Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

Analisis Sentimen Twitter Atas Isu Hak Angket Menggunakan Pembobotan TF-IDF dan Algoritma SVM

Irqi Anbi Fahrezi[1], Rudiman[2], Nauval Azmi Verdikha[3]
[1,2,3] Universitas Muhammadiyah Kalimantan Timur
2011102441036@umkt.ac.id[1], rudiman@umkt.ac.id[2], nav651@umkt.ac.id[3]

ABSTRACT

Social media has become an important platform for voicing public opinion. One of the most popular and frequently used social media is Twitter. Twitter is a popular social media in Indonesia for discussions on political issues. The topic that is being discussed is the "inquiry right" because of the alleged fraud that occurred in the 2024 elections. The alleged fraud in the 2024 elections raised issues related to the rolling of the right of inquiry aimed at finding out the oddity or fraud. Therefore, a method is needed to classify the opinion whether it is classified as a positive or negative sentiment. This research uses 1113 data obtained from Twitter social media by applying crawling techniques. The data goes through several preprocessing stages then feature extraction using Term Frequency-Inverse Document Frequency, split data, and Support Vector Machine algorithms. The test results using these stages obtained an accuracy of 75%, indicating that the applied method is effective in classifying public sentiment related to the inquiry right issue..

Keywords : *Accuracy, Inquiry Rights, Support Vector Machine, TF-IDF, Twitter*

ABSTRAK

Media sosial menjadi platform penting dalam menyuarakan opini publik. Salah satu media sosial yang sering digunakan dan paling populer adalah Twitter. Twitter menjadi media sosial yang populer di Indonesia digunakan untuk berdiskusi termasuk isu politik. Topik yang ramai diperbincangkan adalah "hak angket" karena adanya dugaan kecurangan yang terjadi pada pemilu tahun 2024. Adanya dugaan kecurangan yang terjadi pada pemilu tahun 2024 memunculkan isu terkait bergulirnya hak angket yang ditujukan untuk mengetahui adanya keganjilan atau kecurangan tersebut. Oleh karena itu diperlukan sebuah metode untuk mengklasifikasikan opini tersebut apakah tergolong sentimen positif atau negatif. Penelitian ini menggunakan sebanyak 1113 data yang telah yang didapatkan dari media sosial Twitter dengan menerapkan teknik crawling. Data melewati beberapa tahapan preprocessing kemudian ekstraksi fitur menggunakan Term Frequency-Invers Document Frequency, split data dan algoritma Support Vector Machine. Hasil pengujian menggunakan tahapan tersebut memperoleh hasil akurasi sebesar 75%, menunjukkan bahwa metode yang diterapkan efektif dalam mengklasifikasikan sentimen publik terkait isu hak angket.

Kata kunci : *Akurasi, Hak Angket, Support Vector Machine, TF-IDF, Twitter.*

PENDAHULUAN

Melihat polemik yang terjadi di dunia politik belakangan ini adanya indikasi terjadinya kecurangan pada pemilihan umum tahun 2024 sehingga memunculkan isu terkait bergulirnya hak angket oleh Dewan Perwakilan Rakyat (DPR) untuk menuntaskan adanya dugaan keganjilan dan kecurangan (Aryanti et al., 2024). Kekacauan ini terjadi sebab adanya dugaan ketidaknetralan ASN, politik uang dan lainnya yang mana hal ini sudah semestinya dengan bergulirnya hak hak angket dapat menjadi pembenahan sistem pemilu dan pemerintahan (Supryadi, 2024). Hal ini didasari oleh salah satu hak yang dimiliki oleh DPR sebagaimana tercantum pada Undang-Undang Republik Indonesia Nomor

27 Tahun 2009 Pasal 77 ayat 3 tentang MPR, DPD, DPD, dan DPRD. Hal ini pun akhirnya mengundang berbagai macam respon yang diberikan oleh masyarakat.

Salah satu media dimana masyarakat dapat memberikan komentarnya terhadap suatu isu adalah Twitter. Twitter adalah media sosial yang digunakan untuk mengirimkan pesan singkat atau tweet (Krisdiyanto, 2021). Indonesia dengan jumlah pengguna Twitter yang berkisar 19,5 juta pengguna dengan menduduki peringkat ke-5 media sosial yang sering digunakan pada tahun 2020 (Amelia et al., 2022). Dengan banyaknya pengguna serta kebebasan yang diberikan dalam platform media sosial ini maka penting untuk melakukan analisis sentimen terhadap opini-opini atau pandangan masyarakat terhadap isu-isu tertentu yang di ungkapkan di media sosial Twitter.

Analisis sentimen merupakan sebuah metode dalam melakukan ekstraksi data opini, mengolah dan memahami data yang memiliki basis tekstual secara otomatis untuk melihat sentimen yang tercantum pada sebuah pendapat (Hendra & Fitriyani, 2021). metode analisis ini menerapkan teknik *Natural Language Preprocessing* (NLP) untuk mengidentifikasi kata dan frasa yang menunjukkan emosi tertentu (Munawaroh et al., 2024). Untuk mengatasi pencarian dokumen yang relevan dan mengurangi kesalahan pengambilan, digunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk pencarian keterkaitan antar beberapa dokumen. merupakan jenis pembobotan yang kerap kali digunakan dalam *information retrival* (Yutika et al., 2021). TF-IDF merupakan sebuah metode menghitung atau membobotkan kata melalui teknik *tokenisasi, stopwords, stemming* dan frekuensi terhadap munculnya kata dalam sebuah dokumen yang menunjukan tingkat kepentingan sebuah kata dalam sebuah dokumen (Rofiqi et al., 2019). Untuk melakukan analisi sentimen menggunakan teknik *Text Mining* diperlukan metode klasifikasi yang tepat. Salah satu model algoritma yang kerapkali digunakan dalam klasifikasi yaitu algoritma *Support Vector Machine* (SVM).

