Volume 11 Nomor 01 (2024)

JURNAL REKAYASA SISTEM DAN INDUSTRI
e-ISSN: 2579-9142 p-ISSN: 2356-0843
http://jrsi.sie.telkomuniversity.ac.id

Analisis Sentimen Twitter Tentang Isu Resesi 2023: Studi Komparatif Pendekatan *Machine Learning*

Twitter Sentiment Analysis of Recession 2023: A Comparative Study of Machine Learning Approaches

Virra Retnowati A'izzah[1], Rianto[*1], Vega Purwayoga[1]

[1]Program Studi Informatika, Fakultas Teknik, Universitas Siliwangi

ARTICLE INFO

Article history:
Diterima 13-06-2023
Diperbaiki 03-03-2024
Disetujui 14-05-2024

Kata Kunci:
Bernoulli Naïve Bayes, *Decision Tree*, *K-Nearest Neighbors*, *Sentiment Analysis*, *Support Vector Machine*, Regresi *Linear*

ABSTRAK

Sentiment Analysis adalah metode yang berguna untuk memahami opini publik tentang topik tertentu. Salah satu topik yang menarik perhatian baru-baru ini adalah potensi resesi global pada tahun 2023. Dalam penelitian ini, lima algoritma yang berbeda - *Bernoulli Naive Bayes* (BNB), *Support Vector Machine* (SVM), Regresi *Linear*, *K-Nearest Neighbors* (KNN), dan *Decision Tree* - dibandingkan untuk menentukan algoritma mana yang memberikan analisis sentimen yang paling akurat terhadap data Twitter yang terkait dengan topik ini. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki akurasi tertinggi, dan mayoritas pengguna Twitter memiliki sentimen negatif terhadap topik yang berkaitan dengan potensi resesi di tahun 2023, dengan tingkat prediksi sebesar 81,7% dibandingkan dengan 16,3% untuk sentimen positif. Hasil dari penelitian ini diharapkan dapat digunakan untuk memahami sudut pandang masyarakat umum mengenai resesi yang diprediksi akan terjadi pada tahun 2023 dan untuk memberikan wawasan untuk pengembangan kebijakan dan strategi yang bertujuan untuk memitigasi penurunan ekonomi.

Keywords:
Bernoulli Naïve Bayes, Decision Tree, K-Nearest Neighbors, Linear Regression, Sentiment Analysis, Support Vector Machine

ABSTRACT

Sentiment Analysis helps understand public opinion on a particular topic. One recent topic that has attracted attention is the potential for a global recession in 2023. In this study, five different algorithms - Bernoulli Naive Bayes (BNB), Support Vector Machine (SVM), Linear Regression, K-Nearest Neighbors (KNN), and Decision Tree - were compared to determine which algorithm provided the most accurate sentiment analysis of Twitter data related to this topic. The results showed that the SVM algorithm had the highest accuracy, and most Twitter users had negative sentiments towards topics related to a potential recession in 2023, with a prediction rate of 81.7% compared to 16.3% for positive sentiments. The results of this study are expected to be used to understand the general public's viewpoints regarding the predicted recession in 2023 and to provide insights for developing policies and strategies to mitigate the economic downturn.

1. Pendahuluan

Kondisi ekonomi global saat ini menunjukkan ketidakstabilan dan ketidakpastian belakangan ini. Banyak ekonom memprediksi bahwa penurunan ekonomi akan terjadi pada tahun 2023. Potensi resesi pada tahun ini telah menjadi topik penting yang menjadi perhatian para ekonom dan ahli keuangan di seluruh dunia. Dengan adanya ketidakpastian dan ketidakstabilan global, ada kekhawatiran berkembang yang akan berdampak pada berbagai industri yang berujung pada penurunan ekonomi global di tahun-tahun mendatang. Penurunan ekonomi atau Resesi didefinisikan oleh *National Bureau of Economic Research* (NBER) sebagai penurunan yang cukup besar dalam kegiatan ekonomi yang tersebar luas dan berlangsung lebih lama dari beberapa bulan [1]. Artikel [2] mengeksplorasi kemungkinan resesi global yang akan datang dalam waktu dekat, menggambarkan berbagai faktor yang berkontribusi terhadap kekhawatiran tersebut, yaitu eskalasi inflasi, gangguan rantai pasokan, ketegangan geopolitik, dan pandemi COVID-19 yang masih berlangsung.

Email: 207006020@student.unsil.ac.id (A'izzah, V. R.), rianto@unsil.ac.id (Rianto)*, vega.purwayoga@unsil.ac.id (Purwayoga, V.)

*Penulis korespondensi

DOI: https://doi.org/10.25124/jrsi.v11i01.612

Menurut organisasi riset Ned Davis, menyatakan bahwa potensi resesi global meningkat menjadi 98% pada 2022 dan memungkinkan risiko resesi ini akan meningkat untuk beberapa waktu pada tahun 2023 [3]. Isu resesi 2023 ini telah menimbulkan perhatian dan kekhawatiran di kalangan ekonom, pemerintah, dan masyarakat umum. Pemanfaatan platform media sosial, seperti Twitter, telah muncul sebagai satu alternatif bagi masyarakat global untuk mengartikulasikan perspektif dan sentiment mereka mengenai masalah ekonomi.

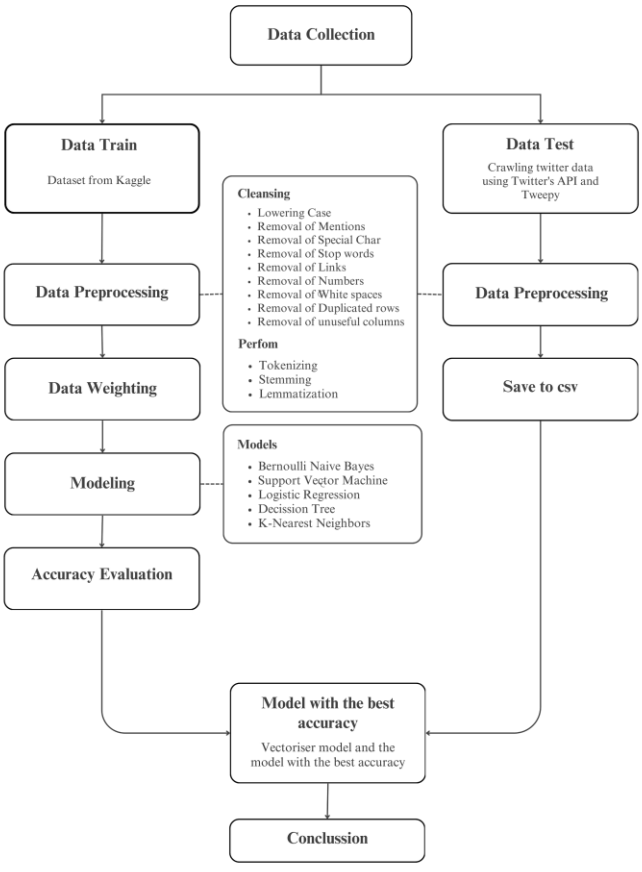
Analisis sentimen adalah proses mendeteksi dan mengklasifikasikan emosi atau opini yang diungkapkan dalam sebuah teks secara otomatis [4]. Analisis ini dapat digunakan untuk berbagai tujuan, seperti menganalisis umpan balik pelanggan, postingan media sosial, ulasan produk, ulasan film, ekonomi, isu internasional dan lainnya [4], [5]. Selain itu penggunaan analisis sentimen pada data Twitter telah menjadi mekanisme yang banyak digunakan untuk memahami perspektif masyarakat tentang beragam subjek [6], seperti pola ekonomi. Analisis sentimen yang diungkapkan dalam *tweet* dapat menghasilkan wawasan signifikan tentang persepsi publik mengenai kemungkinan resesi ekonomi pada tahun 2023.

