidak relevan dan mengekstraksi sentimen yang terkandung dalam tweet-tweet tersebut. Metode Naive Bayes Classifier diterapkan untuk mengklasifikasikan tweet-tweet tersebut ke dalam kategori sentimen positif, negatif, atau netral terhadap perilaku korupsi dari pejabat pemerintah. Hasil penelitian ini menyimpulkan bahwa masyarakat sangat marah, kecewa, serta tingkat kepercayaan rendah terhadap perilaku korupsi yang dilakukan oleh pejabat pemerintah. Dibuktikan dengan kategori sentimen yang paling dominan adalah sentimen negatif dengan jumlah data sebanyak 224 dan terdapat 95 data yang masuk ke dalam kategori sentimen positif.

Kata Kunci: Analisis Sentimen; Korupsi Pejabat Pemerintah; Media Sosial; Tweet; Naive Bayes Classifier

Abstract—The corrupt behavior of government officials is a problem that worries the public and threatens the integrity of the government system. In today's digital era, social media is an important means for the public to voice their opinions and sentiments on social issues, including the corruption of government officials. This study aims to analyze public sentiment toward the corrupt behavior of government officials based on tweet data on social media using the Naive Bayes Classifier method. Tweet data is taken from social media Twitter related to corruption cases involving government officials within a certain period. The data is then processed to remove irrelevant elements and extract the sentiments contained in the tweets. The Naive Bayes Classifier method is applied to classify these tweets of positive, negative, or neutral sentiment toward corrupt behavior from government officials. The results of this study conclude that the public is very angry, disappointed, and has a low level of trust in corrupt behavior committed by government officials. Proven by the most dominant sentiment category is negative sentiment with 224 data and 95 data fall into the positive sentiment category.

Keywords: Sentiment Analysis; Corruption of Government Officials; Social Media; Tweet; Naive Bayes Classifier

1. PENDAHULUAN

Pemberitaan mengenai korupsi sepertinya tak pernah berhenti mengisi tayangan televisi. Banyak dari pelaku korupsi adalah pejabat pemerintahan yang menduduki posisi penting. Korupsi merupakan tindakan yang melanggar hukum dengan maksud untuk memperkaya diri sendiri atau orang lain, sehingga menyebabkan kerugian finansial bagi negara atau perekonomian negara (berdasarkan UU No. 20/2001) [1]. Korupsi telah merugikan masyarakat dengan menghambat pembangunan, merusak kepercayaan publik terhadap lembaga pemerintahan, serta menciptakan ketidakadilan dalam distribusi sumber daya [2]. Salah satu bentuk korupsi yang menjadi perhatian serius adalah perilaku korupsi yang dilakukan oleh pejabat pemerintah.

Perilaku korupsi dari pejabat pemerintah memiliki dampak yang buruk dan merugikan [3]. Dalam ranah sektor publik, [4] korupsi dapat menyebabkan pemborosan sumber daya dan mengakibatkan penurunan kualitas pelayanan publik [5]. Selain itu, korupsi juga memiliki dampak di luar aspek keuangan, seperti merusak kepercayaan pada sistem akuntabilitas [6] pemerintah, merendahkan citra pemerintahan di mata publik, dan mengganggu kualitas pelayanan publik secara keseluruhan [5]. Hal ini berakibat negatif terhadap kepercayaan masyarakat terhadap pemerintahan dan mempengaruhi tingkat partisipasi publik dalam proses politik.

Penggunaan media sosial telah menyebar luas dan sangat pesat dalam berbagai macam masyarakat [7]. Media sosial tidak hanya untuk bersosialisasi, dan komunikasi namun juga untuk menyampaikan aspirasi dan mengutarakan pendapat masyarakat secara mudah dan luas [8], salah satu media sosial yang digunakan untuk mempersentasikan pemikiran masyarakat melalui twitter [9] dalam menanggapi pejabat pemerintah yang korupsi. Hal ini dapat dijadikan sebagai sebuah acuan untuk mengetahui sentimen [10] masyarakat terhadap perilaku korupsi pejabat pemerintah.

Data mining merupakan suatu proses teknik dalam pengolahan data yang bertujuan untuk menemukan hubungan atau pola dari data yang tidak diketahui oleh pengguna, kemudian menyajikannya dalam bentuk yang mudah dipahami [11]. Dari hubungan data yang ditemukan, informasi tersebut dapat digunakan untuk pengambilan keputusan [12]. Data mining terdiri dari beberapa kelompok yaitu: Deskripsi, Estimasi, Prediksi, Klasifikasi, Clustering, dan Asosiasi [11]. Klasifikasi adalah proses pengelompokan objek data ke dalam kelas-kelas tertentu berdasarkan karakteristiknya [13]. Salah satu metode yang sering digunakan dalam klasifikasi adalah



metode naive Bayes [14]. Analisis sentimen adalah suatu teknik untuk mengumpulkan komentar dan ulasan dari orang lain mengenai suatu hal, [15] misalnya informasi yang terdapat pada media sosial berbasis web, dan kemudian mengklasifikasikan tingkat emosi yang diungkapkan oleh orang tersebut, apakah bersifat netral, positif, atau negatif [16]. Beberapa metode Analisis Sentimen yang umum digunakan meliputi K-Nearest Neighbor (K-NN), Naive Bayes Classifier (NBC), Random Forest, dan Decision Tree [17].

