Jurnal Sains dan Teknologi
Volume 12 Number 1, Tahun 2023, pp. 129-137
P-ISSN: 2303-3142 E-ISSN: 2548-8570
Open Access: https://doi.org/10.23887/jstundiksha.v12i1.52358

Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine

Aditiya Hermawan[1]*, Indrico Jowensen[2] , Junaedi[3], Edy[4]

[1,2,3,4]Universitas Buddhi Dharma, Tangerang, Indonesia

ARTICLE INFO

Article history:
Received September 16, 2022
Revised September 20, 2022
Accepted February 18, 2023
Available online April 25, 2023

Kata Kunci:
Analisa Sentimen, Support Vector Machine, Text Mining, Twitter.

Keywords:
Sentiment Analysis, Support Vector Machine, Text Mining, Twitter.

ABSTRAK

Setiap tahun, jumlah orang yang menggunakan media sosial bertambah seiring dengan jumlah orang yang menggunakan internet. Peningkatan tersebut diiringi dengan meningkatnya informasi pada internet yang tentunya informasi tersebut mempunyai nilai jika dilakukan analisa. Untuk menganalisa data dalam jumlah besar dapat menggunakan teknik text mining. Text mining mampu memproses untuk memperoleh informasi berkualitas tinggi dari teks. Text mining juga dapat digunakan untuk menganalisa informasi seperti sentimen dari sebuah kalimat dengan sangat cepat untuk memudahkan dalam mendapatkan informasi yang berkualitas. Informasi diproses berasal dari media sosial berbasis text yaitu twitter yang mana pengambilan data dilakukan dengan bantuan Application Programming Interface dan menggunakan kata kunci berupa sebuah kata atau hashtag. Kalimat tersebut akan dilakukan proses text mining dengan menggunakan algoritma Support Vector machine untuk menghasilkan klasifikasi dari sentimen suatu kalimat ke dalam sentiment positif, netral atau negatif. Tingkat akurasi yang dihasilkan oleh proses ini adalah sebesar 73% berdasarkan data sentimen yang dimiliki. Tingkat akurasi dalam melakukan text mining sangat dipengarui pada proses Pre-Processing karena terdapat banyak kata perlu dilakukan pengelolahan lebih lanjut.

ABSTRACT

Every year, the number of people using social media is increasing along with the number of people using the internet. This increase is accompanied by an increase in information on the internet which of course this information has value if it is analyzed. To analyze large amounts of data, text mining techniques can be used. Text mining is capable of processing to obtain high quality information from text. Text mining can also be used to analyze information such as the sentiment of a sentence very quickly to make it easier to get quality information. Processed information comes from text-based social media, namely Twitter, where data collection is carried out with the help of the Application Programming Interface and using keywords in the form of a word or hashtag. The sentence will be subjected to a text mining process using the Support Vector machine algorithm to produce a classification of the sentiments of a sentence into positive, neutral or negative sentiments. The level of accuracy produced by this process is 73% based on sentiment data. The level of accuracy in carrying out text mining is greatly influenced by the Pre-Processing process because there are many words that need to be further processed.

1. PENDAHULUAN

Indonesia merupakan salah satu negara yang memiliki pertumbuhan pengguna media sosial terbesar di dunia dalam satu tahun, yaitu 20 juta pengguna pada tahun 2019 (Kemp, 2019). Pertumbuhan Media Sosial terjadi dibarengin dengan meningkatnya jumlah pengguna internet di Indonesia yang ditunjukan dengan era keterbukaan informasi yang semakin tersebar luas, sehingga memudahkan masyarakat untuk memperoleh informasi (Pratama & Tjahyanto, 2021). Era keterbukaan informasi juga ditandai dengan pesatnya pertumbuhan data, yang sebagian besar dari data tersebut merupakan data tidak terstruktur yang sering kali berisi teks, termasuk artikle berita, posting media sosial, data Twiitter, data yang ditranskrip dari video, serta dokumen formal (Benchimol et al., 2020, 2022). Pesatnya pertubuhan data yang terdapat di internet meningkat secara besar-besaran dapat dikatakan meningkat secara eksponensial. Hal tersebut tidak lepas karena adanya Platfoam media sosial seperti facebook, Twitter dan Instagram yang terus bertambah penggunanya secara eksponens siap dalam beberapa tahun terakhir dan menyediakan sumber informasi yang berharga untuk menganalisis tren dan opini sosial. (Leelawat et al., 2022) Dengan jumlah data yang besar dan semakin bertambah cepat menyebabkan terjadinya "banjir informasi" dimana data tersebut memiliki nilai jika dapat dianalisis. Informasi yang sangat banyak ini tidak mungkin dianalisis secara manual satu per satu tanpa bantuan teknologi analisis data berbasis komputer.