Penelitian sebelumnya [7] melakukan analisis data Twitter untuk mengidentifikasi variabel yang menjadi faktor terhadap resesi dan reaksi masyarakat umum. Penelitian ini menggunakan teknik *Naive Bayes* dan *Support Vector Machine* (SVM) untuk memastikan sentimen masyarakat umum. Hasil penelitian menunjukkan bahwa pendekatan *Support Vector Machine* (SVM) menunjukkan tingkat akurasi yang lebih unggul, dengan tingkat 79,5%, dibandingkan dengan tingkat akurasi 72,5% yang ditunjukkan oleh metode *Naive Bayes*. Menurut prediksi metode SVM, ada 144 contoh sentimen positif dan 636 contoh sentimen negatif. Dalam penelitian [8] studi ini membahas ancaman resesi global pada tahun 2023 dan mengeksplorasi bagaimana analisis sentimen digunakan untuk mempelajari opini publik tentang masalah ini di Twitter melalui penggunaan tagar #resesi2023 dan metode *Decission Tree*. Tujuan dari penelitian ini adalah untuk mengetahui akurasi suatu metode dalam menangkap opini pengguna terkait resesi 2023. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* mencapai akurasi 89,86%, recall 82,64%, presisi 84,22%, dan f1-score 83,32%, menunjukkan kinerja yang sangat baik. Penelitian lanjutan yang disarankan adalah menambah data dari sosial media lain selain Twitter dan menerapkan beberapa algoritma lain seperti KNN, *Random Forest*, dan *Naïve Bayes*.

Penelitian lainnya [9] yang melakukan perbandingan beberapa pendekatan *machine learning* dalam melakukan sentiment analysis pada data Twitter. Studi ini mengklasifikasikan sikap positif, negatif, dan netral pada data Twitter menggunakan tiga teknik *machine learning* yang berbeda yaitu *Support Vector Machine* (SVM), *Naive Bayes* (NB), dan *Random Forest* (RF). Hasilnya menunjukkan bahwa pendekatan SVM mengungguli teknik pembelajaran mesin lainnya untuk kategorisasi sentimen pada data Twitter[9]. Selain itu, pada penelitian [10] membandingkan teknik *machine learning* untuk analisis sentimen Twitter. Algoritma yang dibandingkan adalah *Naive Bayes, Support Vector Machines* (SVM), *Maximum Entropy,* dan *Decision Tree.* Penelitian ini melibatkan percobaan dengan berbagai jenis fitur dan strategi *preprocessing*. Algoritma ini diuji pada dataset tweet yang terkait dengan Pemilihan Presiden AS 2016 oleh penulis. Dalam hal akurasi, presisi, *recall*, dan skor F1, temuan mengungkapkan bahwa SVM mengungguli metode lain. Menurut penelitian ini, SVM adalah algoritma yang layak untuk analisis sentimen data Twitter dan dapat digunakan untuk sejumlah aplikasi seperti penambangan opini, pemantauan merek, dan manajemen krisis.

Berdasarkan isu yang dipaparkan dan penelitian sebelumnya, penelitian ini bertujuan untuk melakukan analisis sentimen Twitter terhadap resesi 2023 dengan menggunakan algoritma *machine learning* yang telah diuji performanya pada data tweet resesi Q4 2022 dari Kaggle. Algoritma yang digunakan yaitu *Bernoulli Naive Bayes* (BNB), *Support Vector Machine* (SVM), Regresi Linier, *K-Nearest Neighbors* (KNN), dan *Decision Tree*, sebagai studi pembanding. Model dengan performa terbaik kemudian digunakan untuk analisis data Twitter baru pada isu resesi tahun 2023. Penelitian ini diharapkan dapat membantu mengidentifikasi tanggapan dan perspektif orang tentang isu resesi global 2023 dan mengidentifikasi algoritma *machine learning* yang mempunyai performa lebih unggul dalam analisis sentimen. Hasil analisis sentimen Twitter juga dapat membantu para pemangku kepentingan seperti pelaku bisnis, pembuat kebijakan, dan investor dalam memperoleh pengetahuan yang signifikan mengenai pandangan masyarakat terhadap ekonomi dan dapat dijadikan pertimbangan dalam membuat keputusan yang lebih baik dan merumuskan taktik yang lebih efisien untuk mengatasi hambatan yang diakibatkan oleh penurunan ekonomi.

2. Metode Penelitian



Gambar 1 Metodologi penelitian

Penelitian ini akan menganalisis *sentiment* pengguna Twitter tentang isu resesi 2023 menggunakan beberapa algoritma *machine learning* sebagai studi perbandingan. Tahapan dari penelitian ini yaitu tahap *training* dan tahap *testing*. Pada tahap *training* dilakukan *data collection* yang diambil dari Kaggle, *data preprocessing*, *data weighting*, *modeling*, *accuracy evaluation*, lalu menyimpan model vectorizer dan model dengan akurasi terbaik. Pada tahap *testing*, *data collection* diambil dengan *crawling* data *Twitter* lalu *data preprocessing* dan *sentiment analysis* menggunakan model *vectorizer* dan model dengan akurasi terbaik, dari hasil percobaan dilakukan pengambilan kesimpulan. Gambar 1 merupakan diagram metodologi penelitian.

2.1 Data Collection

Tahap pengumpulan data dilakukan dengan *crawling* data menggunakan *API Twitter* yang didapatkan dari Akun *Twitter* Developer dan memanfaatkan modul *Tweepy* pada python. Data yang dikumpulkan merupakan *tweet* berbahasa Inggris dari *Twitter* dengan *keyword* '*Recession 2023'*. Adapun untuk rentang waktu, karena parameter '*since'* pada *tweepy* telah tidak dapat digunakan dan peraturan berbayar untuk *API Twitter*[11] maka dengan keterbatasan *yang* ada, memanfaatkan akun gratis *developer Twitter* dan secara otomatis *tweet* yang terkumpul adalah *tweet* dalam rentang waktu 7 hari yaitu 20 April 2023 s/d 26 April 2023. Proses pengumpulan data ini menghasilkan data dengan 2515 *tweet* dan disimpan dalam format *.csv*. Sedangkan *data train* yang digunakan adalah data *tweet* resesi pada Q4 2022 dari Kaggle dengan jumlah data sebanyak 34850 *tweets* berbahasa Inggris dan diberi label *Positive* atau *Negative*.