Analisis Sentimen berdasarkan tweet di media sosial sudah dilakukan oleh beberapa peneliti diantaranya yaitu I. R. Afandi, F. Noor Hasan, dkk [18] telah melakukan penelitian analisis sentimen berdasarkan tweet di media sosial terkait pelayanan jasa ekspedisi Anteraja. Penelitian ini menggunakan metode Naive Bayes dan mengumpulkan data tweet berbahasa Indonesia dari tanggal 1 Januari 2021 hingga 16 Agustus 2022. Hasil penelitian menyimpulkan bahwa metode Naive Bayes dapat memprediksi sentimen opini masyarakat terhadap pelayanan Anteraja dengan tingkat akurasi sebesar 85.06%. Tujuan dari penelitian untuk membantu dalam evaluasi, analisis, penilaian, dan pemahaman sikap masyarakat terhadap pelayanan Anteraja berdasarkan data dari media sosial. Peneliti lain yaitu F. Noor Hasan dan M. Dwijayanti [10] juga melakukan penelitian terkait analisis sentimen ulasan pelanggan terhadap layanan Grab di Indonesia. Penelitian ini menggunakan metode klasifikasi Multinomial Naive Bayes Classifier (NBC) dengan dataset berisi 1000 tweet dari Twitter. Hasil kesimpulan dari penelitian ini menunjukkan bahwa metode klasifikasi Naive Bayes Classifier (NBC) memberikan akurasi tertinggi, mencapai 92,5%. Untuk sentimen negatif, nilai presisi yang didapat adalah 57%, nilai recall sebesar 67%, dan f1-score sebesar 62%. Sedangkan untuk sentimen positif, nilai presisi mencapai 97%, nilai recall sebesar 95%, dan f1-score sebesar 96%. Hal ini menunjukkan bahwa metode NBC berhasil dalam mengidentifikasi sentimen negatif dan positif dari ulasan pelanggan tentang layanan Grab Indonesia berdasarkan data tweet.

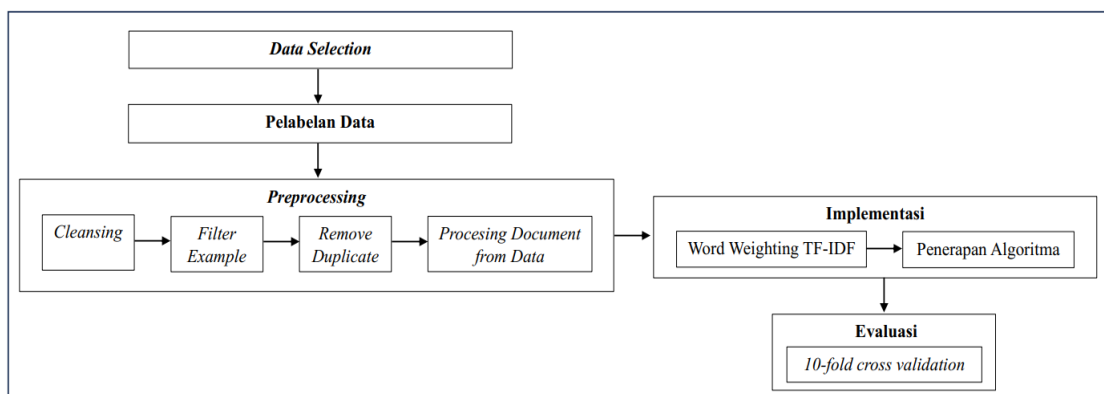
Penelitian terdahulu lainnya dilakukan oleh P. Arsi, B. A. Kusuma, dkk [15] yang meneliti sentimen terkait pemindahan ibu kota menggunakan metode Naive Bayes Classifier. bertujuan untuk mengklasifikasikan sentimen yang bersumber di media sosial Twitter. Dataset berisi 1007 tweet dari twitter. Hasil penelitian menunjukkan tingkat akurasi sebesar 94,33%, nilai presisi yang didapat adalah 0,87% dan recall sebesar 0,99%. Artinya, metode Naive Bayes Classifier berhasil dengan baik dalam mengklasifikasikan sentimen dari tweet-tweet yang berkaitan dengan pemindahan ibu kota, dengan tingkat keberhasilan yang tinggi dan nilai presisi serta recall yang baik. Penelitian lainnya dilakukan oleh Syifa Khairunnisa, Adiwijaya, dkk [19] melakukan penelitian terkait Komentar Masyarakat pada Kasus Pandemi COVID-19. Penelitian menggunakan Support Vector Machine (SVM) untuk mengklasifikasikan komentar-komentar tersebut, menghasilkan tingkat akurasi sebesar 77.77%. Kemudian penelitian lain dilakukan oleh Hakam Febtadianrano Putro, Retno Tri Vlandari, dkk [11] melakukan penelitian terkait penerapan klasifikasi pelanggan. Peneliti menggunakan metode naive bayes untuk mengklasifikasi pelanggan yang menghasilkan akurasi sebesar 92%, presisi 100%, dan recall sebesar 91%.