*Corresponding author.
E-mail addresses: aditiya.hermawan@ubd.ac.id (Aditiya Hermawan)

Untuk menganalisis data dalam jumlah besar anda dapat menggunakan teknik text mining. Text Mining adalah proses menemukan pengetahuan yang tidak diketahui melalui ekstraksi informasi otomatis dari sejumlah besar teks yang tidak terstruktur (Kowsari et al., 2019; Nota et al., 2022; Tandel et al., 2019). Penambangan teks mampu mengambil data dalam jumlah besar dan kemudian diproses untuk mendapatkan informasi yang berguna dari kumpulan teks. Aplikasi yang dapat dilakukan dengan text mining berupa analisis sentimen pada media sosial, ekstraksi informasi untuk mengklasifikasi teks, dan peringkasan teks untuk meringkas teks panjang menjadi bentuk 1 atau 2 paragraf terus dikembangkan. Hal tersebut dapat dilihat dari upaya yang dilakukan dalam penelitian dan pengembangan sistem untuk ekstraksi otomatis informasi dari sebuah ulasan. Data dan informasi dari internet menjadi target utama penelitian di bidang text mining oleh perusahaan untuk berbagai tujuan (Ahuja et al., 2015; Pilar et al., 2022; Rathika & Soranamageswari, 2022; Starosta, 2022; Xue et al., 2021). Objek dari aplikasi text mining yang akan dilakukan pada penelitian ini adalah media sosial. Teks di media sosial dapat diambil dengan cepat dengan Text mining. Media sosial yang mengandung banyak informasi berupa text adalah twitter. Twitter dipilih sebagai objek penelitian karena Twitter memiliki lebih dari 330 juta pengguna aktif bulanan dan 145 juta pengguna aktif harian di seluruh dunia (Leelawat et al., 2022). Twitter banyak dipilih untuk diteliti dari pada sosial netwoks lainnya karena beberapa karakteristik seperti penetrasi populasi yang tinggi, digunakan banyak pengguna untuk berpendapat prihal topik yang hangat ataupun topik khusus yang sedang banyak dibahas (Batista & Ribeiro, 2013; Passi & Motisariya, 2022; Pilar et al., 2022; Villavicencio et al., 2021). Selain itu teks yang terdapat di Twitter dapat dengan mudah ditambang karena Twitter sendiri sudah menyediakan API publik yang dapat digunakan oleh siapa saja.

Sebuah teks di media sosial tidak hanya menyampaikan informasi tentang suatu informasi. Tetapi memiliki informasi lainnya seperti pendapat atau perasaan terhadap sesuatu. Infromasi tersebut dapat dianalisis dan dilakukan klasifikasi menggunakan teknik analisis sentimen. Analisis sentimen adalah sebuah teknik untuk menganalisis opini, sentimen, apresiasi, emosi tentang suatu produk, layanan, organisasi, individu, dan mereka atribut (Medhat et al., 2014). Teknik analisis sentimen juga dapat digunakan dalam menganalisis pendapat orang dalam sebuah potongan teks untuk menentukan apakah sentimen tersebut positif, negatif, atau netral (Fauzi, 2018; Pratama & Tjahyanto, 2021). Analisis sentiman juga melibatkan bilang Natural Language Processing (NLP) untuk mendeteksi informasi subjektif dalam dokumen (Pilar et al., 2022; Pintas et al., 2021). Informasi yang diambil untuk dianalisis adakan dilakukan beberpa proses yaitu dimulai dengan mengektrak data lalu data dibersihkan terlebih dahulu. Proses pembersihan data disebut pre-processing, kemudian setelah dibersihkan, data tersebut kemudian diolah dengan metode yang akan digunakan. Terdapat beberapa metode untuk melakukan klasifikasi analisis sentimen, salah satunya adalah Support Vektor Machine (SVM).

Support Vector Machine (SVM) telah muncul sebagai sebuah teknik yang kuat untuk tujuan umum pengenalan pola (Wenda, 2022). Teknik ini telah diterapkan pada masalah – masalah regresi dan klasifikasi dengan kinerja yang sangat bagus (Altawaier & Tiun, 2016; Chen et al., 2015; Kremer et al., 2014; Qian et al., 2015) serta digunakan untuk mencari teks pertambangan dengan akurasi tinggi seperti pada penelitian yang dilakukan oleh Ira Zulfa dan Edi Winarko (Zulfa & Winarko, 2017). Mereka menyebutkan bahwa menggunakan metode Support Vector Machine dapat menghasilkan akurasi sebesar 92,18% dibandingkan dengan Metode Naive Bayes Classifier dengan akurasi 79,10% untuk Analisis Sentimen Tweet Indonesia (Zulfa & Winarko, 2017). Dengan akurasi yang tinggi yang dihasilkan oleh Algoritma Support Vector Machine pada penelitian yang telah sebelumnya untuk analisis sentiman, maka diimplentasikan dengan menggunakan pendekatan analisis secara langsung data yang ada pada media sosial berbasis dengan algoritma Support Vector Machine. Implementasinya dilakukan dalam bentuk aplikasi website yang dapat digunakan secara praktis untuk menganalisis sentimen khususnya dalam bahasa Indonesia.

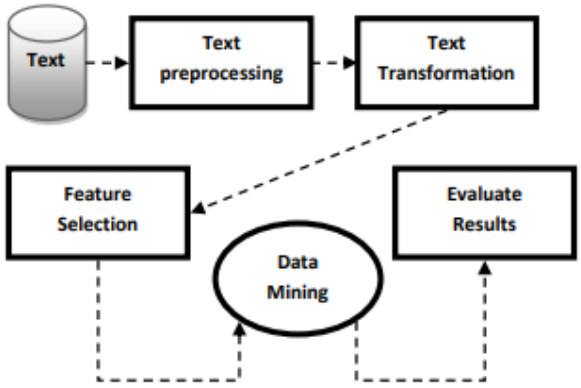
2. METODE

Metode yang digunakan pada penelitian ini adalah eksperimen, dimana data dilakukan ujicoba secara langsung menggunakan pendekatan proses Text Mining yang terdiri dari 5 proses yang dapat dilihat pada Gambar 1.