SVM memiliki konsep utama dalam klasifikasi data, yaitu menentukan *hyperlane* optimal yang memaksimalkan jarak antara dua kelas yang sudah ada (Tineges et al., 2020). *Hyperlane* merupakan fungsi yang dapat digunakan untuk memisahkan antar kelas pada data (Pratiwi et al., 2021). Merujuk pada penelitian sebelumnya yang berjudul "Penerapan Algoritma SVM Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia" Dengan mengadopsi algoritma SVM dan TF-IDF untuk melakukan analisis sentimen dari data hasil *crawling* Twitter terhadap opini publik terhadap KPK, dari pengujian didapatkan hasil akurasi sebesar 82% serta dihasilkan sentimen label negatif 77%, positif 8%, dan netral 25% (Darwis et al., 2020). Penelitian kedua pada analisis sentimen layanan indihome menggunakan SVM didapatkan hasil akurasi 87% (Tineges et al., 2020).

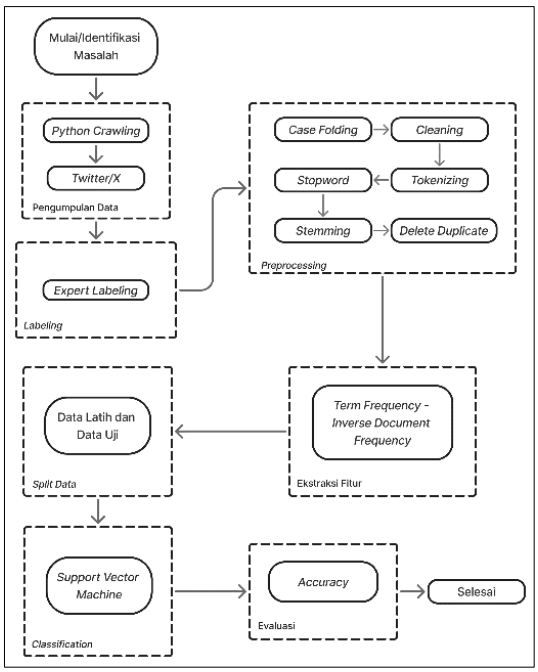
Berdasarkan penjelasan pada paragraf sebelumnya, penelitian ini memiliki urgensi penelitian yang terletak pada pentingnya menganalisis sentimen masyarakat yang ada di media sosial terkait isu hak angket dalam konteks pemilihan umum di Indonesia tahun 2024. Proses analisis sentimen dalam penelitian ini akan menerapkan pembobotan atau ekstraksi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF) dan model klasifikasi *Support Vector Machine* (SVM). Hasil akhir yang didapatkan dari penerapan metode tersebut adalah akurasi yang digunakan melakukan evaluasi. Dengan adanya penelitian ini

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

juga diharapkan dapat berguna bagi lembaga legislatif dalam memahami sentimen terhadap penggunaan hak angket yang dimiliki dalam menjalankan tugas dan kewajiban.

METODE PENELITIAN

Terdapat beberapa tahapan penelitian yang dilakukan yanng digambarkan pada gambar 1. Penelitian diawali dengan pengumpulan data, *labeling, preprocessing*, ektraksi fitur, *split* data, *classification*, dan evaluasi. Keseluruhan tahapan dilakukan memanfaatkan bahasa pemrograman Python dan fungsi serta *library* yang dimilikinya seperti *library pandas, library NLTK, library sastrawi, library matplotlib, library sckit-learn*. Untuk menuliskan program digunakan *google colaboratory* sebagai *text editor*.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Metode yang digunakan untuk mengumpulkan data adalah metode *crawling* pada platform *Twitter* dengan menerapkan kata kunci "hak angket" dalam konteks pemilihan umum Indonesia tahun 2024. *Crawling* dilakukan dengan memanfaatkan *tools tweet-harvest* yang dikembangkan dengan Node.js. *Tweet harvest* adalah *tools* untuk pengambilan data di *Twitter* dengan memanfaatkan *auth_token* Twitter (Yuniarossy et al., 2024).

2.2 Labeling

Data yang telah diekstrak kemudian masuk kedalam tahap labeling. Tujuannya adalah memberikan label pada dataset berdasarkan keadaan emosi dan bahasa pengguna yang ada pada setiap tweet (Aldisa & Maulana, 2022). Pemberian label sentimen dilakukan dengan memberikan label positif ataupun negatif pada komentar berdasarkan acuan dasar kebahasaan. Oleh sebab itu pada penelitian ini labeling dilakukan oleh *expert* dalam di dalam bidangnya dengan memanfaatkan *website* projects.co.id, tempat transaksi antara pengguna jasa dan tenaga ahli.