Dalam penelitian ini, peneliti memilih menggunakan Naive Bayes Classifier (NBC) sebagai metode pengelompokkan data berdasarkan pada teorema Bayes [20]. Alasan pemilihan metode ini adalah karena NBC memiliki tingkat akurasi yang lebih baik dibandingkan dengan metode K-Nearest Neighbor (K-NN), Random Forest, dan Decision Tree [15].

Tujuan Penelitian ini, Peneliti akan melakukan penentuan klasifikasi sentimen positif atau negative berdasarkan Tweet di Media Sosial dalam mengenai Perilaku Korupsi dari Pejabat Pemerintah. Hasil klasifikasi yang didapat akan di hitung akurasinya pada Metode Naive Bayes Classifier dengan menggunakan confusion matriks [18]. Diharapkan penelitian ini akan memberikan manfaat dalam melakukan penelitian tentang opini masyarakat berdasarkan data tweet yang berkaitan dengan Korupsi Pejabat Pemerintah dan mengidentifikasi sentimen positif atau negatif yang terkandung dalam tweet tersebut [21].

2. METODOLOGI PENELITIAN

Penelitian yang dilakukan oleh Peneliti melibatkan beberapa tahapan, yang ditunjukkan pada gambar di bawah ini

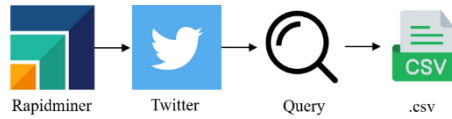


Gambar 1. Metode Penelitian



2.1 Data Selection

Tahap awal dimulai dengan proses pengumpulan data (crawling data) [22] dengan menggunakan tools RapidMiner untuk mengambil data yang bersumber dari media sosial Twitter. Seperti ditunjukkan dalam ilustrasi berikut.



Gambar 2. Proses Ekstraksi Data Tweet

Dalam penelitian ini, peneliti melakukan proses ekstraksi data dengan menggunakan kata kunci "Korupsi Pejabat Pemerintah." Data tersebut diambil dari tanggal 29 Mei 2023 hingga 7 Juli 2023, dengan menggunakan teks berbahasa Indonesia. Sebanyak 599 data tweet berhasil dikumpulkan dan disimpan dalam format file dokumen csv

Text
@Stev3n_Peg3l Selama ini bangun jalan berbanyar pake hutang, lantas uang pajak yg seharusnya di pergunakan...
@BI00dandsteel @kumparan Mending kita minta bikin search engine buat korupsi bisa bantu KPK dan setidaknya ...
@rezarezarezare @aldyfiqri @bhulukhuduktv Disini uang pajak rakyat dikorupsi, yg korupsi gatau ?;"166037...
Ada korupsi besar lagi ya ..? Halo Johny G Plate ? diduga beberapa parpol terima aliran dana. Halooo..? Parpol m...
@Relawananies Jgn jd pejabat bermental penjajah, uang hasil ngutang di korupsi kroni-kroninya, begitu negara A...
RT @CutSarina5: Dari segala sudut di korupsi dan tidak henti henti di rezim ini dari satu ke satu kasus korupsi belu...
Terus hasil pajak rakyat kemana? Hasil kekayaan alam Indonesia untuk apa? Korupsi ya? https://t.co/YKDBs20sUK...

Gambar 3. Output Ekstraksi Data Tweet

Ilustrasi di atas menampilkan beberapa contoh output proses ekstraksi data tweet yang terkait dengan sentimen masyarakat terhadap perilaku korupsi dari pejabat pemerintah.

2.2 Pelabelan Data

Setelah data didapatkan dari tahap Data Selection [23], tahap selanjutnya adalah menentukan pelabelan data secara sistem manual dengan 2 peneliti. Proses pelabelan ini dilakukan dengan memberikan sentimen positif dan negatif pada data menggunakan aplikasi Microsoft Excel 2019, seperti yang terlihat pada gambar berikut.

	Text	Sentimen
1	@Stev3n_Peg3l Sementara para pejabat kita kaya raya dengan uang pajak rak	Negatif
2	@BI00dandsteel @kumparan Mending kita minta bikin search engine buat ko	Positif
3	@rezarezarezare @aldyfiqri @bhulukhuduktv Disini uang pajak rakyat dikoru	Negatif
4	Ada korupsi besar lagi ya ..? Halo Johny G Plate ? diduga beberapa parpol teri	Negatif
5	@Relawananies Jgn jd pejabat bermental penjajah, uang hasil ngutang di koru	Positif
6	RT @CutSarina5: Dari segala sudut di korupsi dan tidak henti henti di rezim in	Negatif
7		

Gambar 4. Output Proses pelabelan data tweet

2.3 Preprocessing

Tahap selanjutnya Preprocessing merupakan langkah mengubah teks yang tidak teratur menjadi teks yang terstruktur dalam mengolah data menggunakan perangkat lunak rapidminer studio [24]. Pada tahapan ini dilakukan untuk membersihkan data yang diinginkan sebelum dilakukannya analisis lebih lanjut.

