**BRIDGE : Jurnal publikasi Sistem Informasi dan Telekomunikasi**
**Volume. 2 No. 2 Mei 2024**
e-ISSN:3046-725X, dan p-ISSN : 3046-7268, Hal 176-193
DOI: https://doi.org/10.62951/bridge.v2i2.180

# Implementasi Model Support Vector Machine Dalam Analisa Sentimen Masyarakat Mengenai Kebijakan Penerapan Aplikasi Mypertamina

**Salsabila Dwi Fitri[1], Dewi Lestari[2], Rizqa Raaiqa Bintana[3*], Reni Aryani[4], Mohamad Ilhami[5], Yolla Noverina[6]**
[1-6] Program Studi Sistem Informasi, Universitas Jambi, Indonesia

Alamat: Jl. Jambi-Muara Bulian KM. 15, Mendalo Darat, Kec. Jambi Luar Kota, Kabupaten Muaro Jambi, Jambi
**Korespondensi penulis: rizqa.raaiqa.bintana@unja.ac.id*

***Abstract****. The policy for using the MyPertamina application issued does not rule out the possibility of differences of opinion due to changes in the policy. There are many positive, neutral, and negative responses to the MyPertamina application implementation policy. To see the public's reaction to the MyPertamina application implementation policy, it can be seen through various media, including social media. Twitter is a social network that is widely used by people in Indonesia. The number of Twitter users in Indonesia reached 18.45 million in 2022, making Indonesia the fifth largest Twitter user country in the world. Researchers conducted a sentiment analysis of the search results for tweets containing the keyword "MyPertamina" using the support vector machine algorithm. 382 tweet data were obtained and classified using the support vector machine algorithm. Support vector machine is a supervised learning algorithm for data classification. SVM is very fast and effective in solving text data problems. Text data is suitable for classification with the SVM algorithm because the basic nature of text tends to be high-dimensional. Of the 382 data analyzed, the support vector machine classification using the RBF kernel with parameter C=2 gave the highest accuracy value of 80.51%, precision value of 81%, recall value of 81%, and F1 score value of 80%.*

***Keywords****: classification, sentiment analysis, support vector machine, text mining.*

**Abstrak**. Kebijakan penggunaan aplikasi MyPertamina yang dikeluarkan tidak menutup kemungkinan terjadinya perbedaan pendapat akibat perubahan kebijakan yang terjadi. Banyaknya tanggapan positif, netral, dan negatif terhadap kebijakan implementasi aplikasi MyPertamina. Untuk melihat reaksi masyarakat terhadap kebijakan implementasi aplikasi MyPertamina dapat dilihat melalui berbagai media, termasuk media sosial. Twitter merupakan jejaring sosial yang banyak digunakan oleh masyarakat di Indonesia. Jumlah pengguna Twitter di Indonesia mencapai 18,45 juta pada tahun 2022, menjadikan Indonesia sebagai negara pengguna Twitter terbesar kelima di dunia. Peneliti melakukan analisis sentimen terhadap hasil pencarian tweet yang mengandung kata kunci “MyPertamina” dengan menggunakan algoritma support vector machine. Data tweet sebanyak 382 data diperoleh dan diklasifikasi menggunakan algoritma support vector machine. Support vector machine adalah algoritma supervised learning untuk klasifikasi data. SVM sangat cepat dan efektif dalam menyelesaikan permasalahan data teks. Data teks cocok untuk klasifikasi dengan algoritma SVM karena sifat dasar teks cenderung berdimensi tinggi. Dari 382 data yang dianalisis, klasifikasi support vector machine menggunakan kernel RBF dengan parameter C=2 memberikan nilai akurasi tertinggi sebesar 80,51%, nilai presisi sebesar 81%, nilai recall sebesar 81%, dan nilai skor F1 sebesar 80%.

**Kata kunci**: analisis sentimen, klasifikasi, support vector machine, text mining.

## 1. LATAR BELAKANG

Aplikasi MyPertamina merupakan aplikasi yang diluncurkan oleh PT Pertamina sebagai upaya pembatasan untuk mengawasi penyaluran bahan bakar bersubsidi. Melalui Aplikasi MyPertamina, masyarakat diimbau untuk membuat akun dan mendaftar di Situs MyPertamina. Oleh karena itu, pembelian bahan bakar bersubsidi seperti solar dan Pertalite hanya dapat dilakukan oleh konsumen yang terdaftar di situs resmi MyPertamina, dan tujuan utamanya adalah untuk memastikan penggunaan bahan bakar bersubsidi tepat sasaran. Menurut Anggota Komisi VI DPR Nevi Zuairina, langkah penyaluran subsidi BBM melalui