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

2.3 *Preprocesssing*

Tahap *preprocessing* bertujuan untuk mengubah bentuk dokumen yang memiliki data tidak terstruktur menjadi terstruktur agar dapat diolah lebih lanjut (Ridwansyah, 2022). Dalam arti lain *preprocessing* berfungsi untuk mengurangi volume kosa kata dengan menghapus noise dan menyeragamkan bentuk kata pada data (Normawati & Prayogi, 2021). Tahap ini juga menggunakan *library sastrawi* untuk melakukan pemrosesan teks berbasis bahasa Indonesia. Adapun proses *preprocessing* yang diterapkan adalah *case folding*, *cleaning*, *tokenizing*, *stopword*, *stemming*, dan *delete duplicate* (Supriyadi & Sibaroni, 2023); (Fitriyah & Kartikasari, 2023).

2. 3.1 *Case Folding*

Data berbasis teks seringkali terjadi inkonsistensi dalam penggunaan huruf kapital. Dengan menerapkan fugsi 'lower()' yang dimiliki python seluruh kata didalam dokumen akan diubah menjadi huruf kecil (*lowercase*).

2. 3.2 *Cleaning*

Cleaning merupakan fungsi guna menghilangkan noise dari kata-kata yang tidak dibutuhkan dalam proses klasifikasi. Kata-kata yang merupakan karakter pada teks yang tidak penting dan tidak mempengaruhi sentimen seperti *hashtag, url, mention*, simbol ataupun karakter non-alfanumerik.

2. 3.3 *Tokenizing*

Tokenizing adalah proses pemisahan teks panjang dapat berupa kalimat, paragraf ataupun dokumen menjadi potongan-potongan yang disebut juga dengan token (kata). Contohnya pada kalimat "python memiliki banyak library" setelah ditokenisasi menjadi "python, memiliki, banyak, library".

2.3.4. *Stopword*

Stopword atau *stopword removal* adalah proses mengambil kata kata penting dalam dataset. Kata-kata yang bersifat umum, konjungsi, dan tidak memiliki makna akan dihapus, untuk menjalankan tahap ini digunakan *library sastrawi* untuk pemrosesan data teks berbasis bahasa indonesia.

2. 3.5 *Stemming*

Stemming adalah tahapan yang bertujuan untuk menghapus awalan, akhiran, ataupun gabungan dari keduanya dan merubahnya menjadi kata dasar atau root word. Tahap ini memanfaatkan 'StemmerFactory' dari *library sastrawi*.

2. 3.6 *Delete Duplicate*

Penerapan *delete duplicate* dapat mencegah adanya data berulang yang memiliki makna yang sama. Dengan menjaga kebersihan data dapat menghemat sumber daya dan menghemat waktu pemrosesan.

2.4 Ekstraksi Fitur

Proses selanjutnya, dilakukan pembobotan kata menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF) adalah proses mentransformasi data yang memiliki basis tekstual menjadi data numerik untuk memeberi bobot pada setiap kata atau fitur dengan mengabungkan perhitungan frekuensi kemunculan sebuah kata dan *inverse* frekuensi dokumen yang mengandung kata tersebut (Mahendra., 2019; Septian et al., 2019).

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

Tujuan TF-IDF adalah melakukan identifikasi kata penting yang ada didalam dokumen atau kata kunci dalam sebuah dokumen atau kumpulan dokumen (Wati et al., 2023).

Nilai *Term Frequency* (TF) merupakan teknik dalam mencari bobot dari dokumen dengan mencari banyaknya kemunculan *term* pada dokumen. Semakin sering sebuah *term* muncul maka akan mempengaruhi nilai pembobotan (Mahendra., 2019). Untuk menghitung *term frequency* suatu dokumen dapat dihitung dengan persamaan (1) sebagai berikut (Ananda & Suryono, 2024):

$$tf_{t,d} = \frac{Jumlah\ kemunculan\ term\ t\ dalam\ dokumen\ d}{Total\ jumlah\ term\ dalam\ dokumend} \quad (1)$$

TF memberikan nilai frekuensinya dari kemunculan suatu term dalam sebuah dokumen. Sedangkan *Inverse Document Frequency* adalah metode untuk menghitung penyebaran term pada keseluruhan dokumen (Mahendra., 2019). IDF dapat dihitung menggunakan persamaan (2) (Yutika et al., 2021):

$$idf_t = \log\left(\frac{N}{df_t}\right) \quad (2)$$

Dimana N adalah total jumlah dokumen dalam kumpulan dokumen tersebut. Setelah didapatkan nilai TF dan IDF maka langkah selanjutnya adalah menghitung nilai TF-IDF dengan mengalikan kedua nilai TF dan nilai IDF. Dari kedua persamaan diatas maka nilai TF-IDF dapat dihitung menggunakan persamaan (3) (Yutika et al., 2021):

$$tfidf_{t,d} = tf_{t,d} * idf_t \quad (3)$$

2.5 *Split* Data

Merupakan metode yang dapat diterapkan untuk membagi dataset dan merupakan salah satu dari banyak aspek dapat menyebabkan pengaruh dari kinerja optimal suatu model dapat bekerja pada algoritma *machine learning* (Oktafiani et al., 2023). Data latih berfungsi untuk *mentraining* algoritma, sedangkan data uji berfungsi untuk memeriksa kinerja algoritma (Putri et al., 2023). Rasio pembagian data yang diterapkan pada penelitian ini yaitu sebesar 80% data akan digunakan sebagai data latih dan 20% sebagai data uji. Rasio ini mengacu pada penelitian sebelumnya dengan rasio 80:20 digunakan dapat menghasilkan performa yang baik dalam model klasifikasi sentimen dengan akurasi sebesar 87% (Pratiwi et al., 2021). Tahap *split data* memanfaatkan *library* 'train_test_split' dari *library sklearn.*