2.2 Data Preprocessing

Preprocessing data sebelum analisis *sentiment* sangat penting. Proses ini dilakukan dengan melibatkan *cleaning*, *normalizing*, dan menghapus informasi yang sifatnya *noisy* dari *tweet* yang didapatkan untuk diklasifikasi [12]. *Preprocessing* data memungkinkan model analisis sentimen untuk mengenali dan mengklasifikasikan sikap, pandangan, dan perasaan di balik teks dengan akurat [13].

Pada tahap ini beberapa langkah dalam *preprocessing* yaitu mengubah semua *case* huruf ke huruf kecil, menghapus *mentions* '@', menghapus *special character* (seperti ', ;, !, dll), menghapus *stop words* ('*the'*, '*is'*, dll), menghapus *links*/tautan, menghapus angka, menghapus *white spaces*, menghapus baris yang sama (duplicated rows) dan menghapus kolom yang tidak digunakan dalam model. Selain itu diterapkan *tokenizing*, *stemming* dan *lemmatization* untuk mempermudah dalam proses analisis dan membantu meningkatkan akurasi text analisis. Hasil dari *data preprocessing* ini adalah adanya pengurangan data dari 2515 *tweet* menjadi 1167 data *tweet.*

2.3 Data Weighting

Untuk meningkatkan kinerja *sentiment analysis*, *data weighting* dilakukan sebagai strategi dengan memberikan bobot pada istilah yang ditentukan oleh signifikansi relatifnya [14]. Dalam penelitian ini metode yang digunakan adalah pendekatan TF-IDF. TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode umum untuk pembobotan data dalam *sentiment analysis*. Teknik ini melibatkan konversi dokumen menjadi vector [15]. TF-IDF melibatkan penugasan bobot untuk kata-kata individual dalam dokumen tertentu, dengan mempertimbangkan frekuensi mereka dalam dokumen serta di seluruh korpus dokumen. Teknik ini mengurangi pentingnya kata-kata yang sering digunakan dan menyoroti kata atau frasa yang signifikan.

2.4 Modeling

Proses analisis sentimen memerlukan tahap penting yang dikenal sebagai pemodelan, yang menggunakan algoritma pembelajaran mesin untuk mengkategorikan sentimen teks tertentu [13]. Berbagai algoritma digunakan dalam analisis sentimen, seperti *Naive Bayes*, *Support Vector Machine* (SVM), dan *clustering* [16]. Pemilihan algoritma dan teknik pemodelan yang tepat bergantung pada berbagai faktor, termasuk sifat data, besarnya dataset, dan tingkat presisi yang diperlukan untuk mencapai tujuan analisis sentimen [17].

Dalam kasus ini, pemodelan melibatkan pelatihan algoritma pembelajaran mesin pada kumpulan data *tweet* resesi 2023 pada Q4 2022 dari Kaggle dengan jumlah data sebanyak 34850 *tweets* yang mempunyai label *positive*, *negative*, dan *neutral*. Pelatihan ini bertujuan untuk mengembangkan model yang dapat secara akurat mengklasifikasikan *tweet* baru yang belum dikategorikan. Beberapa algoritma yang digunakan dalam penelitian adalah *Bernoulli Naive Bayes* (BNB), *Support Vector Machine* (SVM), Regresi *Linear, K-Nearest Neighbors* (KNN), dan *Decision Tree*.

2.5 Accuracy Evaluation and Comparison

Evaluasi model analisis sentimen melibatkan pemanfaatan akurasi sebagai metrik penting [18]. Penilaian presisi model analisis sentimen dilakukan melalui pemanfaatan *confusion matrix*, yang dihasilkan dari penghitungan komparatif sentimen [19]. Kinerja model dievaluasi menggunakan metrik seperti akurasi, skor AUC dan skor F1. Skor akurasi sentimen adalah metrik yang digunakan untuk menilai kinerja penilaian sentimen [20].

3. Hasil dan Pembahasan

3.1 Data Collection

Pada penelitian ini digunakan dua dataset yaitu dataset untuk *training* dan dataset untuk *testing/prediction*. *Dataset Training* adalah dataset untuk pelatihan model yang diambil dari Kaggle berisi *tweet* berbahasa Inggris tentang isu resesi 2023 pada Q4 2022. Data ini berjumlah 34581 *tweets* dengan perbandingan *sentiment* pada Tabel 1.

Tabel 1.
Label pada Dataset *Train*

Sentiment	Total	*Percentage*
Negative	28452	81,64%
Positive	6139	17,62%

Dataset *sentiment prediction* adalah dataset yang digunakan untuk analisis *sentiment* dari isu resesi 2023 yang dikumpulkan dengan *crawling* data menggunakan *API Twitter* memanfaatkan modul *Tweepy* pada python. Data yang dikumpulkan merupakan *tweet* berbahasa Inggris dari *Twitter* dengan *keyword* '*Recession 2023'* dalam rentang waktu 20

April 2023 s/d 26 April 2023. Proses pengumpulan data ini menghasilkan data dengan 2515 *tweet* dan disimpan dalam format *.csv.*

3.2 Data Preprocessing

Data *preprocessing* dilakukan pada kedua dataset (dataset *training* dan dataset *prediction*).

3.2.1 Cleansing

a. *Lowering case*

Lowering case atau mengubah semua jenis huruf pada dokumen menjadi huruf kecil bertujuan untuk menstandarisasi *text* dan untuk mempermudah dalam proses analisis. Tabel 2 merupakan perbandingan dari hasil proses *lowering case*.

Tabel 2.
Hasil Proses *Lowering Case*

	Tweet awal	*Tweet* hasil
Data Train	#OOTT #WTI	#oott #wti
	Implication 1 could be due to the threat of #recession overriding the worry of #inflation, as the #CPI data was benign this morning. The fed's likely pivot Wednesday shall boost this implication. https://t.co/9fWRLaYPJU	implication 1 could be due to the threat of #recession overriding the worry of #inflation, as the #cpi data was benign this morning. the fed's likely pivot wednesday shall boost this implication. https://t.co/9fwrlaypju
Data Test/ Prediction	Money Supply contraction hasn't happened in 90 years.	money supply contraction hasn't happened in 90 years.
	Periods of #M2 Contraction other than 2023: -Great Depression 1929 -Depression of 1921 -Panic of 1893 -1870s Banking Crisis	periods of #m2 contraction other than 2023: -great depression 1929 -depression of 1921 -panic of 1893 -1870s banking crisis
	All previous situations had #unemployment rate north of 10% and massive bank failures.	all previous situations had #unemployment rate north of 10% and massive bank failures.
	#Recession	#recession

b. *Removal of Mentions*

Mentions dalam Twitter biasanya digunakan untuk menyebutkan *user* ke dalam *tweet*. Menghapus *mentions* atau identitas *user* Twitter dilakukan untuk menghilangkan bias atau *noisy* dalam data. Tabel 3 merupakan hasil dari penghapusan mentions dan 'rt'.