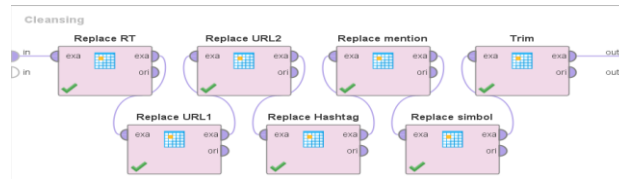
a. Cleansing yaitu tahapan untuk membersihkan karakter-karakter yang tidak dibutuhkan [24].

Pada tahapan cleansing terdapat beberapa sub-proses, yaitu :

1. Replace RT
Merupakan tahapan yang berfungsi untuk menghapus kata – kata RT dalam data komentar Twitter.
2. Replace URL 1
Merupakan tahapan yang berfungsi untuk menghapus kata URL yang terletak diawal dan ditengah tweet dalam data komentar Twitter.
3. Replace URL 2
Merupakan tahapan yang berfungsi untuk menghapus kata URL yang terletak diakhir tweet dalam data komentar Twitter.
4. Replace Hashtag
Merupakan tahapan yang berfungsi untuk menghapus simbol Hashtag pada tweet dalam data komentar Twitter.
5. Replace Mention
Merupakan tahapan yang berfungsi untuk menghapus simbol mention pada tweet dalam data komentar Twitter.



6. Replace simbol
Merupakan tahapan yang berfungsi untuk menghapus simbol – simbol pada tweet dalam data komentar Twitter.
7. Trim
Merupakan tahapan yang berfungsi untuk menghapus wide space atau spasi – spasi yang tidak perlu pada data komentar Twitter.



Gambar 5. Operator cleansing yang digunakan

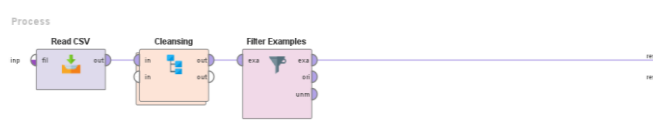
Berikut tabel proses cleansing tweet.

Tabel 1. Hasil Cleansing

Before	After	Sentimen
@Stev3n_Peg3l Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya;"1659885144667484160"	Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya1659885144667484160	Negatif
@Relawananies Jgn jd pejabat bermental penjah,uang hasil ngutang di korupsi kroni-kroninya, begitu negara APBNnya minus,tinggal naikan pajak rakyatnya,,;"1660306175722938370"	Jgn jd pejabat bermental penjahang hasil ngutang di korupsi kroni kroninya begitu negara APBNnya minus tinggal naikan pajak rakyatnya1660306175722938370	Positif
@rezarezarezare @aldyfiqri @bhulukhuduktv Disini uang pajak rakyat dikorupsi, yg korupsi gatau malu ?;"1660370110631247872"	Disini uang pajak rakyat dikorupsi yg korupsi gatau malu 1660370110631247872	Negatif

b. Filter Example

Merupakan tahapan penyaringan data yang berupa tweet komentar kosong atau tweet yang tidak memiliki karakter satupun [25]. Pada tahap ini data awal yang berjumlah 599 data komentar, setelah dilakukan proses filter data komentar berubah jumlahnya menjadi 579 data.



Gambar 6. Operator Filter Example yang digunakan

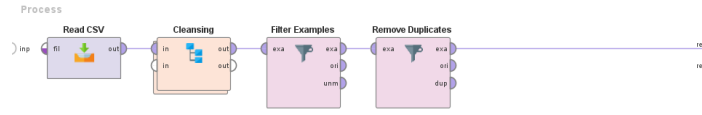
Berikut tabel proses Filter Example tweet.

Tabel 2. Hasil Filter Example

Before	After	Sentimen
?	Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya1659885144667484160	Negatif
Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya1659885144667484160	Jgn jd pejabat bermental penjahang hasil ngutang di korupsi kroni kroninya begitu negara APBNnya minus tinggal naikan pajak rakyatnya1660306175722938370	Positif
Jgn jd pejabat bermental penjahang hasil ngutang di korupsi kroni kroninya begitu negara APBNnya minus tinggal naikan pajak rakyatnya1660306175722938370	Disini uang pajak rakyat dikorupsi yg korupsi gatau malu 1660370110631247872	Negatif

c. Remove Duplicate

Merupakan tahapan untuk menghilangkan data komentar yang memiliki value yang sama [24]. Pada tahap ini data sebelumnya berjumlah 579 data, setelah dilakukan proses remove duplicate jumlah nya menjadi 319 data. Dan jumlah kategori yang paling banyak setelah proses remove duplicate yaitu kategori sentimen negatif sebesar 224 dan kategori yang paling sedikit yaitu sentimen Positif sebanyak 95.



Gambar 7. Operator Remove Duplicate yang digunakan

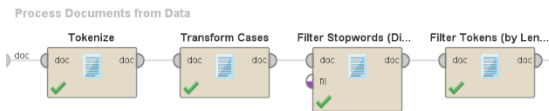
Berikut tabel proses Remove Duplicate tweet.