2.6 *Classification*

SVM adalah metode klasifikasi yang memungkinkan data dipisahkan kedalam kelas yang berbeda dengan mencari *hyperplane* optimal yang memaksimalkan jarak antar kelas-kelas tersebut (Ananda & Suryono, 2024). Tujuannya adalah untuk menemukan batasan keputusan optimal yang memisahkan data sebaik mungkin. Penerapan SVM memiliki tujuan untuk menemukan *hyperlane* optimal melalui cara memaksimalkan jarak antar kelas. *Hyperplane* adalah fungsi yang dapat digunakan untuk memisahkan beberapa kelas (Pratiwi et al., 2021). SVM sangat efektif dalam memproses data yang kompleks dan cocok untuk ruang fitur yang luas. Pada dasarnya, SVM merupakan suatu *linear classifier*, namun SVM dapat dikembangkan menjadi *nonlinear classifier* (Ade Dwi Dayani et al., 2024). Berikut adalah persamaan (4) merupakan rumus kernel linear pada SVM (Rahayu et al., 2022):

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

$$k(x_i, x) = x^t x \quad (4)$$

Secara default dari penerapan *'sckit-learn'* untuk memodelkan klasifikasi menggunakan kernel *Radial Basis Function* (RBF), tetapi dengan menambahkan parameter *(kernel='linear')* pada program maka kernel linear dapat digunakan dalam proses klasifikasi. Kernel linear menghitung produk titik *(dot product)* dari dua vektor input di ruang asli tanpa mentransformasikannya ke ruang fitur yang lebih tinggi (Rabbani et al., 2023). Berikut adalah persamaan (5) merupakan persamaan *support vector machine* (Oktavia et al., 2023):

$$f(x) = \sum_{i=1}^{\infty} ai\ yi\ K(x.x^i) + b \quad (5)$$

Keterangan:
ai = alfa ke-i
yi = kelas data latih ke-i
∑ = jumlah data
K(x.x') = fungsi kernel yang digunakan, dengan
x = data uji
xi = data latih ke-i
b = bias

Hasil permodelan dapat dituangkan dalam bentuk tabel *confusion matrix*. *Confusion matrix* adalah tabel yang menunjukan Jumlah data hasil pengujian klasifikasi yang benar dan jumlah data pengujian yang salah (Normawati & Prayogi, 2021). Berikut adalah tabel 1 merupakan *confusion matrix*:

Tabel 1. Confusion Matrix

		Prediksi	
		Negative	*Positive*
Aktual	*Negative*	TN (*True Negative*)	FP (*False Positive*)
	Positive	FN (*False Negative*)	TP (*True Positive*)

Keterangan:

a. *True Positive (TP),* adalah nilai dari kelas positif yang diprediksi dengan benar.
b. *True Negative (TN),* adalah nilai dari kelas negatif yang diprediksi dengan benar.
c. *False Positive* (FP), adalah nilai dari kelas negatif yang diprediksi sebagai label positif.
d. *False Negative (FN)*, adalah nilai dari kelas positif yang diprediksi sebagai label negatif.

2.6 Evaluasi

Pada penelitian ini penilaian kinerja algoritma hanya menentukan nilai akurasi untuk mengevaluasi kinerja yang diberikan dari algoritma. Akurasi merupakan nilai rasio data *tweet* yang benar terdeteksi di dalam data pengujian. Dalam arti lain akurasi adalah ukuran yang menunjukkan seberapa dekat hasil atau nilai prediksi yang diberikan sistem dengan nilai hasil prediksi manusia (Azhari et al., 2021). berikut persamaan rumus untuk menghitung nilai akurasi:

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pengambilan data dilakukan dengan rentang waktu pencarian 14 November 2023 - 10 Mei 2024 agar data yang diambil merupakan data yang relevan dalam masa pemilu Indonesia tahun 2024. Dari proses pengambilan data tersebut didapatkan total data mentah sebanyak 1113 data, kemudian data disimpan dalam format file "hakangket.csv" yang akan digunakan pada proses selanjutnya. Hasil *crawling* terdiri dari 15 kolom namun untuk mempermudah proses analisis data pada tahapan selanjutnya, maka hanya kolom *full_text* saja yang digunakan. Berikut adalah gambar 2 menampilkan data yang berhasil dikumpulkan.

conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	tweet_url	user_id_str	username
1798819876690362827	Fri Jun 07 12:12:27 +0000 2024	0	@MichelAdam7__ Sudah wkt nya Hak Angket digul...	1799081797220499672	NaN	MichelAdam7__	in	NaN	0	0	0	https://x.com/AchmadSuchaemi2/status/179908179...	1600952077283234688	AchmadSuchaemi2
1798164334222405993	Fri Jun 07 10:39:41 +0000 2024	0	Hak Angket - Kalo mksalnya ga pusa sama penjel...	1799028450017349986	NaN	prasastisultan	in	Maastricht	0	1	0	https://x.com/prasastisultan/status/1799028480...	364291144	prasastisultan
1798642096914457062	Fri Jun 07 08:21:15 +0000 2024	1	@ika_hw71 Yang ngotot UU Tapera disahkan siapa...	1798992611645985269	NaN	ika_hw71	in	NaN	0	0	0	https://x.com/OppaHan4/status/1798992611645985269	1529006292152774656	OppaHan4
1798977857538175216	Fri Jun 07 07:18:39 +0000 2024	1	@detikcom AH MHOSOK,IBUKT1NYA HAK ANGKET MANA...	1798977857538175216	NaN	detikcom	in	NaN	0	0	0	https://x.com/ayubt6531/status/1798977857538177...	1770812338872963560	ayubt6531
1798674873335865444	Fri Jun 07 06:52:20 +0000 2024	0	@geloraco PDI-P gak usah banyak drakor gasak d...	1798971238038113818	NaN	geloraco	in	Indonesia	0	0	0	https://x.com/jokonsembrono/status/1798971238...	1432907677693267971	jokonsembrono
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1793527384899285257	Thu May 23 11:59:38 +0000 2024	0	@Miduk17 Aaahh... kalau golkar mah belum rupi ...	1793611745577296166	NaN	Miduk17	in	NaN	0	0	0	https://x.com/akuerooo/status/1793611745577296166	1751896507024658832	akuerooo
1793592415124914444	Thu May 23 11:33:34 +0000 2024	0	@msald_didu Mungkinkah kasus UKT itu settingan ...	1793606190280450382	NaN	msald_didu	in	NaN	0	0	0	https://x.com/BuanaJannah/status/1793606190280...	1610307065998200835	BuanaJannah
1793445869892817189	Thu May 23 11:09:52 +0000 2024	0	@KaikaVtuber @PngAdlinR4kyt @berkasih_m63350 @...	1793600229708841065	NaN	KaikaVtuber	in	NaN	0	0	0	https://x.com/BuanaJannah/status/1793600229708...	1610307065998200835	BuanaJannah
1793576706681111783	Thu May 23 09:36:24 +0000 2024	2	SAVE NKRI apabila Jokowi TIDAK HADIR SAAT DIPA...	1793576706681111783	NaN	NaN	in	NaN	0	0	0	https://x.com/ENS83311139/status/1793576706681...	1168102826424483329	ENS83311139
1793492088119783739	Thu May 23 09:18:34 +0000 2024	0	@PDI_Perjuangan OBOR HAK ANGKET SDH PADAM...	1793571464077267397	NaN	PDI_Perjuangan	in	NaN	0	0	0	https://x.com/ahadnet392613/status/1793571464O...	1682778973662986624	ahadnet392613

Gambar 2. Data hasil Crawling

3.2 Labeling

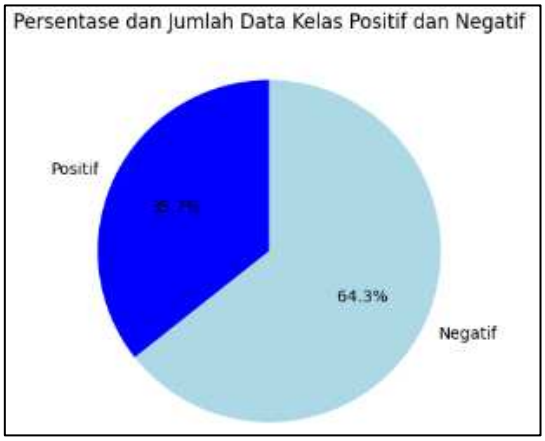
Sebanyak 1113 data diberi label sentimen positif dan negatif oleh *expert* yang hasilnya ditampilkan pada tabel 2 dibawah ini.

Tabel 2. Dataset Terlabeli

	Full_text	Label
1	@democrazymedia Njirr katanya hak angket? Manaaaaa ?? Penipu lu	Negatif
...	...	...
1113	@gorunbiraz3 @jokowi @prabowo @gibran_tweet @AgusYudhoyono @bengkeldodo @cocolatetwo @KangayamLombok Padahal tinggal lakuin hak angket	Positif

Hasil pemberian label dapat divisualisasikan dalam bentuk grafik lingkaran untuk menampilkan persentase kelas. Gambar 3 dibawah ini menunjukan bahwa dalam dataset terdapat 64,3% data tergolong kelas negatif dan 35,7% data adalah kelas positif, apabila dinyatakan dalam angka yaitu 397 merupakan kelas positif dan 716 adalah kelas negatif. Artinya banyak pengguna *Twitter* yang condong memberikan respon negatif dalam mengomentari isu hak angket yang ramai diperbincangkan.

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526



Gambar 3. Persentase Kelas Positif dan Negatif

3.3 *Preprocessing*

Data yang belum diolah dapat ditampilkan dalam bentuk *wordcloud* untuk melihat pola kata yang sering muncul. Gambar 4 adalah *wordcloud* sebelum dataset melalui proses *preprocessing*.

Gambar 4. Wordcloud Sebelum Preprocessing

Dari gambar diatas menunjukan kata-kata yang yang paling sering muncul dalam dataset yang berkaitan dengan topik Hak Angket seperti “https”, “dan”, “yg”, “PDIP”,”co”, “itu”, “Puan Maharani”, “Jokowi” dan kata-kata lainnya juga muncul dalam *wordcloud* menunjukkan bahwa didalam dataset terdapat kata-kata bersifat umum, tidak baku, bahkan *url* yang tidak memiliki pengaruh apapun. Kata-kata ini perlu dihilangkan agar tidak menjadi *noise* dan mempengaruhi hasil akhir yang didapatkan. Keberadaan kata yang tidak memiliki makna ini menunjukan bahwa pentingnya tahapan *preprocessing* untuk membuat dataset yang digunakan menjadi lebih terstruktur.

Sebanyak 1113 data mentah yang telah diberi label positif dan negatif diproses melewati tahap *case folding*, *cleaning*, *tokenizing*, *stopword*, *stemming*, dan *delete duplicate* untuk menciptakan data yang terstruktur. Hasil dari tahapan preprocessing ditampilkan dalam bentuk *wordcloud* gambar 5.