Tabel 3.
Hasil Proses Hapus *Mentions*

	Tweet awal	*Tweet* hasil
Data Train	a homebuilder stock that could more than double from here. #meritagehomes #homebuilderstocks #investing #recession #mortgagerate #stockmarketnews #cnbc #powerlunch #housing market #stephenkim #usec onomy #ratehikes @evercoreisi https://t.co/vdcn7kkw4x https://t.co/ihqotlfkaw	a homebuilder stock that could more than double from here. #meritagehomes #homebuilderstocks #investing #recession #mortgagerate #stockmarketnews #cnbc #powerlunch #housing market #stephenkim #usec onomy #ratehikes https://t.co/vdcn7kkw4x https://t.co/ihqotlfkaw
Data Test/ Prediction	rt @businessinsider: the us economy is slowing down - just look at the freight recession that's sent diesel prices tumbling https://t.co/gh	the us economy is slowing down - just look at the freight recession that's sent diesel prices tumbling https://t.co/gh

c. *Removal of Special Char*

Pada tahap ini dilakukan proses menghilangkan karakter khusus apa pun yang tidak termasuk dalam kategori karakter alfanumerik atau spasi seperti Karakter non-alfanumerik, termasuk tanda baca, simbol matematika, dan lainnya. Tujuan dari proses ini adalah untuk menjamin bahwa data tekstual tidak memiliki karakter asing yang berpotensi mempengaruhi kinerja algoritma yang digunakan untuk analisis teks. Tabel 4 merupakan hasil proses penghapusan karakter spesial.

Tabel 4.
Hasil Proses Hapus *Special Char*

	Tweet awal	*Tweet* hasil
Data Train	#oott #wti	oott wti
	implication 1 could be due to the threat of #recession overriding the worry of #inflation, as the #cpi data was benign this morning. the fed's likely pivot wednesday shall boost this implication. https://t.co/9fwrlaypju	implication 1 could be due to the threat of recession overriding the worry of inflation as the cpi data was benign this morning the feds likely pivot wednesday shall boost this implication httpstco9fwrlaypju
Data Test/ Prediction	u.s. consumer confidence fell to a nine-month low in april led by a darkening outlook that augers a recession beginning in the near future, a survey showed on tuesday. https://t.co/xr0wz3kd5e	us consumer confidence fell to a ninemonth low in april led by a darkening outlook that augers a recession beginning in the near future a survey showed on tuesday httpstcoxr0wz3kd5e

d. *Removal of Stop Words*

Stopwords adalah kata-kata yang paling umum dalam *natural language*. Untuk tujuan menganalisis data teks dan membangun model NLP, *stopwords* ini tidak menambah banyak nilai pada makna dokumen. Maka kata-kata yang termasuk dalam *stop words* seperti "the", "is", "in", "for", "where", "when", "to", "at" dan lain-lain dihilangkan. Pada penelitian ini *library* Gensim digunakan dalam menghapus *stop words*. Gensim adalah *library* yang sangat berguna untuk digunakan dalam tugas-tugas NLP. Tabel 5 merupakan hasil dari proses penghapusan *stop words*.

Tabel 5.
Hasil proses stop words menggunakan gensim

	Tweet awal	*Tweet* hasil
Data Train	oott wti	oott wti implication 1
	implication 1 could be due to the threat of recession overriding the worry of inflation as the cpi data was benign this morning the feds likely pivot wednesday shall boost this implication httpstco9fwrlaypju	threat recession overriding worry inflation cpi data benign morning feds likely pivot wednesday shall boost implication httpstco9fwrlaypju
Data Test/ Prediction	as economists and investors scour data on inflation jobs housing	economists investors scour data inflation jobs housing banking bellwether

	Tweet awal	*Tweet* hasil
	banking and other bellwether indicators to determine whether the united states is headed for a recession a visit to the nations largest foodbank warehouse offers some ominous clues	indicators determine united states headed recession visit nations largest foodbank warehouse offers ominous clues httpstcopnhrf0e61i

e. *Removal of Links*

Links/tautan tidak memberikan wawasan berharga yang substansial untuk analisis sentimen dan berpotensi menimbulkan bias pada hasil jika tidak dihilangkan. Maka proses penghapusan tautan dilakukan dalam penelitian ini. Tabel 6 merupakan hasil proses penghapusan tautan.

Tabel 6.
Hasil Proses Hapus Tautan

	Tweet awal	*Tweet* hasil
Data Train	oott wti implication 1 threat recession overriding worry inflation cpi data benign morning feds likely pivot wednesday shall boost implication httpstco9fwrlaypju	oott wti implication 1 threat recession overriding worry inflation cpi data benign morning feds likely pivot wednesday shall boost implication
Data Test/ Prediction	economists investors scour data inflation jobs housing banking bellwether indicators determine united states headed recession visit nations largest foodbank warehouse offers ominous clues httpstcopnhrf0e61i	economists investors scour data inflation jobs housing banking bellwether indicators determine united states headed recession visit nations largest foodbank warehouse offers ominous clues

f. *Removal of Numbers*

Pada kasus ini, nilai numerik tidak memiliki arti penting. Menghilangkan nilai numerik dari data akan meningkatkan konsistensi dan memudahkan penanganannya dalam model algoritma pembelajaran mesin. Tabel 7 adalah hasil dari proses penghapusan numerik.

Tabel 7.
Hasil Proses Hapus Numerik

	Tweet awal	*Tweet* hasil
Data Train	oott wti implication 1 threat recession overriding worry inflation cpi data benign morning feds likely pivot wednesday shall boost implication	oott wti implication threat recession overriding worry inflation cpi data benign morning feds likely pivot wednesday shall boost implication
Data Test/ Prediction	dust settles 2023 think recession fed rate cuts latin america outperform colombia highest rate differential orange followed mexico red brazil blue chile black carry king 2023 paolaguscoffiif	dust settles think recession fed rate cuts latin america outperform colombia highest rate differential orange followed mexico red brazil blue chile black carry king paolaguscoffiif

g. *Removal of White Spaces*

Penghapusan *white spaces* berkaitan dengan prosedur menghilangkan area atau interval kosong di tengah-tengah istilah dalam kumpulan data tekstual. Pada penelitian ini tujuan penghapusan *white spaces* adalah untuk meningkatkan keseragaman data teks dan mengurangi kerumitan komputasi yang terkait dengan analisis data.

h. *Removal of Duplicated Rows*

Duplicated rows atau adanya duplikasi *row* berpotensi mempengaruhi analisis statistik dan hasil pemodelan, yang pada akhirnya menghasilkan kesimpulan yang kurang akurat. Penghapusan entri duplikat dari kumpulan data merupakan langkah penting dalam memastikan ketepatan dan keseragamannya, sehingga mengurangi potensi ketidakakuratan dalam prosedur analisis selanjutnya. Tabel 8 merupakan hasil proses penghapusan data *tweet* yang sama atau duplikasi *row*.

Tabel 8.
Hasil Proses Hapus Duplikasi *Row*

	Jumlah *Tweet* awal	Jumlah *Tweet* hasil
Data Train	34591	28793
Data Test/ Prediction	2515	1171

i. *Removal of Unuseful Columns*

Menghilangkan kolom yang tidak relevan berpotensi mengurangi kerumitan dataset dan meningkatkan ketepatan dan kemanjuran model yang dibangun. Kolom yang akan digunakan dalam *data train* adalah hanya kolom *text* dan *targer* (*sentiment*) sedangkan pada *data test/prediction* hanya membutuhkan kolom *text*.