Text Pre-Processing

Tahapan *pre-processing* merupakan tugas penting yang harus dilakukan sebelum data digunakan dalam pembuatan model. Persiapan text dan penghapusan merupakan bagian dari pemrosesan data dan proses analisis yang melibatkan informasi dari kumpulan data yang digunakan dan merencanakan *pre-processing* yang dilakukan (Kashina et al., 2020) . Pada penelitian ini terdapat beberapa *pre-processing*

yang dilakukan pada data *tweet* yang terkumpul sebelum diklasifikasikan, tahapan prosesnya yaitu : *lowercasing*, *noise removal*, dan penghapusan *stopword*.



Gambar 1. Proses Text Mining (Rashid & Shoaib, 2016)

Lowercasing

Lowercasing dilakukan agar semua data yang akan digunakan adalah huruf kecil tanpa ada huruf besar, yang dapat dilihat pada Tabel 1. Hal ini karena komputer membedakan jika teks huruf besar dan huruf kecil. Data latih yang akan dianalisis memiliki huruf kecil semua. Sehingga bila terdapat data dengan huruf kapital maka akan terlewati oleh *data training* yang ada.

Tabel 1. Lowercasing

Raw	Lower cased
AMAN AmAn Aman	Aman

Noise Removal

Noise removal ini dilakukan untuk menghilangkan simbol-simbol yang ada sebelum dilakukan analisa. Pada saat pengambil data dengan Tweet terdapat banyak simbol misalnya \n yang berarti *new line* atau baris baru, kemudian simbol tanda tanya, emotikon dihilangkan agar *tweet* dapat dianalisa *data training*. Selain simbol pada *noise removal* dilakukan penghilangan *link*, yang terambil oleh *twitter* API karena tag tersebut akan memberatkan proses analisa. Hal ini dapat dilihat pada Tabel 2.

Tabel 2. Noise Removal

Tweet	Tweet tanpa Noise
@MVSOLAR Ada psbb segala sih 	mvsolar ada psbb segala sih
@steffislsbl Kalo di tempat gue udh ga psbb jd dah banyak yang nongki	steffislsbl alo di tempat gue udh ga psbb jd dah banyak yang nongki

Stopword

Stopword mengandung kata-kata yang kurang berarti untuk dianalisa dan dapat mengurangi kualitas data yang di analisa. *Tujuan penhilangan stopword* dilakukan untuk mengurangi jumlah teks yang akan diproses, sehingga proses yang dilakukan menjadi lebih cepat. Untuk menghilangkan *stopword* dalam bahasa Indonesia digunakan *library Python* Sastrawi. Sastrawi adalah sebuah modul NLP dalam bahasa Indonesia yang dibuat oleh (Robbani, 2016). *Library* tersebut juga memiliki fitur penghapusan *stopword* yang dapat digunakan dengan class *StopWordRemoverFactory*.

Text Transformation

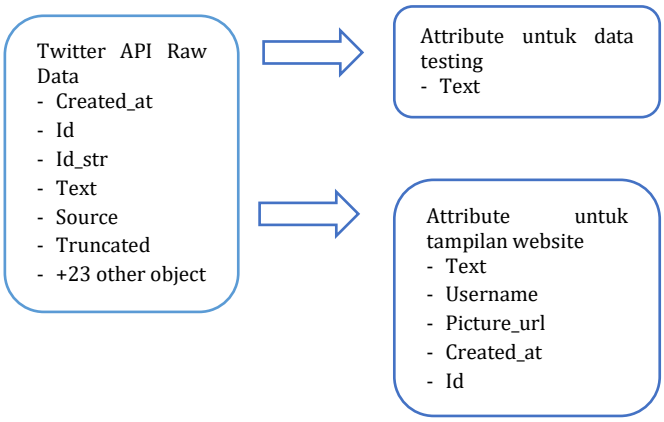
Setelah data dibersihkan, selanjutnya dilakukan proses *Teks Tranformasi* atau dikenal dengan Tokenisasi. Tokenisasi merupakan proses mengubah teks yang lebih besar menjadi sekumpulan teks yang kecil atau disebut token. Token dapat berupa kata, kalimat atau paragraf (Riesener et al., 2022). Tokenisasi dapat memisahkan tulisan yang ada dengan mencari pembatas dalam sebuah kata.Pembatasan

kata yang umum digunakan adalah tanda spasi (*Whitespace*). Tokenisasi dilakukan agar setiap potongan teks dalam sebuah tweet dapat dibersihkan.