*Received: April 30, 2024; Accepted: Mei 30, 2024; Published: Mei 31, 2024;*
** Salsabila Dwi Fitri rizqa.raaiqa.bintana@unja.ac.id*

aplikasi MyPertamina untuk pembelian Partalite dan solar tidak efektif dan akan semakin menyulitkan masyarakat. Kebijakan ini dapat menyebabkan masyarakat di pedesaan atau daerah tidak dapat menerima subsidi. Pasalnya, banyak dari mereka yang tidak bisa menggunakan aplikasi MyPertamina, kesulitan mengakses internet, dan terkendala kepemilikan perangkat. Hal ini berarti penerapan kebijakan penggunaan aplikasi MyPertamina ini memiliki pengaruh dan dampak pada masyarakat.

Untuk melihat reaksi masyarakat terhadap penerapan Kebijakan Penggunaan Aplikasi MyPertamina dapat dilihat melalui berbagai sarana dan media, termasuk media sosial. Media sosial merupakan sumber informasi dan media bertukar pendapat serta kehidupan sehari-hari. Ada sejumlah media sosial populer, dan sebagian besar penggunanya adalah Facebook, Instagram, dan Twitter. Twitter merupakan jejaring sosial yang banyak digunakan oleh masyarakat di Indonesia. Menurut survey *We Are Social*, jumlah pengguna Twitter di Indonesia diperkirakan mencapai 18,45 juta pada tahun 2022. Hal ini menjadikan Indonesia sebagai negara pengguna Twitter terbesar kelima di dunia. Jumlah ini mewakili 4,23% dari seluruh pengguna Twitter di seluruh dunia atau 436 juta orang. Untuk mengetahui secara terukur opini dan sentimen masyarakat mengenai penggunaan aplikasi MyPertamina, maka data opini yang digunakan untuk analisis sentimen dalam penelitian ini diperoleh dari tweet yang kemudian diklasifikasi menjadi kelas positif, negatif, atau netral.

Secara umum, ada dua pendekatan untuk melakukan analisis sentimen: pendekatan *supervised learning* dan pendekatan *lexicon-based*. Supervised learning didasarkan pada data pelatihan, sedangkan pendekatan berbasis kamus (lexicon-based) menggunakan kamus sentimen dari kata-kata opini dan membandingkannya dengan data untuk menentukan nilai kata tersebut. Analisis sentimen merupakan model klasifikasi data yang menggunakan pendekatan supervised learning pada pembelajaran mesin (machine learning). Supervised learning adalah jenis algoritma pembelajaran mesin yang digunakan untuk mengklasifikasikan data dan melakukan operasi pembelajaran menggunakan data masukan berlabel. Algoritma supervised learning mencakup Naive Bayes Classifier, Support Vector Machine (SVM), dan Artificial Neural Network (ANN). Melakukan analisis sentimen memerlukan algoritma yang sesuai untuk menghasilkan keluaran yang diinginkan. Salah satu metode klasifikasi dokumen yang populer saat ini adalah metode support vector machine.

Model support vector machine sangat akurat karena dapat menangani model nonlinier yang kompleks. SVM sangat cepat dan efektif dalam menyelesaikan permasalahan data teks. Data teks cocok untuk diklasifikasi menggunakan algoritma SVM. Sifat dasar teks cenderung berdimensi tinggi, karena memiliki banyak fitur yang tidak berhubungan namun cenderung berkorelasi satu sama lain, dan umumnya masuk ke dalam kategori yang terpisah secara linier. Hal ini telah dibuktikan dalam penelitian sebelumnya. Model support vector machine telah terbukti mencapai akurasi yang sangat tinggi saat melakukan analisis sentimen. Dalam penelitian bertajuk "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek", sebuah metode yang menggabungkan model support vector machine dan ekstraksi fitur TF-IDF ditemukan 86% lebih akurat daripada metode naive bayes. Metode ini juga berkolaborasi dengan TF-IDF untuk menganalisis opini masyarakat tentang Gojek di Twitter dengan membaginya menjadi dua kelas: positif dan negatif (Windasari et al., 2017). Selain itu, dalam penelitian sebelumnya, "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film", dilakukan pengujian menggunakan algoritma klasifikasi Naive Bayes (NB), support vector machine, dan artificial neural network (ANN). Diambil kesimpulan bahwa dibandingkan ANN dan NB, support vector machine memberikan hasil terbaik dengan nilai akurasi 81.10% dan nilai AUC = 0.904, dan SVM memberikan akurasi dan nilai AUC tertinggi (Chandani et al., 2015).