3.2.2 Lexical analysis

a. *Tokenizing*

Tokenisasi adalah proses segmentasi urutan string menjadi unit-unit diskrit yang dikenal sebagai token, yang dapat mencakup kata, frasa, simbol, kata kunci, dan elemen-elemen lainnya. Metode yang dikenal sebagai `word_tokenize () ` digunakan untuk tokenisasi teks dalam pemrosesan bahasa alami. Fungsi ini secara efektif mengambil komponen suku kata dari unit leksikal individu. Tabel 9 adalah *sample* hasil dari proses tokenisasi.

Tabel 9.
Hasil Proses *Tokenization*

	Tweet awal	*Tweet* hasil
Data Train	barking loud pay debt interesting biden gives company bails em blackrock recession interestrates	['barking', 'loud', 'pay', 'debt', 'interesting', 'biden', 'gives', 'company', 'bails', 'em', 'blackrock', 'recession', 'interestrates']
Data Test/ Prediction	economists investors scour data inflation jobs housing banking bellwether indicators determine united states headed recession visit nations largest foodbank warehouse offers ominous clues	['economists', 'investors', 'scour', 'data', 'inflation', 'jobs', 'housing', 'banking', 'bellwether', 'indicators', 'determine', 'united', 'states', 'headed', 'recession', 'visit', 'nations', 'largest', 'foodbank', 'warehouse', 'offers', 'ominous', 'clues']

b. *Stemming*

Stemming adalah teknik dalam pemrosesan bahasa alami yang melibatkan pengurangan kata menjadi bentuk dasar atau akarnya, yang biasa disebut sebagai *stem*. Tujuan *stemming*

adalah untuk merampingkan data tekstual dengan menghilangkan sufiks morfologis. Tabel 10 adalah *sample* hasil dari *stemming*.

Tabel 10.
Hasil Proses *Stemming*

	Tweet awal	*Tweet* hasil
Data Train	barking loud pay debt interesting biden gives company bails em blackrock recession interestrates	bark loud pay debt interest biden give compani bail em blackrock recess interestr
Data Test/ Prediction	economists investors scour data inflation jobs housing banking bellwether indicators determine united states headed recession visit nations largest foodbank warehouse offers ominous clues	economist investor scour data inflat job hous bank bellweth indic determin unit state head recess visit nation largest foodbank warehous offer omin clue

c. *Lemmatization*

Lemmatisasi mengacu pada proses linguistik yang mengubah kata menjadi bentuk dasar atau bentuk dasarnya, yang juga dikenal sebagai lemma. Proses ini melibatkan pemeriksaan morfologi istilah dan konteks di sekitarnya untuk memastikan bentuk dasarnya. Tabel 11 adalah *sample* hasil dari lemmatisasi.

Tabel 11.
Hasil Proses *Lemmatization*

	Tweet awal	*Tweet* hasil
Data Train	barking loud pay debt interesting biden gives company bails em blackrock recession interestrates	bark loud pay debt interest biden give compani bail em blackrock recess interestr
Data Test/ Prediction	economists investors scour data inflation jobs housing banking bellwether indicators determine united states headed recession visit nations largest foodbank warehouse offers ominous clues	economist investor scour data inflat job hous bank bellweth indic determin unit state head recess visit nation largest foodbank warehous offer omin clue

3.3 *Weighting Data*

Proses pembobotan data dilakukan dengan pendekatan TF-IDF. Gambar 2 merupakan *script* yang digunakan.

```
# Fit the TF-IDF Vectorizer :
vectoriser = TfidfVectorizer(ngram_range=(1,2), max_features=10000)
vectoriser.fit(X_train)
print('No. of feature_words: ', len(vectoriser.get_feature_names_out()))
```

```
No. of feature_words:  10000
```

```
# Transform the data using TF-IDF Vectorizer :
X_train = vectoriser.transform(X_train)
X_test  = vectoriser.transform(X_test)
```

Gambar 2 *Syntax* TF-IDF

3.4 *Modeling*

Splitting data dilakukan dengan persentase 85% untuk data latih dan 15% untuk data pengujian. Gambar 3 merupakan syntax yang digunakan untuk memisahkan data.

```
# Memisahkan 85% data untuk data pelatihan dan 15% untuk data pengujian
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.15,
                                                    random_state=100)
```

Gambar 3 *Syntax splitting data*

Gambar 4 merupakan *script* yang digunakan untuk melakukan *modeling* menggunakan algoritma *Bernoulli Naive Bayes* (BNB), *Support Vector Machine* (SVM), Regresi *Linear*, *K-Nearest Neighbors* (KNN), dan *Decision Tree*.

```
def model_Evaluate(model):
    # Predict values for Test dataset
    y_pred = model.predict(X_test)
```

```
# Model-1 : Bernoulli Naive Bayes.
BNBmodel = BernoulliNB()
BNBmodel.fit(X_train, y_train)
model_Evaluate(BNBmodel)
y_pred1 = BNBmodel.predict(X_test)
# Model-2 : SVM (Support Vector Machine).
SVCmodel = LinearSVC()
SVCmodel.fit(X_train, y_train)
model_Evaluate(SVCmodel)
y_pred2 = SVCmodel.predict(X_test)
# Model-3 : Logistic Regression.
LRmodel = LogisticRegression(C = 2, max_iter = 1000, n_jobs=-1)
LRmodel.fit(X_train, y_train)
model_Evaluate(LRmodel)
y_pred3 = LRmodel.predict(X_test)
# Model-4 : k-nearest neighbors.
int(sqrt(len(df)))
knn = KNeighborsClassifier(n_neighbors=int(sqrt(len(df))))
knn.fit(X_train, y_train)
y_pred4 = knn.predict(X_test)
# Model-5 : Decision Tree
clf = DecisionTreeClassifier()
start1 = time.time()
clf = clf.fit(X_train, y_train)
LRmodel.fit(X_train, y_train)
model_Evaluate(clf)
y_pred5 = clf.predict(X_test)
```