Contoh proses tokenisasi :
"Kondisi Indonesia pada saat ini sangat Aman."
Maka setelah proses tokenisasi kalimat tersebut berubah menjadi
"Kondisi" "Indonesia" "pada" "saat" "ini" "sangat" "Aman" – **Jumlah Token 7**

Feature Selection

Feature selection adalah pemilihan atribut yang perlu sebelum dilakukan proses analisis data. Proses ini terjadi karena banyak data yang tersedia dan jumlahnya terus bertambah sehingga perlu dilakukan hal tersebut agar dapat meningkatkan nilai akurasi, serta meningkakan performace pada data berdimensi sangat tinggi (Kartiwi et al., 2018). Selain itu feature selection attribut dilakukan agar bentuk data yang akan dianalisa menjadi lebih sederhana dan mengurangi jumlah ukuran data untuk di proses saat pengumpulan data. Attribut selection ini dilakukan karena tidak semua atribut yang diambil oleh Tweepy akan digunakan untuk dianalisa dan membuat ukuran data yang ada menjadi lebih kecil sehingga dapat mempercepat proses pembersihan data dan proses menganalisa kata pada sebuah tweet. Attribut selection dapat dilihat pada Gambar 2.



Gambar 2. Attribute Selection

Data Mining

Pada tahap ini dilakukan proses klasifikasi, data akan diklasifikasi menjadi data kalimat positif dan negatif dengan menggunakan algoritma *Support Vector Machine* (SVM). SVM adalah teknik statistik dan pembelajaran mesin dengan tujuan utama prediksi. Mereka dapat diterapkan untuk terus menerus, hasil biner, dan kategoris analog dengan Gaussian, logistik, dan multinomial regresi (Guenther & Schonlau, 2016). Proses klasifikasi ini dimulai dengan membuat vektor fitur untuk data latih. Fitur Vector ini dibuat untuk mengubah tulisan menjadi bentuk vektor sehingga proses klasifikasi dapat dilakukan oleh Support Vector Machine. Proses ini menggunakan library milik sklearn dengan modul feature_extraction.text pada fungsi TfidfVectorizer.fit_transform.

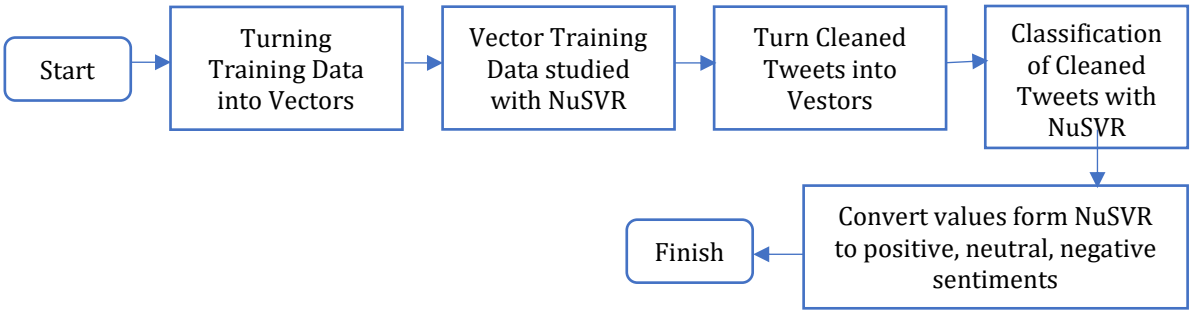
Setelah data pelatihan diubah menjadi vektor, vektor ditempatkan di suatu tempat untuk mempelajari pola dengan Support Vector Machine. Klasifikasi yang digunakan dalam penelitian ini menggunakan Nu Support Vector Regression. NuSVR memiliki properti yang sama dengan Support Vector Regression biasa, namun pada NuSVR dalam proses regresi, parameter nu digunakan untuk mengontrol jumlah support vector. Pada data latih yang digunakan kalimat positif bernilai 1, sedangkan kalimat negatif bernilai -1. Karena menggunakan NuSVR, nilai hasil klasifikasi memiliki nilai maksimal 1 dan nilai terkecil -1. Pemberian nilai pada kata positif dan negatif dilakukan agar dalam pengklasifikasian dapat mengeluarkan output sentimen netral ketika tidak ada kata positif atau negatif dalam kalimat tersebut. Hasil perhitungan SVM untuk setiap tweet memiliki output berupa angka desimal. Berdasarkan nilai yang dikeluarkan, dibuat tabel konversi untuk menentukan tweet tersebut memiliki sentimen positif, netral atau negatif, yang dapat dilihat dalam Tabel 3.

Tabel 3. Value Conversion

Classification	Value Threshold
Positive	$x \geq 0.5$
Neutral	$-0.5 \leq x \leq 0.5$
Negative	$x \leq -0.5$

Setelah data dilatih, dilakukan proses klasifikasi data dari tweet. Data pengujian berupa tweet yang sudah dibersihkan diubah menjadi bentuk vektor juga. Setelah diubah menjadi vektor data akan diklasifikasikan. Jika tweet memiliki kata sentimen negatif maka penempatan vector support berada pada area yang memiliki nilai negatif seperti -0,6 sedangkan jika memiliki sentimen positif nilai nilai positifnya adalah 0,56. Nilai support vector dari tweet sangat dipengaruhi oleh data training yang digunakan dan support vector yang ada sangat berhubungan satu sama lain. Misalnya ada 1 sentimen positif yang bernilai 1. Jika ada kata yang bukan sentimen atau sentimen negatif, maka nilai vector support yang ada digeser. Jika hanya ada satu sentimen dalam sebuah tweet sentimen, tetapi sisanya tidak mengirimkan sama sekali dalam jumlah besar, secara keseluruhan tweet tersebut dianggap tidak memiliki sentimen. Oleh karena itu, dilakukan penghilangan kata depan (stopword) agar hasil klasifikasi lebih baik.