Tabel 3. Hasil Remove Duplicate

Before	After	Sentimen
iya enak aja uang pajak kita t alias milyar di korupsi jamaah belum yg lain	Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya1659885144667484160	Negatif
iya enak aja uang pajak kita t alias milyar di korupsi jamaah belum yg lain	Jgn jd pejabat bermental penjajah uang hasil ngutang di korupsi kroni kroninya begitu negara APBNnya minus tinggal naikan pajak rakyatnya1660306175722938370	Positif
Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya1659885144667484160	Disini uang pajak rakyat dikorupsi yg korupsi gatau malu 1660370110631247872	Negatif

d. Processing Document from Data

Setelah print data, selanjutnya melakukan Tahapan Processing Document from Data yang terdapat beberapa sub-proses [18], yaitu : Tokenize, Transform Cases, Filter Stopwords, dan Filter Tokens.



Gambar 8. Operator Processing Document from Data yang digunakan

1. Tokenize

Merupakan sub-proses yang berfungsi untuk mempermudah proses stopwords, kalimat perlu dipecah menjadi kata-kata yang terpisah [19].

Berikut tabel proses tokenize tweet.

Tabel 4. Hasil Tokenize

Before	After	Sentimen
Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya1659885144667484160	Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya	Negatif
Jgn jd pejabat bermental penjajah uang hasil ngutang di korupsi kroni kroninya begitu negara APBNnya minus tinggal naikan pajak rakyatnya1660306175722938370	Jgn jd pejabat bermental penjajahuang hasil ngutang di korupsi kronikroninya begitu negara APBNnya minustinggal naikan pajak rakyatnya	Positif
Disini uang pajak rakyat dikorupsi yg korupsi gatau malu 1660370110631247872	Disini uang pajak rakyat dikorupsi yg korupsi gatau malu	Negatif

2. Transform Cases

Merupakan sub –proses yang berfungsi untuk merubah kata-kata yang berkapital dalam dataset menjadi huruf kecil [26]. Berikut tabel proses transform case tweet.

Tabel 5. Hasil Transform Cases

Before	After	Sentimen
Sementara para pejabat kita kaya raya dengan uang pajak rakyat nya	sementara para pejabat kita kaya raya dengan uang pajak rakyat nya	Negatif
Jgn jd pejabat bermental penjajahuang hasil ngutang di korupsi kronikroninya begitu negara APBNnya minustinggal naikan pajak rakyatnya	jgn jd pejabat bermental penjajahuang hasil ngutang di korupsi kronikroninya begitu negara apbnnya minustinggal naikan pajak rakyatnya	Positif
Disini uang pajak rakyat dikorupsi yg korupsi gatau malu	disini uang pajak rakyat dikorupsi yg korupsi gatau malu	Negatif

3. Filter Stopwords



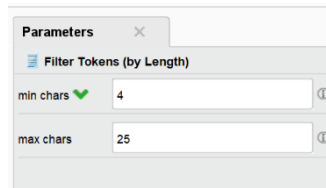
Merupakan sub-proses yang berfungsi untuk menghapus semua kata penghubung dan kata-kata yang tidak relevan dari dataset [18], peneliti menggunakan kamus stopwords yang diperoleh dari situs web www.keagle.com. Berikut tabel proses filter stopwords tweet.

Tabel 6. Hasil Filter Stopwords

Before	After	Sentimen
sementara para pejabat kita kaya raya dengan uang pajak rakyat nya	pejabat kaya raya uang pajak rakyat nya	Negatif
jgn jd pejabat bermental penjahauang hasil ngutang di korupsi kronikroninya begitu negara apbnnya minustinggal naikan pajak rakyatnya disini uang pajak rakyat dikorupsi yg korupsi gatau malu	jgn jd pejabat bermental penjahauang hasil ngutang korupsi kronikroninya negara apbnnya minustinggal naikan pajak rakyatnya uang pajak rakyat dikorupsi yg korupsi gatau malu	Positif Negatif

4. Filter Tokens

Merupakan sub-proses yang berfungsi untuk penghapusan kata-kata dengan jumlah huruf yang telah ditentukan [18]. Data yang digunakan hanya terdiri dari kalimat dengan panjang minimal 4 kata dan maksimal 25 kata.



Gambar 9. Parameters Filter Tokens yang dipergunakan

Berikut tabel proses filter tokens by length tweet.

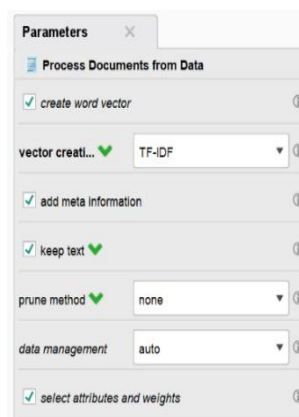
Tabel 7. Hasil Filter Tokens

Before	After	Sentimen
pejabat kaya raya uang pajak rakyat nya	pejabat kaya raya uang pajak rakyat	Negatif
jgn jd pejabat bermental penjahauang hasil ngutang korupsi kronikroninya negara apbnnya minustinggal naikan pajak rakyatnya	pejabat bermental penjahauang hasil ngutang korupsi kronikroninya negara apbnnya minustinggal naikan pajak rakyatnya	Positif
uang pajak rakyat dikorupsi yg korupsi gatau malu	uang pajak rakyat dikorupsi korupsi gatau malu	Negatif