Gambar 4 Algoritma model

3.5 *Accuracy Evaluation and Comparison Model*

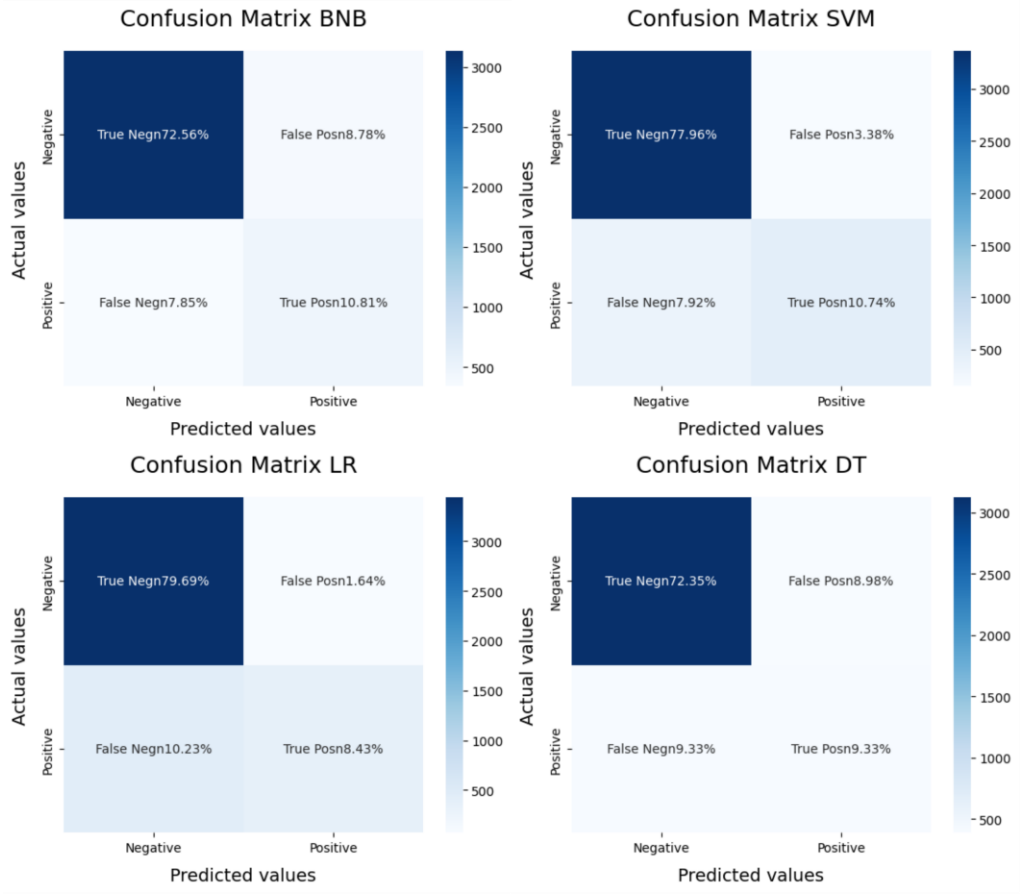
Pada tahap ini menampilkan metrik kinerja dari lima model klasifikasi yang berbeda yaitu *Bernoulli Naive Bayes*, *Support Vector Machine*, Regresi Logistik, *K-nearest Neighbour*, dan *Decision Tree* yang pada tahap sebelumnya dilakukan *training* dengan menggunakan dataset *train*. Model-model ini dinilai berdasarkan nilai akurasi, nilai F1 untuk kelas 0 dan kelas 1, nilai AUC, serta waktu eksekusi pelatihan dan pengujian. Perbandingan dari performa kelima model ditampilkan pada Tabel 12.

Tabel 12.
Hasil Perbandingan Model

Model Id	*Model Name*	*Accuracy*	*F1-score (class 0)*	*F1-score (class 1)*	*AUC Score*	*Training execution time in seconds*	*Testing execution time in seconds*
1	*Bernoulli Naïve Bayes* (BNB)	83%	90%	57%	74%	0,01	0,19
2	*Support Vector Machine* (SVM)	89%	93%	66%	77%	0,25	0,10
3	*Logistic Regression* (LR)	88%	93%	59%	72%	4,87	0,42
4	*K-nearest Neighbors* (KNN)	82%	NaN	NaN	52%	0,00	NaN
5	*Decision Tree* (DT)	82%	89%	51%	70%	23,86	0,12

Berdasarkan Tabel 12, *Support Vector Machine* menunjukkan tingkat presisi terbesar yaitu 89% dan skor F1 tertinggi untuk kedua kategori, yang menunjukkan kemampuan prediksi yang unggul untuk kategori *positive* dan *negative*. Pengklasifikasi *Bernoulli Naive Bayes* menunjukkan tingkat akurasi yang tinggi, namun, skor F1-nya untuk kelas 1/*negative* rendah, yang menunjukkan kinerja yang tidak memadai dalam memprediksi kelas *positive*. Model Regresi Logistik menunjukkan tingkat presisi tinggi kedua yaitu 88% setelah Model *Support Vector Machine* dan F1-*score* yang tinggi dalam kaitannya dengan kelas 0/*positive*, tetapi memberikan hasil yang kurang optimal untuk kelas 1/*positive* dibandingkan model *Support Vector Machine*. Algoritma *K-Nearest Neighbors* tidak memiliki F1-*score*, sedangkan algoritma *Decision Tree* menunjukkan tingkat akurasi dan F1-*score* yang cukup baik untuk kelas 0/*negative*, tetapi menunjukkan kinerja yang kurang optimal untuk kelas 1/*positive*. Hasil penelitian menunjukkan bahwa waktu eksekusi pelatihan untuk model *Logistic Regression* dan *Decision Tree* sangat tinggi, manandakan kebutuhan yang lebih besar untuk sumber daya komputasi selama pelatihan. Namun, waktu eksekusi pengujian untuk semua model relatif rendah, yang mengindikasikan kinerja yang efisien dalam hal sumber daya komputasi selama pengujian.

Pada penelitian ini *Confussion Matrix* digunakan untuk melihat performa model dalam memprediksi. Gambar 5 merupakan perbandingan *confusion matrix* yang dihasilkan dari model yang diuji.