Data yang telah dilatih akan disimpan menggunakan pickle. Pickle digunakan untuk menggunakan kembali model dan vektor dari data latih yang telah dilakukan sehingga pada saat ingin melakukan klasifikasi ulang data tidak perlu melakukan latih ulang agar proses klasifikasi data dapat dilakukan lebih cepat. Proses pickle ini dilakukan dengan cara dumping model yang telah dilatih dan vektor yang sudah ada. Dump data pelatihan ini disebut bentuk vektor dump classifier.sav dan vectorizer.sav. Proses klasifikasi sentimen dapat diliahat pada Gambar 3.



Gambar 3. Proses Klasifikasi Sentimen

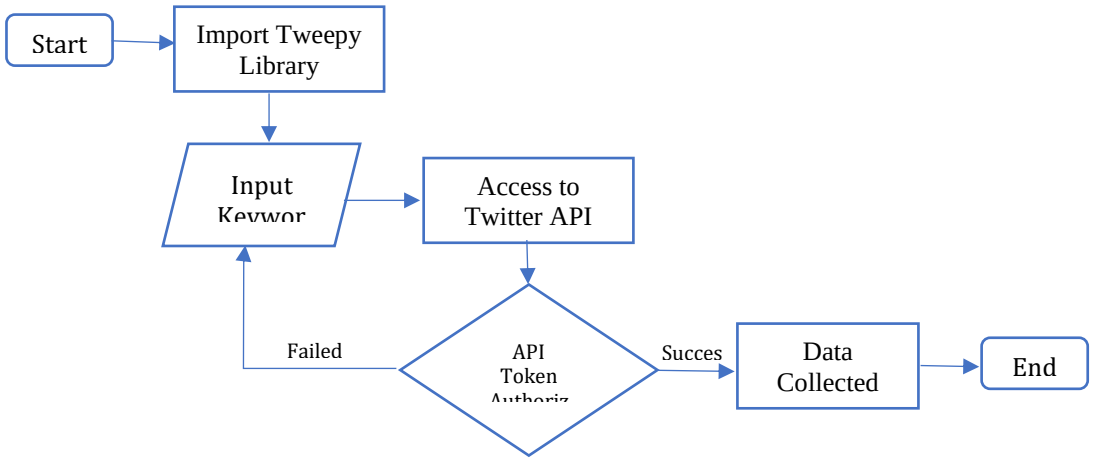
Evaluasi

Hasil dari *tweet* yang sudah dianalisa yaitu *tweet* yang memiliki sentimen positif atau sentimen negatif kemudian dilakukan pengukuran dengan menggunakan perhitungan *confusion matrix* didapatkan hasil algoritma yang efektif untuk melihat tingkat akurasi, presisis, recall dan prevalancenya (Saputra, 2021).

Metode Pengumpulan Data

Dalam penelitian ini ada 2 jenis data yang dikumpulkan. Data Pelatihan berupa kata-kata yang mengandung sentimen positif atau negatif dan Data Pengujian dari tweet yang dikumpulkan oleh API Twitter. Tujuan dari pengumpulan data pelatihan ini adalah agar mesin dapat mempelajari kata-kata yang memiliki sentimen sehingga mesin dapat mengklasifikasikan data di Twitter yang memiliki sentimen positif atau negatif. Data pelatihan yang digunakan berasal dari Daftar Kata Opini Bing Liu (Liu et al., 2005) yang telah dimodifikasi oleh Wahid, D.H ke dalam bahasa Indonesia. Kata-kata positif dan negatif ini dapat diunduh di (Masdevid, 2021). Jumlah kata negatif yang akan dilatih adalah 2436 kata dan jumlah kata positif sebanyak 1197 kata sehingga total data latih yang digunakan adalah 3633 kata. Baik kata positif maupun negatif yang digunakan tidak memiliki huruf kapital, simbol sehingga pada saat ingin melakukan klasifikasi, data yang akan diuji terlebih dahulu harus diubah menjadi huruf kecil. Kata yang bermakna positif misalnya acungan jempol, adaftif, adil afinitas dan afirmasi. Sedang kata-kata yang bermakna negatif seperti abnormal, absurd, acak, acak-acakan dan acuh. Kemudian data pengujian yang dikumpulkan dalam penelitian ini adalah data primer. Data pengujian ini digunakan sebagai pengujian terhadap klasifikasi yang dilakukan. Data ini dikumpulkan menggunakan pustaka python tweepy. Pustaka tweepy di python dapat digunakan untuk mengakses API Publik yang disediakan oleh Twitter.

Selama proses pengambilan data menggunakan Twitter API dilakukan proses retweet filter. Filter retweet ini bertujuan untuk menghindari redundansi data karena pada saat pengumpulan data dengan Twitter API terdapat data yang memiliki isi yang sama. API Twitter menganggap retweet sebagai tweet biasa. Tweet yang di Retweet diawali dengan tulisan RT di depannya. Proses filter retweet dilakukan dengan menambahkan "–filter: retweets" saat menggunakan api.search dari tweepy. Selain filter retweet, parameter lang = id digunakan untuk mengambil tweet dalam bahasa Indonesia. Proses pengumpulan data dapat dilihat pada Gambar 4.