3. HASIL DAN PEMBAHASAN.

3.1 Word Weighting TF-IDF

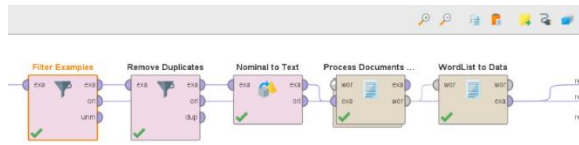
Proses ini bertujuan untuk menentukan nilai bobot pada beberapa tweet yang telah melalui proses preprocessing. Metode yang digunakan yaitu vector creation TF-IDF di operator proses dokumen from data yang telah menggabungkan langkah-langkah Tokenize, Transform Cases, Filter Stopwords, dan Filter Tokens [18]. Seperti dalam ilustrasi gambar berikut.



Gambar 10. Parameters Process Documents from Data yang digunakan



Tahap selanjutnya Operator Process Document from Data dihubungkan dengan Operator WordList to Data untuk dilakukan perhitungan bobot nilai kata dan frekuensi yang bermunculan [18] pada kumpulan data, seperti yang di ilustrasikan pada gambar berikut.



Gambar 11. Mengubungkan Operator Process Document from Data ke Operator WordList to Data

Row No.	word	in docu... ↓	total	in class (Po...	in class (Ne...
1168	pejak	240	269	61	208
779	korupsi	191	217	57	160
1199	pejabat	89	102	32	70
1353	rakyat	69	87	16	71
1618	uang	50	55	12	43
121	bayar	47	51	1	50
1092	negara	46	52	19	33
1157	orang	24	27	6	21
1218	pemerintah	19	20	0	20
547	hukum	17	18	4	14

Gambar 12. Setelah hasil Proses Word Weighting TF-IDF

Gambar di atas menunjukkan bahwa kata "Pajak" adalah kata yang paling sering muncul dalam dataset dengan frekuensi sebanyak 269 kata. Setelah itu, diikuti oleh kata "Korupsi" dengan 217 kemunculan, dan kata "Pejabat" dengan 102 kemunculan. Selain itu, peneliti telah melakukan visualisasi data dengan membuat wordcloud untuk menampilkan 40 kata dengan frekuensi kemunculan yang sering digunakan, seperti yang di ilustrasikan pada gambar berikut.



Gambar 13. Wordcloud dari 40 kata yang paling sering muncul

Dari wordcloud yang dihasilkan, dapat dilihat bahwa ukuran kata dalam wordcloud mencerminkan jumlah kemunculan kata tersebut dalam dataset. Semakin besar ukuran kata, semakin banyak kata tersebut muncul dalam dataset. Oleh karena itu, dapat disimpulkan bahwa kata-kata "pajak", "korupsi", dan "pejabat" adalah kata-kata yang paling sering muncul dalam dataset.

3.2 Penerapan Algoritma

Penelitian ini memanfaatkan algoritma Naive Bayes untuk melakukan analisis pada dataset yang telah mengalami tahap pembobotan dengan metode TF-IDF. Dataset yang digunakan terdiri dari dua kategori sentimen, yaitu sentimen positif dan sentimen negatif [23]. Untuk membagi data menjadi data test dan data Training, peneliti menggunakan perangkat lunak RapidMiner Studio yang melakukan proses pembagian secara otomatis. Berikut ini adalah contoh langkah manual dalam menghitung algoritma Naive Bayes untuk tweet yang dipilih secara acak dari kumpulan data (dataset) yang telah melalui tahap preprocessing dan pembobotan TF-IDF [18]. Tweet yang digunakan sebagai contoh adalah "Pejabat Korupsi Uang Negara," yang telah diberi label manual sebagai kategori sentimen negatif. Di bawah ini adalah tabel yang menunjukkan frekuensi kata-kata tersebut.

Tabel 8. Frekuensi Kata Pada Tweet

Kata	Class Positif	Class Negatif	Total
Pejabat	32	70	102
Korupsi	57	160	217
Uang	12	43	55
Negara	19	33	52

Keterangan :



		Pred. Positif	Pred. Negatif
Predicted Values	Pred. Positif	[True Positif] 210	[False Positif] 36
	Pred. Negatif	[False Negatif] 14	[True Negatif] 188

Gambar 18. Visualisasi Confusion matrix

Berikut cara perhitungan manual mencari nilai akurasi dengan menggunakan rumus berikut:

$$\begin{aligned}
 \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \\
 &= \frac{212+190}{212+190+34+12} \times 100\% \\
 &= \frac{402}{448} \times 100\% \\
 &= 89.73
 \end{aligned}$$

Pada gambar di atas, terlihat hasil pengujian dataset menggunakan metode 10-Fold cross validation dan algoritma naive bayes dalam aplikasi RapidMiner Studio menghasilkan akurasi nilai yang diperoleh adalah 89.73%. Dari pengolahan dataset tersebut, masyarakat sangat marah dan kecewa atas perilaku korupsi yang dilakukan oleh pejabat pemerintah. Oleh karena itu, pemerintah harus meningkatkan transparansi dalam pengelolaan dana, menegakkan hukum secara adil dan tegas terhadap pelaku korupsi untuk memberikan efek jera dan memberantas praktik korupsi.