Gambar 5. Wordcloud Sebelum Preprocessing

Setelah dilakukannya *preprocessing* menyisakan 939 data bersih yang akan digunakan dalam tahapan selanjutnya. Data-data ini dapat divisualisasikan dalam bentuk

Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

wordcloud untuk melihat kata kata apa saja yang sering muncul pada dataset baik sebelum ataupun sesudah proses *preprocesssing*. Gambar *wordcloud* diatas memberikan informasi adanya perbedaan pola pada kata-kata yang sering muncul dalam dataset selain itu perubahan juga lain juga terdapat pada berubahnya setiap kata yang memiliki awalan. Namun secara keseluruhan proses *preprocessing* masih memiliki kekurangan. Kekurangan tersebut terdapat pada tahap tahap *stopword*, dimana masih terdapat kata-kata yang tidak memiliki arti belum terhapus secara menyeluruh contohnya adalah kata "yg", "gak", "ya", "nya", "sdh", "ga", dan sebagainya. Hal ini dikarenakan dari *library sastrawi* yang digunakan untuk memproses data teks berbasis bahasa Indonesia tidak sepenuhnya mencakup kata-kata istilah, bahasa gaul, atau singkatan yang digunakan oleh masyarakat dalam memberikan opininya. Permasalahan ini dapat menyebabkan pengaruh pada hasil dari pembobotan TF-IDF serta penerapannya pada algoritma SVM.

3.4 Ektraksi Fitur

Setiap kata dalam dataset diproses dan dikonversi menjadi nilai vektor numerik. Hasil dari penerapan fungsi TF-IDF menggunakan persamaan rumus 3. output yang diberikan pada baris pertama menghasilkan (0, 2636) menunjukkan bahwa kata dalam dokumen indeks ke-0 dan indeks *term* dalam *vocabulary* ke-2636 memiliki nilai TF-IDF 0,4142567525057198. Untuk output (938, 151) yaitu kata dalam dokumen indeks ke-938 dan indeks term ke-151 dalam *vocabulary* memiliki nilai TF-IDF 0,11694591337534087. Proses TF-IDF menghasilkan hasil akhir jumlah baris data sebanyak 939 baris dengan jumlah 2846 *term*. Hasil perhitungan TF-IDF ditampilkan pada gambar 6.

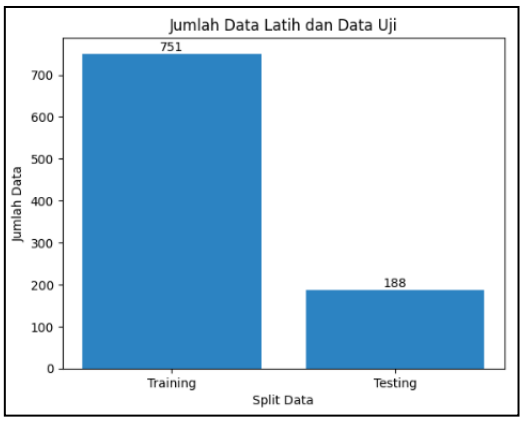
Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

```
TF-IDF Vocabulary Size: 2846 stemmed words
(Indeks dokumen, Indeks Vocabulary) Nilai TF-IDF


  (0, 2636)     0.4142567525057198
  (0, 1802)     0.5690740930139379
  (0, 1485)     0.5690740930139379
  (0, 1432)     0.4087676538015373
  (0, 854)      0.08257928762885247
  (0, 151)      0.0824036575914
  (1, 2269)     0.41203315089023
  (1, 2204)     0.43679355410911896
  (1, 2200)     0.43679355410911896
  (1, 2085)     0.43679355410911896
  (1, 2084)     0.2656016759152822
  (1, 2065)     0.29618235425927586
  (1, 877)      0.30261636990091302
  (1, 854)      0.06338383873384698
  (1, 151)      0.06324903367206515
  (2, 2734)     0.24778353561026176
  (2, 2690)     0.23373749441987526
  (2, 2526)     0.2972634901490242
  (2, 2500)     0.44754333919301686
  (2, 2414)     0.22377166959650843
  (2, 2331)     0.23373749441987526
  (2, 2161)     0.21604157531601023
  (2, 2000)     0.17798366811187039
  (2, 1618)     0.11515250600275552
  (2, 921)      0.1997598035827551
  :     :
  (935, 854)    0.06445018789121043
  (935, 772)    0.3328844538418887
  (935, 489)    0.444142027279415
  (935, 218)    0.4011017226690048
  (935, 151)    0.06431311491276544
  (935, 129)    0.4189650630627091
  (936, 1944)   0.37981042210341653
  (936, 1517)   0.53181716427325
  (936, 854)    0.11381033640741206
  (936, 659)    0.404278042561801
  (936, 151)    0.11356828402091597
  (936, 148)    0.6193732828753261
  (937, 1944)   0.23711954576897687
  (937, 1839)   0.37946860029539026
  (937, 1062)   0.48964335662390107
  (937, 854)    0.07105296143082653
  (937, 727)    0.2472454811414237
  (937, 311)    0.48964335662390107
  (937, 151)    0.07090184564095328
  (937, 95)     0.4421936717817633
  (937, 78)     0.2313774031323151
  (938, 2631)   0.5659840532013405
  (938, 1350)   0.8076205781514263
  (938, 854)    0.11719516463124831
  (938, 151)    0.11694591337534087

Jumlah Baris: 939
Jumlah terms: 2846
```

Gambar 6. Hasil TF-IDF

3.5 *Split* Data

Penelitian ini melakukan *split* data dengan porsi dari data latih dan data uji yaitu 80:20 dengan 80% data latih dan 20% data uji dengan mengambil data hasil dari pembobotan TF-IDF. Gambar 7 menampilkan hasil pembagian data dengan output dari proses *split* data, dari 939 data 751 data adalah data latih dan 188 data merupakan data uji.