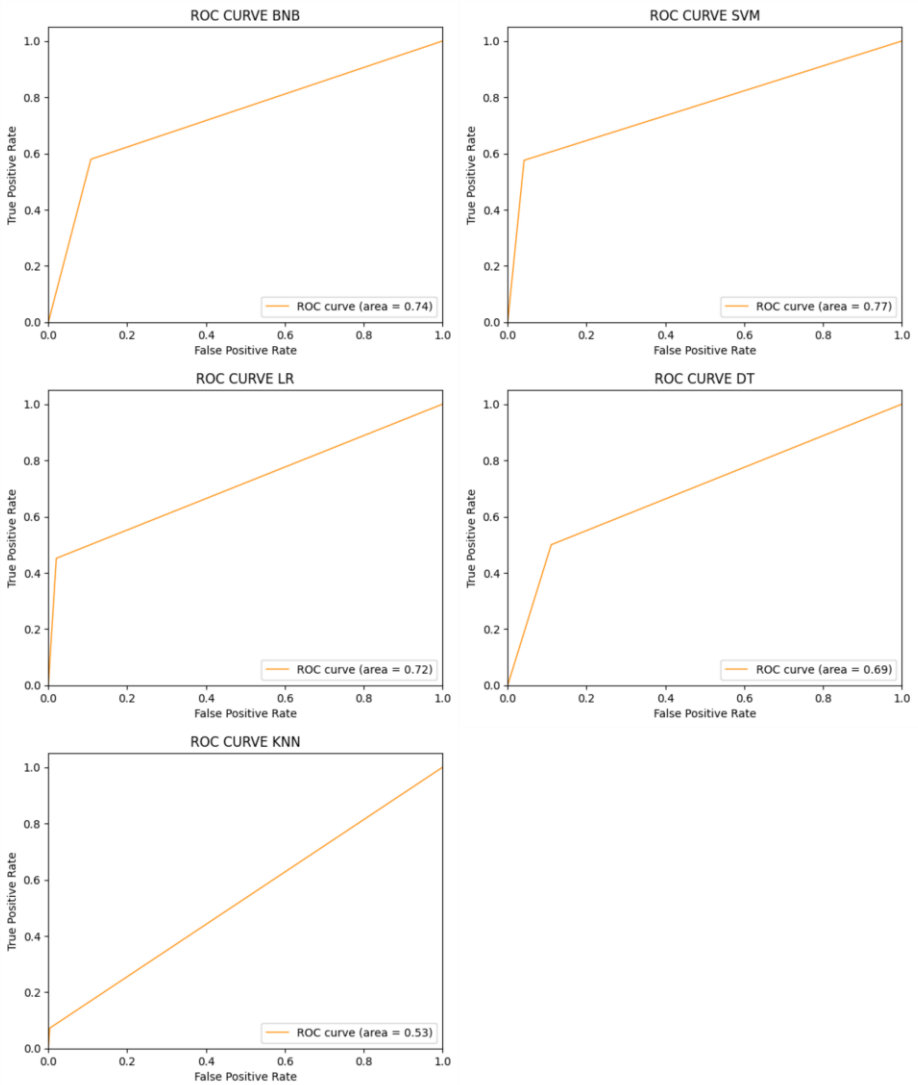


Gambar 5 Perbandingan *confusion matrix model*

Dari Gambar 5, model *Bernoulli Naive Bayes* mempunyai tingkat *True Negative* (TN) yang tinggi yaitu 72,56%, menunjukkan bahwa model ini mampu memprediksi *sentiment negative* dengan baik. Namun, model ini memiliki tingkat *False Negative* (FN) yang relatif tinggi yaitu 7,85%, menunjukkan bahwa model ini kurang baik dalam mengidentifikasi *sentiment* positif. Tingkat *True Positive* (TP) adalah 10,81%, menunjukkan kapasitas moderat untuk meramalkan kejadian positif, sedangkan tingkat *False Positive* (FP) adalah 8,78%, menunjukkan jumlah moderat prediksi positif yang tidak akurat. Pada model *Support Vector Machine*, model ini berkinerja sedikit lebih baik dengan tingkat *True Negative* (TN) yang lebih tinggi yaitu 77,96% dan tingkat *False Negative* (FN) yang lebih rendah yaitu 7,92%. Selain itu, model ini memiliki tingkat *False Positive* (FP) yang rendah yaitu 3,38 persen, menunjukkan kapasitas yang lebih tinggi untuk prediksi sentiment positif. Namun, tingkat *True Positive* (TP) adalah 10,74%, yang sebanding dengan model *Bernoulli Naive Bayes*.

Model *Logistic Regression* memiliki tingkat *True Negative* (TN) yang lebih tinggi dari model lain yaitu 79,69%, menunjukkan bahwa model ini bekerja dengan baik dalam meramalkan *sentiment* negatif. Namun dengan tingkat *False Negative* (FN) yang cukup tinggi yaitu 10,23%, kurang baik dalam menemukan *sentiment* yang positif. Berbeda dengan model lainnya, tingkat *False Positive* (FP) model ini rendah yaitu 1,64%, sementara tingkat *True Positive* (TP) model lebih tinggi yaitu 8,43, yang menunjukkan kurangnya kemampuan dalam prediksi *sentiment* positif. Sedangkan model *Decision Tree* memiliki tingkat *True Negative* (TN) yaitu 72,35% dan tingkat *False Negative* (FN) yaitu 9,33%, yang menunjukkan bahwa model kesulitan dalam memprediksi kejadian positif dan negatif secara akurat. Tingkat *True Positive* (TP) adalah 9,33%, menunjukkan kemampuan moderat untuk mengantisipasi kejadian positif, sedangkan tingkat *False Positive* (FP) adalah 8,98%, menunjukkan proporsi moderat dari prediksi positif yang salah.

Secara keseluruhan, Dengan tingkat *True Negative* (TN) yang lebih besar, tingkat *False Negative* (FN) yang lebih rendah, dan tingkat FP yang relatif rendah, model SVM memiliki kinerja terbaik secara keseluruhan di antara model-model lainnya. Gambar 6 merupakan bentuk kurva ROC yang dihasilkan masing-masing model yang diuji.



Gambar 6 Perbandingan kurva roc model

Berdasarkan perbandingan visual dari kurva ROC Gambar 6, model *Support Vector Machine* mempunyai daerah atas kurva lebih kecil dibanding dengan model lainnya. Selain itu model dengan nilai AUC tertinggi adalah model *Support Vector Machine* yaitu sebesar 0,77. Perbandingan ini menunjukkan jika model *Support Vector Machine* mempunyai kemampuan yang lebih unggul dalam membedakan antara *sentiment negative* dan *positive* dibanding model lainnya.

3.6 Saving Model with Pickle

Tahap selanjutnya adalah menyimpan model yang telah dibuat. Penyimpanan model ini dilakukan untuk digunakan kembali pada predikasi *sentiment* twitter dengan data *tweets* yang telah dikumpulkan dengan *crawling* menggunakan API Twitter sebelumnya. Modul yang dipakai pada penelitian ini adalah Pickle.

3.7 Twitter Sentiment Prediction of Recession 2023

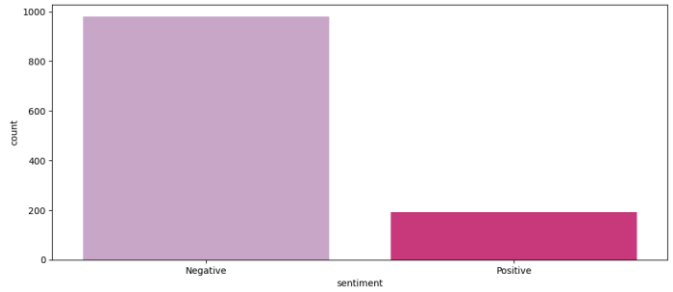
Model yang digunakan dalam *sentiment* ini adalah menggunakan Model *Support Vector Machine* yang mempunyai nilai akurasi lebih tinggi yaitu 89% dibandingkan model lainnya. Hasil prediksi *sentiment* yang didapatkan dapat dilihat pada Tabel 13.

Tabel 23.
Hasil Prediksi *Sentiment*

Sentiment	Total	*Percentage*
Negative	980	81,7%
Positive	191	16,3%

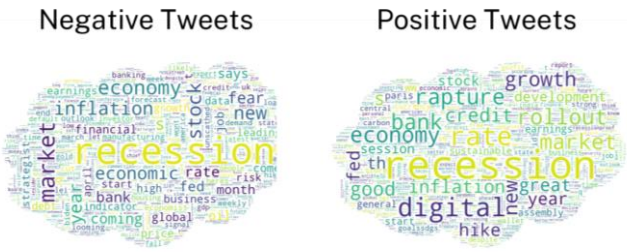
Berdasarkan data yang disajikan pada tabel, menunjukkan bahwa mayoritas pengguna Twitter mempunyai *sentiment negative* terhadap topik-topik yang berkaitan dengan resesi 2023 dengan hasil presentasi prediksi 81,7% lebih tinggi daripada *sentiment positive* yang memiliki presentasi 16,3%.

Gambar 7 merupakan grafik yang menggambarkan distribusi sentimen yang diekspresikan dalam *tweet* yang dianalisis.