4. KESIMPULAN

Berdasarkan analisis penelitian pada pemrosesan dataset yang dikumpulkan dari tanggal 29 Mei 2023 hingga 7 Juli 2023 berjumlah 599 data tweet, ditemukan bahwa masyarakat memiliki sentimen negatif terhadap Perilaku Korupsi Pejabat Pemerintah. Dibuktikan dengan kategori sentimen yang paling dominan adalah sentimen negatif dengan jumlah data sebanyak 224 dan terdapat 95 data yang masuk ke dalam kategori sentimen positif. Pada Proses Pembobotan TF-IDF terdapat 3 kata yang sangat mendominasi dan sering muncul pada sentimen-sentimen masyarakat yaitu kata "Pajak", "Korupsi" dan kata "Pejabat". Hal ini menunjukkan bahwa masyarakat sangat marah, kecewa, serta tingkat kepercayaan rendah terhadap perilaku korupsi yang dilakukan oleh pejabat pemerintah. Khususnya yang dilakukan oleh pejabat pajak. Selain itu, tingkat kepercayaan masyarakat terhadap pejabat pemerintah semakin berkurang. Dalam penerapan algoritma Naive Bayes dan pengujian menggunakan metode 10-fold cross validation, didapatkan hasil akurasi sebesar 89.73%. True Positif 210, True Negatif 188, False Positif 36, dan False Negatif 14.

REFERENCES

- [1] E. Maria and A. Halim, "E-Government dan Korupsi: Studi di Pemerintah Daerah, Indonesia dari Perspektif Teori Keagenan," *Jurnal Ekonomi dan Keuangan*, vol. 5, no. 1, pp. 40–58, May 2021, doi: 10.24034/j25485024.y2021.v5.i1.4789.
- [2] N. Asyikin, "Pengawasan Publik terhadap Pejabat Publik yang Melakukan Tindakan Korupsi: Perspektif Hukum Administrasi," *Jurnal Wawasan Yuridika*, vol. 4, no. 1, pp. 80–102, Mar. 2020, doi: 10.25072/jwy.v4i1.316.
- [3] J. Indrawan, A. Ilmar, and H. Simanihuruk, "Korupsi dalam Pengadaan Barang dan Jasa di Pemerintah Daerah," *Jurnal Transformative*, vol. 6, no. 2, pp. 127–147, Sep. 2020, doi: 10.21776/ub.transformative.2020.006.02.1.
- [4] R. R. Fadhilah, H. Husni, and D. Dahlan, "Pemeriksaan Kasus Tindak Pidana Korupsi terhadap Pejabat Pemerintah Kabupaten oleh Penyidik," *Kanun Jurnal Ilmu Hukum*, vol. 22, no. 1, pp. 163–178, May 2020, doi: 10.24815/kanun.v22i1.11879.
- [5] V. Mahardhika, "Pertanggungjawaban Pidana Pejabat Pembuat Komitmen Sebagai Upaya Pencegahan Korupsi Pengadaan Barang/Jasa Pemerintah," *Jurnal Hukum Samudra Keadilan*, vol. 16, no. 1, pp. 140–155, Jun. 2021, doi: 10.33059/jhsk.v16i1.2636.
- [6] R. A. Maulana, "Penyalahgunaan Kewenangandiskresi Pejabat Eksekutif Pemerintahindonesia: Studi Kasus Ratu Atut Chosiyah," *Jurnal Supremasi Hukum*, vol. 18, no. 1, pp. 84–91, Jan. 2022, doi: 10.33592/jsh.v18i01.2467.
- [7] F. Fitriana, E. Utami, and H. Al Fatta, "Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 1, pp. 19–25, Jul. 2021, doi: 10.31603/komtika.v5i1.5185.
- [8] S. Juanita, "Analisis Sentimen Persepsi Masyarakat Terhadap Pemilu 2019 Pada Media Sosial Twitter Menggunakan Naive Bayes," *Jurnal Media Informatika Budidarma*, vol. 4, no. 3, pp. 552–558, Jul. 2020, doi: 10.30865/mib.v4i3.2140.
- [9] N. M. A. J. Astari, G. H. D. Dewa, and I. Gede, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem dan Informatika (JSI)*, vol. 15, no. 1, pp. 27–29, Nov. 2020, doi: 10.30864/jsi.v15i1.332.