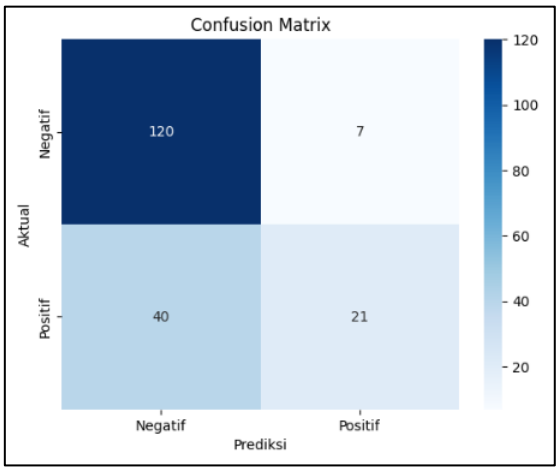


Gambar 7. Hasil Split Data

Sci-Tech Journal
Volume 3 Nomor 2 (2024) 179 – 192 E-ISSN 2830-6759
DOI: 10. 56709/stj.v3i2.526

3.6 Klasifikasi

Pada tahap klasifikasi 188 data diuji menggunakan model *Support Vector Machine* menghasilkan prediksi yang dapat disajikan dalam bentuk confusion matrix. Gambar 8 menampilkan *confusion matrix.*



Gambar 8. Confusion Matrix

Berdasarkan gambar dapat diketahui bahwa dari penerapan model algoritma didapatkan *confusion matrix* yaitu *True Positive* 21 (21 data yang diklasifikasikan kelas positif dan benar kelas positif), *True Negative* 120 (120 data yang diklasifikasikan kelas negatif dan benar kelas negatif), *False Positive* 7 (7 data yang sebenarnya kelas negatif, tetapi salah diklasaifikasikan sebagai kelas positif), *False Negative* (FN) 40 (40 data yang sebenarnya adalah kelas positif, tetapi salah diklasifikasikan sebagai kelas negatif).

3.7 Evaluasi

Berdasarkan *confusion matrix* yang dihasilkan, maka dapat dihitung nilai akurasi dari hasil permodelan dengan menggunakan persamaan rumus 6 yang dapat dijabarkan dalam bentuk perhitungan sebagai berikut.

$$Accuracy = \frac{21+120}{21+7+40+120} = \frac{141}{188} = 0{,}75 \qquad (7)$$

Dari penerapan persamaan rumus diatas dapat dijelaskan secara rinci dari perhitungan yang ada. Nilai penjumlahan 21 (TP) + 120 (TN) adalah 141 merupakan jumlah prediksi benar kemudian dibagi dengan total keseluruhan prediksi 21(TP) + 7 (FP) + 40 (FN) + 120 (TN) adalah 188. Jumlah data tersebut kemudian dibagi sehingga didapatkan nilai akhir yaitu 0,75. Menunjukkan bahwa model yang digunakan mampu melakukan prediksi yang benar sebesar 75% dari keseluruhan data yang diuji.

KESIMPULAN DAN SARAN

Berdasarkan penelitian yang telah berhasil dilakukan maka dapat disimpulkan penelitian ini menggunakan data yang diambil dari media sosial *Twitter* mengenai topik Hak Angket menggunakan teknik *crawling*. Didapatkan 1113 kemudian diberikan label sentimen yang menghasilkan 397 positif dan 716 negatif. Data kemudian diproses pada tahap *preprocessing*, tahap ini menghasilkan data bersih siap olah sebanyak 939 data. Selanjutnya data dibobotkan menggunakan TF-IDF untuk dibobotkan dengan merubah data menjadi nilai vektor numerik, yang menghasilkan 939 baris dan 2846 kata unik.

Selanjutnya dengan mengimplementasikan metode *Support Vector Machine* dengan distribusi data latih 80% dan data uji 20% didapatkan akurasi sebesar 75%. Dengan hasil *confusion matrix true positif* 21, *true negatif* 120, *false positif* 7, dan *false negatif* 40. Namun pada penelitian ini terdapat kekurangan yaitu tidak optimalnya hasil yang diberikan pada tahap *preprocessing* sebab masih adanya kata-kata yang tidak memiliki makna belum terhapus secara menyeluruh pada tahap *stopword*. Hal ini dapat menyebabkan kata-kata yang tidak penting mendapatkan bobot yang tidak seharusnya dalam model TF-IDF, sehingga mengaburkan kata-kata yang benar-benar penting. Selain itu hasil *stopword* yang tidak optimal ini juga dapat menyebabkan dimensionalitas fitur, sehingga membuat model lebih kompleks dan adanya potensi *overfitting* pada algoritma.