Gambar 4. Proses Pengumpulan Data

3. HASIL DAN BAHASAN

Hasil

Model yang dihasilkan untuk melakukan klasifikasi sentiman analisis menggunakan SVM diterapakn secara langsung untuk mengklasifikasi sentiment dari suatu topik pada twitter berdasarkan kata yang ingin dicari klasfikasinya dan mendapatkan informasi berupa kalsifikasi dari kalimat tersebut. Hasil dari model tersebut dapat melihat jumlah tweet dengan sentimen positif, netral dan negatif. Berdasarkan hasil tersebut, kita dapat melihat klasifikasi dari suatu topik pada twitter untuk waktu tertentu mempunyai Sentiment yang lebih banyak positif, negatif ataupun netral dengan lebih mudah dan cepat. Untuk mengetahui tingkat akurasi dari Klasifikasi yang dihasilkan oleh SVM dalam mengklasifikasi Sentimen pada data twitter maka dilakukan percobaan sebanyak 10 kali dengan *keyword* dan waktu pengambilan data yang berbeda kemudian dan dievaluasi menggunakan *confusion matriks* untuk mengukur tingkat akurasi yang dihasilkan antara percobaan, presisi yang didapatkan, nilai *recall* dan *prevelance*. Masing-masing percobaan akan mengevaluasi *tweet* sebanyak 50 *tweet* sehingga total data yang akan dievaluasi sebanyak 500 *tweet*. Proses evaluasi dilakukan dengan cara membandingkan hasil dari model sentiment analisis menggunkan SVM dengan pengecekan secara manual apakah sebuat tweet tersebut masuk ke dalam kategori Posifit, Negatif atau Netral. Setelah itu dibuat Conffusion Matrik yang disajikan pada Tabel 4.

Tabel 4. Conffusion Matrik

N = 50		**Predicted**			**Total**
		Positif	**Netral**	**Negatif**	
Actual	Positif	5	3	0	8
	Netral	1	20	0	21
	Negatif	0	9	12	21
	Total	**6**	**32**	**12**	
Actual Positif		8	Akurasi		74%
Actual Netral		21	Misclasifikasi		36%
Actual Negatif		21	True Positif Rate		63%
Predicted Positif		6	True Neutral Rate		95%
Predicted Netral		32	True Negatif Rate		57%
Predicted Negatif		12	False Positif Rate		38%
Presisi		83%	False Neutral Rate		5%
Prevalence		16%	False Negatif Rate		43%

Berikut ini adalah sampel *tweet* yang berhasil diklasifikasi dengan benar. *Tweet* "keluarga harta paling berharga cinta ibu sepanjang masa kisah adalah salah satu alasan indahnya" memberikan prediksi positif sesuai dengan aktual sentimen yang terjadi. Pemberian sentimen posifit pada *tweet* tersebut dipengaruhi oleh kata "berharga" dan "indah" yang merupakan kata yang mengandung sentimen Positif. Kemudian *tweet* "update wni positif covid luar negeri senin juni orang terinfeksi virus corona via tribun_palu" memberikan prediksi negative karena terdapat kata "terinfeksi" yang membuat sentiment pada tweet tersebut menjadi negative. Pada proses klasifikasi terdapat juga beberapa data yang salah dalam proses klasifikasinya. Contohnya pada tweet "perusahaan terkenal dunia menutup ribuan toko per juni akibat virus corona telah merugikan perusahaan" di mana tweet tersebut seharusnya mempunyai sentimen negatif, sedangkan hasil prediksinya adalah netral. Pada *tweet* tersebut terdapat kata "terkenal" yang memiliki sentimen positif namun kata "menutup" disini memiliki konotasi negatif. Kata "menutup" ini tidak terdapat pada *data training* yang digunakan serta kata "merugikan" memiliki sentimen negatif. Namun karena hanya ditemukan 1 kata positif dan negatif dalam sebuah *tweet* yang dikenali dalam proses klasifikasi maka sentiment yang dihasilkan menjadi netral.

Contoh lainnya pada tweet "personil polsek sibolga sambas melaksanakan patroli seputaran kota sibolga memberikan rasa aman sekali" yang mempunyai sentiment positif tetapi di prediksi netral. Pada t*weet* tersebut terdapat kata "aman" yang memiliki sentiment positif namun dikarenakan jumlah kata yang terdapat pada *tweet* tersebut cukup panjang. Sentimen positif yang dimiliki oleh *tweet* tersebut bisa terlewatkan. Hal ini disebabkan karena terdapat nama sebuah subjek yang cukup panjang dan tidak memiliki arti dalam proses analisa sentiment yakni "personil polsek sibolga sambas" dan "kota sibolga". Setelah dilakukan 10 percobaan pada data tweet, kemudian dihitung rata-rata akurasi, presisi recall dan prevalencenya. Hasil keseluruhan dari percobaan yang telah dilakukan disajikan pada Tabel 5.