- [10] F. Noor Hasan and M. Dwijayanti, "Analisis Sentimen Ulasan Pelanggan Terhadap Layanan Grab Indonesia Menggunakan Multinomial Naïve Bayes Classifier," *Jurnal Linguistik Komputasional*, vol. 4, no. 2, pp. 52–58, Sep. 2021, doi: 10.26418/jlk.v6i1.
- [11] H. F. Putro, R. T. Velandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *Jurnal Teknologi Informasi dan Komunikasi (TIKOMSiN)*, vol. 8, no. 2, pp. 19–24, Oct. 2020, doi: 10.30646/tikomsin.v8i2.500.
- [12] N. L. W. S. R. Ginantra, C. P. Yanti, G. D. Prasetya, I. B. G. Sarasvananda, and I. K. A. G. Wiguna, "Analisis Sentimen Ulasan Villa di Ubud Menggunakan Metode Naive Bayes, Decision Tree, dan K-NN," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 205–215, Dec. 2022, doi: 10.23887/janapati.v11i3.49450.
- [13] G. N. Aulia and E. Patriya, "Implementasi Lexicon Based dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019," *Jurnal Ilmiah Informatika Komputer*, vol. 24, no. 2, pp. 140–153, Aug. 2019, doi: 10.35760/ik.2019.v24i2.2369.
- [14] A. Turmudi Zy, A. Nugroho, A. Rivaldi, and I. Afriantoro, "Analisis Sentimen Terhadap Pembobolan Data pada Twitter dengan Algoritma Naive Bayes," *Jurnal Teknologi Informatika dan Komputer MH. Thamrin*, vol. 8, no. 2, pp. 202–213, Sep. 2022, doi: 10.37012/jtik.v8i2.1240.
- [15] P. Arsi, B. A. Kusuma, and A. Nurhakim, "Analisis Sentimen Pindah Ibu Kota Berbasis Naive Bayes Classifier," *Jurnal Informatika Upgris*, vol. 7, no. 1, pp. 1–6, Jun. 2021, doi: 10.26877/jiu.v7i1.7636.
- [16] F. Sidik, I. Suhada, A. Haikal Anwar, and F. Noor Hasan, "Analisis Sentimen Terhadap Pembelajaran Daring dengan Algoritma Naïve Bayes Classifier," *Jurnal Linguistik Komputasional*, vol. 5, no. 1, pp. 34–43, Mar. 2022, doi: 10.26418/jlk.v6i1.
- [17] R. S. Amardita, A. Adiwijaya, and M. D. Purbolaksono, "Analisis Sentimen terhadap Ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung Menggunakan Algoritma KNN," *Jurnal Riset Komputer*, vol. 9, no. 1, p. 62, Feb. 2022, doi: 10.30865/jurikom.v9i1.3793.
- [18] I. R. Afandi, F. Noor Hasan, A. A. Rizki, N. Pratiwi, and Z. Halim, "Analisis Sentimen Opini Masyarakat Terkait Pelayanan Jasa Ekspedisi Anteraja Dengan Metode Naive Bayes," *Jurnal Linguistik Komputasional*, vol. 5, no. 2, pp. 63–70, Sep. 2022, doi: 10.26418/jlk.v6i1.
- [19] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 406, Apr. 2021, doi: 10.30865/mib.v5i2.2835.
- [20] R. W. Utami, A. Jazuli, and T. Khotimah, "Analisis Sentimen Terhadap Xiaomi Indonesia Menggunakan Metode Naïve Bayes," *Indonesian Journal of Technology, Informatics and Science (IJTIS)*, vol. 3, no. 1, pp. 21–30, Dec. 2021, doi: 10.24176/ijtis.v3i1.7514.
- [21] F. Setya Ananto and F. N. Hasan, "Implementasi Algoritma Naïve Bayes Terhadap Analisis Sentimen Ulasan Aplikasi MyPertamina pada Google Play Store," *Jurnal ICT : Information Communication & Technology*, vol. 23, no. 1, pp. 75–80, 2023, [Online]. Available: <https://ejournal.ikmi.ac.id/index.php/jict-ikmi>
- [22] A. Tirtana, A. Zulkarnain, Y. D. Listio, and S. Malang, "Pembuatan Sistem Pencarian Pekerjaan Menggunakan TF-IDF," *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 13, no. 2, 2019, doi: 10.47065/josh.v3i4.1944.
- [23] T. A. Sari, E. Sinduningrum, and F. Noor Hasan, "Analisis Sentimen Ulasan Pelanggan Pada Aplikasi Fore Coffee Menggunakan Metode Naïve Bayes," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 6, pp. 773–779, Jun. 2023, doi: 10.30865/klik.v3i6.884.
- [24] A. Saputra and F. Noor Hasan, "Analisis Sentimen Terhadap Aplikasi Coffee Meets Bagel Dengan Algoritma Naïve Bayes Classifier," *Sibatik Journal: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan*, vol. 2, no. 2, pp. 465–474, Jan. 2023, doi: 10.54443/sibatik.v2i2.579.
- [25] Nurdin, M. Suhendri, Y. Afrilia, and Rizal, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (Nbc)," *Sistem Informasi (SISTEMASI)*, vol. 10, no. 2, pp. 268–279, May 2021, doi: 10.32520/stmsi.v10i2.1193.
- [26] M. Saraswati and D. Rimirasih, "Analisis Sentimen Terhadap Pelayanan Krl Commuterline Berdasarkan Data Twitter Menggunakan Algoritma Bernoulli Naive Bayes," *Jurnal Ilmiah Informatika Komputer*, vol. 25, no. 3, pp. 225–238, Dec. 2020, doi: 10.35760/ik.2020.v25i3.3256.