DAFTAR PUSTAKA

Ade Dwi Dayani, Yuhandri, & Widi Nurcahyo, G. (2024). Analisis Sentimen Terhadap Opini Publik pada Sosial Media Twitter Menggunakan Metode Support Vector Machine. *Jurnal KomtekInfo*, *11*, 1–10. https://doi.org/10.35134/komtekinfo.v11i1.439

Aldisa, R. T., & Maulana, P. (2022). Analisis Sentimen Opini Masyarakat Terhadap Vaksinasi Booster COVID-19 Dengan Perbandingan Metode Naive Bayes, Decision Tree dan SVM. *Building of Informatics, Technology and Science (BITS)*, *4*(1), 106–109. https://doi.org/10.47065/bits.v4i1.1581

Amelia, R., Darmansah, D., Prastiwi, N. S., & Purbaya, M. E. (2022). Impementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Masyarakat Indonesia Mengenai Drama Korea Pada Twitter. *JURIKOM (Jurnal Riset Komputer)*, *9*(2), 338. https://doi.org/10.30865/jurikom.v9i2.3895

Ananda, D., & Suryono, R. R. (2024). *Analisis Sentimen Publik Terhadap Pengungsi Rohingya di Indonesia dengan Metode Support Vector Machine dan Naïve Bayes*. *8*(April), 748–757. https://doi.org/10.30865/mib.v8i2.7517

Aryanti, D., Aeni, Q., Razi, F., & Qalban, A. A. (2024). Framing Pemberitaan Wacana Hak Angket Dpr Pasca Pemilu Di Media Online. *Jurnal Mutakallimin : Jurnal Ilmu Komunikasi*, *7*(1), 49–57. https://doi.org/10.31602/jm.v7i1.14537

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Darwis, D., Pratiwi, E. S., & Pasaribu, A. F. O. (2020). Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Edutic - Scientific Journal of Informatics Education*, *7*(1), 1–11. https://doi.org/10.21107/edutic.v7i1.8779

Fitriyah, Z., & Kartikasari, M. D. (2023). Text Classification of Twitter Opinion Related To Permendikbud 30/2021 Using Bidirectional Lstm. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(2), 1113–1122. https://doi.org/10.30598/barekengvol17iss2pp1113-1122

Hendra, A., & Fitriyani, F. (2021). Analisis Sentimen Review Halodoc Menggunakan Naï¨ve Bayes Classifier. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(2), 78–89. https://doi.org/10.14421/jiska.2021.6.2.78-89

Krisdiyanto, T. (2021). Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *7*(1), 32. https://doi.org/10.24014/coreit.v7i1.12945

Mahendra. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mnedeteksi Kemiripan Dokumen. *Seminar Hasil Pengabdian Masyarakat 2019*, *2*(1), 49–54.

Munawaroh, A., Ridhoi, R., & Rudiman, R. (2024). Sentiment Analysis Dengan Naïve Bayes Berbasis Orange Terhadap Resiko Pembangunan Ikn. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 587–592. https://doi.org/10.36040/jati.v8i1.8454

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI)*, *5*(2), 697–711.

Oktafiani, R., Hermawan, A., & Avianto, D. (2023). Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning. *Jurnal Sains Dan Informatika*, *9*(April), 19–28. https://doi.org/10.34128/jsi.v9i1.622

Oktavia, D., Ramadahan, Y. R., & Minarto. (2023). Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM). *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(1), 407–417. https://doi.org/10.30865/klik.v4i1.1040

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)*, *4*(1), 40–46. https://doi.org/10.20895/inista.v4i1.387

Putri, A., Hardiana, C. S., Novfuja, E., Siregar, F. T. P., Rahmaddeni, R., Fatma, Y., & Wahyuni, R. (2023). Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(1), 20–26. https://doi.org/10.57152/malcom.v3i1.610

Rabbani, S., Safitri, D., Rahmadhani, N., Sani, A. A. F., & Anam, M. K. (2023). Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 153–160. https://doi.org/10.57152/malcom.v3i2.897

Rahayu, A. S., Fauzi, A., & Rahmat, R. (2022). Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(2), 349. https://doi.org/10.30865/json.v4i2.5398

Ridwansyah, T. (2022). Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *2*(5), 178–185. https://doi.org/10.30865/klik.v2i5.362

Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied*

Informatics, *1*(2), 58–64. https://doi.org/10.28926/ilkomnika.v1i2.18

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent System and Computation*, *1*(1), 43–49. https://doi.org/10.52985/insyst.v1i1.36

Supriyadi, P. F., & Sibaroni, Y. (2023). Xiaomi Smartphone Sentiment Analysis on Twitter Social Media Using IndoBERT. *Jurnal Riset Komputer)*, *10*(1), 2407–389. https://doi.org/10.30865/jurikom.v10i1.5540

Supryadi, A. (2024). Urgensi Hak Angket Dewan Perwakilan Rakyat Republik Indonesia Guna Menyelidiki Dugaan Kecurangan Pemilu. *Ganec Swara*, *18*(1), 491. https://doi.org/10.35327/gara.v18i1.785

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM). *Jurnal Media Informatika Budidarma*, *4*(3), 650. https://doi.org/10.30865/mib.v4i3.2181

Wati, R., Ernawati, S., & Rachmi, H. (2023). Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH. *Jurnal Manajemen Informatika (JAMIKA)*, *13*(1), 84–93. https://doi.org/10.34010/jamika.v13i1.9424

Yuniarossy, B. A., Hindrayani, K. M., & Damaliana, A. T. (2024). Analisis Sentimen Terhadap Isu Feminisme Di Twitter Menggunakan Model Convolutional Neural Network (Cnn). *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *5*(1), 477–491. https://doi.org/10.46306/lb.v5i1.585

Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 422. https://doi.org/10.30865/mib.v5i2.2845