Tabel 5. Hasil Evaluasi

Percobaan	Accuracy	Precision	Recall	Prevalence
1	74%	83%	63%	16%
2	66%	57%	50%	16%
3	74%	95%	72%	58%
4	70%	17%	50%	8%
5	74%	75%	50%	24%
6	66%	20%	14%	14%
7	76%	100%	75%	16%
8	74%	86%	75%	16%
9	76%	100%	67%	30%
10	78%	33%	25%	8%
Rata-rata	73%	67%	54%	21%

Pembahasan

Analisa Sentimen pada data twitter berbahasa indonesia menggunakan SVM memberikan akurasi yang cukup baik pada beberapa kali percobaan. Tingkat akurasi pada SVM juga sangat dipengarui dari data tweet dan hasil dari pre-proccesing dari data tersebut. Beberapa hasil uji coba yang yang salah prediksi terlihat karena beberapa kata pada tweet tidak dapat didefinisikan dengan baik sentimennya sehingga mempengaruhi hasil sentimen pada tweet tersebut. Kemudian hal lainnya yang mempengaruhi tingkat akurasi dalam mengklasifikasikan sentimen pada suatu kalimat, diantaranya adalah terdapat kata kalimat yang tidak tegas menggambarkan suatu sentimen positif ataupun negatif, karena terdapat kata positif dan negatif secara bersamaan, kemudian penggunaan kata daerah, bahasa gaul dan juga kesalahan pengetikan menjadi kendala dalam penentuan sentimen pada suatu kalimat. Selain itu penggunaan suatu kata pada nama lokasi, bangungan dan perusahaan juga berpengaruh hasil akurasi klasifikasinya. Secara keseluruhan Algoritma SVM mempunyai akurasi yang cukup baik untuk melakukan klasifikasi analisa sentimen pada data Twetter berbahasa Indonesia sejalan dengan penelitian yang dilakukan sebelumnya (Zulfa & Winarko, 2017). Beberapa hal yang perlu ditambahkan adalah pada level pre-processing yang lebih mendetail dan juga referensi kata dengan sentimen yang lebih lengkap. Serta beberapa temuan yang membuat kesalahan dalam prediksi adalah kata-kata yang digunakan untuk nama jalan, nama perusahaan atau pun nama orang. Seperti kata "Jalan Suka Hati" jika dilakukan klasifikasi mempunyai makna Positif, yang seharusnya kata tersebut bersifat netral karena sebuah nama jalan. Untuk mengatasi hal tersebut dan meningkatkan akurasi terhadap hasil klasifikasi sentimen tentunya diperlukan data training yang lebih lengkap serta dilakukan penambahan data referensi sentimen agar dapat memberikan hasil analisa yang lebih baik dan juga penyaringan kata yang lebih detail agar kata yang diklasifikasikan sudah tidak ada lagi kata yang tidak terdapat dalam data training. Selain penambahan dan perbaikan pada data

training, dapat juga menambahkan jenis sentimen yang diklasfikasifkan karena banyak juga kata yang tidak termasuk kedalam sentimen positif atau negatif saja, sehingga sentimen dari suatu kalimat bisa lebih jelas. Sentimen lainnya misalnya seperti Marah, Sedih, Senang, Semangat, dan lainnya.

4. SIMPULAN

Penggunaan text mining menggunakan SVM dalam melakukan klasifikasi pada tweet berbahasa Indonesia mempunya akurasi 73% berdasarkan pada 10 kali percobaan yang dilakukan dengan keyword dan waktu yang berbeda-beda. Kemudian nilai presisi yang didapatkan adalah 67% dan nilai *recall* yang didapatkan 54%. Tingkat akutasi, presisi dan recall yang dihasilkan pada penelitian ini sangat dipengaruhi dengan kualitas *data training* dan data referensi sentiment yang digunakan. Selain hal tersebut, ratio yang digunakan dalam membuat model juga harus dibuat seimbang untuk data yang sentimen negative dan sentiment positif. Dengan bantuan SVM dalam melakukan klasifikasi sentiment pada data twitter, hal tersebut memudahkan untuk mencari sentiment pada suatu topik menjadi lebih cepat dilakukan tanpa perlu melihat satu persatu kalimat tersebut. Sehingga mendapat informasi sentiment dari suatu topik tertentu pada kalimat yang berjumlah besar menjadi lebih mudah dengan tingkat akurasi yang cukup baik.

5. DAFTAR PUSTAKA

Ahuja, R., Rastogi, H., Choudhuri, A., & Garg, B. (2015). Stock market forecast using sentiment analysis. *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1008–1010.

Altawaier, M. M., & Tiun, S. (2016). Comparison of machine learning approaches on Arabic twitter sentiment analysis. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 1067–1073. https://doi.org/10.18517/ijaseit.6.6.1456.

Batista, F., & Ribeiro, R. (2013). Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Procesamiento Del Lenguaje Natural*, *50*, 77–84.

Benchimol, J., Kazinnik, S., & Saadon, Y. (2020). Communication and transparency through central bank texts. *132nd Annual Meeting of the American Economic Association*.

Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. *Machine Learning with Applications*, *8*(March 2021), 100286. https://doi.org/10.1016/j.mlwa.2022.100286.

Chen, D., Wang, L., & Li, L. (2015). Position computation models for high-speed train based on support vector machine approach. *Applied Soft Computing*, *30*, 758–766. https://doi.org/https://doi.org/10.1016/j.asoc.2015.01.017.

Fauzi, M. A. (2018). Random forest approach fo sentiment analysis in Indonesian language. *Indonesian Journal of Electrical Engineering and Computer Science*, *12*(1), 46–50. https://doi.org/10.11591/ijeecs.v12.i1.pp46-50.

Guenther, N., & Schonlau, M. (2016). Support Vector Machines. *The Stata Journal: Promoting Communications on Statistics and Stata*, *16*(4), 917–937. https://doi.org/10.1177/1536867X1601600407.