Gambar 7 Visualisasi perbandingan *sentiment* pada *tweet* resesi 2023

Pada Gambar 8 dapat dilihat hasil dari visualisasi *word cloud* yang dihasilkan dari analisis *sentiment*. Pada *tweet negative* kata yang muncul adalah *recession, fear, risk*. Sedangkan pada *tweet* positif kata yang muncul adalah *development, growth, good* dan lain-lain.



Gambar 8 Visualisasi *word cloud sentiment* pada *tweet* resesi 2023

4. Kesimpulan

Dataset yang digunakan dalam pelatihan model diambil dari Kaggle tentang isu resesi 2023 pada Q4 tahun 2022 yang telah diberi label *negative* dan *positive*. Hasil dari perbandingan akurasi model adalah model SVM mempunyai nilai akurasi dan skor F1 pada kedua *class* lebih besar daripada model lainnya yaitu akurasi 89%, skor F1 kelas *negative* 93%, skor F1 kelas positive 77%. Waktu eksekusi pelatihan untuk model Regresi Logistik dan *Decision Tree* membutuhkan waktu lebih lama dibanding model lain, manandakan kebutuhan yang lebih besar untuk sumber daya komputasi selama pelatihan. Namun, waktu eksekusi pengujian untuk semua model relatif rendah, yang mengindikasikan kinerja yang efisien dalam hal sumber daya komputasi selama pengujian. Hasil prediksi *sentiment* menggunakan model SVM mengenai isu resesi 2023 pada dataset *tweet* yang diambil dalam rentang waktu 20 April 2023 s/d 26 April 2023 menunjukkan bahwa mayoritas pengguna Twitter mempunyai *sentiment negative* terhadap topik-topik yang berkaitan dengan resesi 2023 dengan hasil presentasi prediksi 81,7% lebih tinggi daripada sentiment positive yang memiliki presentasi 16,3%. Fenomena presentasi *sentiment negative* lebih tinggi ini dapat mencerminkan kecemasan akan potensi dampak ekonomi yang merugikan dan ketidakstabilan keuangan secara umum. Penelitian selanjutnya dapat melibatkan perbandingan algoritma pada penelitian ini dengan algoritma *machine learning* lainnya dan menggunakan pendekatan lain seperti *deep learning*. Selain itu dapat menggunakan dataset yang berbeda dan menerapkan teknik *sampling* data untuk mencapai keseimbangan data, yang berpotensi meningkatkan akurasi analisis sentimen.

Ucapan Terima Kasih

Terima kasih kepada dosen pembimbing dan rekan-rekan yang telah memberikan bantuan yang sangat berharga dalam pembuatan dan penyelesaian penelitian ini.

Referensi

[1] "Business Cycle Dating Procedure: Frequently Asked Questions," NBER. Accessed: Apr. 28, 2023. [Online]. Available: https://www.nber.org/research/business-cycle-dating/business-cycle-dating-procedure-frequently-asked-questions

[2] M. Egan, "There's a 98% chance of a global recession, research firm warns | CNN Business," CNN. Accessed: Apr. 28, 2023. [Online]. Available: https://www.cnn.com/2022/09/28/economy/recession-global-economy/index.html

Email: 207006020@student.unsil.ac.id (A'izzah, V. R.), rianto@unsil.ac.id (Rianto)*, vega.purwayoga@unsil.ac.id (Purwayoga, V.)

*Penulis korespondensi

DOI: https://doi.org/10.25124/jrsi.v11i01.612

[3] A. Morrow, "5 signs the world is headed for a recession | CNN Business," CNN. Accessed: Apr. 28, 2023. [Online]. Available: https://www.cnn.com/2022/10/02/business/global-recession-fears-explained/index.html

[4] A. Rahman and Md. S. Hossen, "Sentiment Analysis on Movie Review Data Using Machine Learning Approach," in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sep. 2019, pp. 1–4. doi: 10.1109/ICBSLP47725.2019.201470.

[5] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Nov. 2019, pp. 266–270. doi: 10.1109/SMART46866.2019.9117512.

[6] S. Mishra, M. Aggarwal, S. Yadav, and Y. Sharma, "Comparison of Machine Learning Techniques for Sentiment Analysis," in *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, May 2023, pp. 184–191. doi: 10.1109/ACCESS57397.2023.10200806.

[7] S. A. Sutresno, "Analisis Sentimen Masyarakat Indonesia Terhadap Dampak Penurunan Global Sebagai Akibat Resesi di Twitter," *bits*, vol. 4, no. 4, Mar. 2023, doi: 10.47065/bits.v4i4.3149.

[8] G. A. Trianto, T. Y. Sihotang, M. F. Marzuki, and H. Irsyad, "Klasifikasi Opini Terhadap Resesi Indonesia 2023 pada Twitter Menggunakan Algoritma Decesion Tree," *MDP-SC*, vol. 2, no. 1, pp. 1–9, Apr. 2023, doi: 10.35957/mdp-sc.v2i1.3997.

[9] M. S. Kalaivani, S. Jayalakshmi, and R. Priya, "Comparative analysis of sentiment classification using machine learning techniques on Twitter data," *ijhs*, pp. 8273–8280, May 2022, doi: 10.53730/ijhs.v6nS2.7098.

[10] Y. Indulkar and A. Patil, "Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis," in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India: IEEE, Mar. 2021, pp. 295–299. doi: 10.1109/ESCI50559.2021.9396925.

[11] "Standard search API." Accessed: Apr. 28, 2023. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

[12] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Computational Social Networks*, vol. 8, Jan. 2021, doi: 10.1186/s40649-020-00080-x.

[13] S. Wankhede, R. Patil, S. Sonawane, and Prof. A. Save, "Data Preprocessing for Efficient Sentimental Analysis," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Apr. 2018, pp. 723–726. doi: 10.1109/ICICCT.2018.8473277.

[14] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3506–3513, Jun. 2014, doi: 10.1016/j.eswa.2013.10.056.

[15] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation." arXiv, Jun. 17, 2018. doi: 10.48550/arXiv.1806.06407.

[16] "What is Sentiment Analysis? A Complete Guide for Beginners," freeCodeCamp.org. Accessed: May 04, 2023. [Online]. Available: https://www.freecodecamp.org/news/what-is-sentiment-analysis-a-complete-guide-to-for-beginners/

[17] E. Tan, "How To Train A Deep Learning Sentiment Analysis Model," Medium. Accessed: May 04, 2023. [Online]. Available: https://towardsdatascience.com/how-to-train-a-deep-learning-sentiment-analysis-model-4716c946c2ea

[18] J. C. Gonzalez, "Accuracy measures in Sentiment Analysis: the Precision of MeaningCloud's Technology," MeaningCloud. Accessed: May 04, 2023. [Online]. Available: https://www.meaningcloud.com/blog/accuracy-in-sentiment-analysis

[19] D. Z. Solan, "Evaluation of Sentiment Analysis: A Reflection on the Past and Future of NLP," Medium. Accessed: May 04, 2023. [Online]. Available: https://towardsdatascience.com/evaluation-of-sentiment-analysis-a-reflection-on-the-past-and-future-of-nlp-ccfd98ee2adc

[20] "Sentiment Accuracy: Explaining the Baseline and How to Test It - Lexalytics." Accessed: May 04, 2023. [Online]. Available: https://www.lexalytics.com/blog/sentiment-accuracy-baseline-testing/