Kartiwi, M., Gunawan, T. S., Arundina, T., & Omar, M. A. (2018). Feature Selection for Financial Data Classification: Islamic Finance Application. *2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, 1–4. https://doi.org/10.1109/ICSIMA.2018.8688803.

Kashina, M., Lenivtceva, I. D., & Kopanitsa, G. D. (2020). Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification. *Procedia Computer Science*, *178*(2019), 284–290. https://doi.org/10.1016/j.procs.2020.11.030.

Kemp, S. (2019). *DIGITAL 2019: GLOBAL DIGITAL OVERVIEW*. https://datareportal.com/reports/digital-2019-global-digital-overview.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, *10*(4). https://doi.org/10.3390/info10040150.

Kremer, J., Steenstrup Pedersen, K., & Igel, C. (2014). Active learning with support vector machines. *WIREs Data Mining and Knowledge Discovery*, *4*(4), 313–326. https://doi.org/https://doi.org/10.1002/widm.1132.

Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., Yudha, A. K., & Tang, J. (2022). Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. *Heliyon*, *8*(10), e10894.

https://doi.org/10.1016/j.heliyon.2022.e10894.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th International Conference on World Wide Web*, 342–351. http://dl.acm.org/citation.cfm?id=1060797.

Masdevid. (2021). *Kata Positif dan Negatif*. https://github.com/masdevid/US-OpinionWords.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Nota, G., Postiglione, A., & Carvello, R. (2022). Text mining techniques for the management of predictive maintenance. *Procedia Computer Science*, *200*, 778–792. https://doi.org/10.1016/j.procs.2022.01.276.

Passi, K., & Motisariya, J. (2022). Twitter Sentiment Analysis of the 2019 Indian Election. In *IOT with Smart Systems* (pp. 805–814). Springer. https://doi.org/10.1007/978-981-16-3945-6_79.

Pilar, G. D., Isabel, S. B., Diego, P. M., & José Luis, G. Á. (2022). A novel flexible feature extraction algorithm for Spanish tweet sentiment analysis based on the context of words. *Expert Systems with Applications*, *212*(September 2022). https://doi.org/10.1016/j.eswa.2022.118817.

Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review*, *54*(8), 6149–6200. https://doi.org/10.1007/s10462-021-09970-6.

Pratama, R. P., & Tjahyanto, A. (2021). The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia. *Procedia Computer Science*, *197*(2021), 143–150. https://doi.org/10.1016/j.procs.2021.12.128.

Qian, Y., Zhou, W., Yan, J., Li, W., & Han, L. (2015). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing*, *7*(1), 153–168. https://doi.org/10.3390/rs70100153.

Rashid, A., & Shoaib, U. (2016). Knowledge Discovery in Database using intention mining. *Sci.Int.(Lahore)*, *28*(6), 5145–5151.

Rathika, J., & Soranamageswari, M. (2022). Intensified Gray Wolf Optimization-based Extreme Learning Machine for Sentiment Analysis in Big Data. In P. S. R. Chowdary, J. Anguera, S. C. Satapathy, & V. Bhateja (Eds.), *Evolution in Signal Processing and Telecommunication Networks* (pp. 103–114). Springer Singapore.

Riesener, M., Kuhn, M., Lauf, H., Manoharan, S., & Schuh, G. (2022). Concept for the identification of product innovation potentials by the application of text mining. *Procedia CIRP*, *109*(June), 281–286. https://doi.org/10.1016/j.procir.2022.05.250.

Robbani, H. A. (2016). *Sastrawi 1.0.1*. Https://Pypi.Org/Project/Sastrawi/. https://pypi.org/project/Sastrawi/.

Saputra, P. S. (2021). Perbandingan Algoritma Fuzzy C-Means Dan Algoritma Naive Bayes Dalam Menentukan Keluarga Penerima Manfaat (Kpm) Berdasarkan Status Sosial Ekonomi (Sse) Terendah. *JST (Jurnal Sains Dan Teknologi)*, *10*(1), 1–8. https://doi.org/10.23887/jstundiksha.v10i1.23340.

Starosta, K. (2022). Sentiment Analysis as a New Source of Information. In *Measuring the Impact of Online Media on Consumers, Businesses and Society* (pp. 33–48). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36729-9_4.

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). A Survey on Text Mining Techniques. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, *March*, 1022–1026. https://doi.org/10.1109/ICACCS.2019.8728547.

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes. *Information (Switzerland)*, *12*(5). https://doi.org/10.3390/info12050204.

Wenda, A. (2022). Support Vector Machine Untuk Pengenalan Bentuk Manusia Menggunakan Kumpulan Fitur Yang Dioptimalkan. *JST (Jurnal Sains Dan Teknologi)*, *11*(1), 77–84. https://doi.org/10.23887/jstundiksha.v11i1.44437.

Xue, L., Wang, H., Wang, F., & Ma, H. (2021). Sentiment Analysis of Stock Market Investors and Its Correlation with Stock Price Using Maximum Entropy. In R. Lee (Ed.), *Computer and Information Science 2021---Summer* (pp. 29–44). Springer International Publishing. https://doi.org/10.1007/978-3-030-79474-3_3.

Zulfa, I., & Winarko, E. (2017). Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *11*(2), 187. https://doi.org/10.22146/ijccs.24716.