Penelitian yang dilakukan ini bertujuan untuk mengetahui tingkat akurasi hasil analisa sentimen masyarakat mengenai kebijakan penerapan aplikasi MyPertamina terhadap data teks yang diolah menggunakan model support vector machine. Sumber data teks berasal dari opini sosial media twitter berbahasa Indonesia.

## 2. KAJIAN TEORITIS

### A. Analisis Sentimen

Analisis sentimen atau *opinion mining* adalah proses memahami, mengekstraksi, dan mengolah data teks secara otomatis untuk memperoleh informasi emosional yang terkandung dalam teks opini (Buntoro, 2017). Analisis sentimen dapat dibedakan berdasarkan sumber datanya. Penelitian analisis sentimen seringkali menggunakan beberapa tingkatan: analisis sentimen tingkat dokumen dan analisis sentimen tingkat kalimat (Pertiwi, 2019). Tugas dasar analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek apakah pendapat

yang dikemukakan dalam dokumen, kalimat atau fitur entitas atau aspek bersifat positif, netral atau negatif (Mesut et all, 2012).

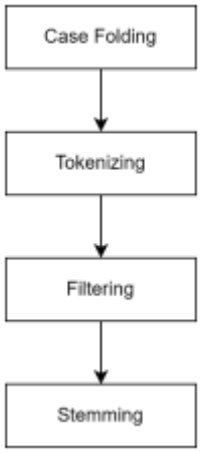
Analisis sentimen dapat diklasifikasikan ke dalam kelas sentimen bersifat positif, negatif dan netral.

1. Sentimen Positif. Menurut Kamus Besar Bahasa Indonesia (KBBI), sentimen positif merupakan reaksi atau sikap yang setuju, sependapat dan meningkatkan nilai seseorang atau sesuatu.
2. Sentimen Negatif. Menurut KBBI, sentimen negatif merupakan reaksi atau sikap yang menurunkan nilai seseorang sehingga menyebabkan penyurutan terhadap sesuatu dan membentuk tren down. Umumnya kalimat sentimen negatif ditandai dengan penggunaan kata negasi.
3. Sentimen Netral. Menurut KBBI, sentimen netral berarti tidak berpihak. Kalimat bersentimen netral merupakan ekspresi kalimat yang tidak bersifat positif maupun negatif.

**B. Text Mining**

Text mining merupakan analisis teks yang sumber datanya biasanya berasal dari dokumen dan tujuannya adalah untuk mencari kata-kata yang menggambarkan isi dokumen, sehingga memungkinkan analisis keterhubungan, keterkaitan, dan kelas antar dokumen (Lesmeister, 2015). Menurut Feldman & Sanger (2007), text mining adalah teknik yang dapat digunakan untuk melakukan klasifikasi, dan text mining adalah jenis data mining yang berupaya menemukan pola menarik dalam data teks dalam jumlah besar. *Text mining* memiliki tujuan dan menggunakan proses yang sama dengan *data mining*, namun memiliki masukan yang berbeda.

Masukan untuk *text mining* adalah adalah data yang tidak (atau kurang) terstruktur, seperti dokumen Word, PDF, kutipan teks, dan lain-lain, sedangkan masukan untuk *data mining* adalah data yang terstruktur. Cara yang digunakan dalam mempelajari struktur data teks adalah dengan terlebih dahulu menentukan fitur yang mewakili setiap kata untuk pada dokumen. Sebelum menentukan fitur-fitur yang mewakili, diperlukan tahap *preprocessing* yang dilakukan secara umum dalam *text mining* pada dokumen, yaitu *case folding, tokenizing, filtering, dan stemming* (Mooney, 2006), seperti terlihat pada Gambar 1.




Sumber: (Mooney, 2006)

**Gambar 1. Proses *text mining***

Tahapan Text Mining:

1. Cleansing: proses penghapusan karakter non-alfabetis untuk mengurangi *noise*. Karakter yang dihapus berupa tanda baca, simbol-simbol seperti tanda '@' untuk nama pengguna, hashtag (#), emoticon dan url dari situs web.
2. Case Folding: proses mengubah semua karakter alfabet yang sudah dibersihkan sebelumnya menjadi huruf kecil (*lower case*).
3. Tokenizing: merupakan proses pemisahan kata-kata dari kalimat.
4. Filtering: proses menghapus kata yang sering muncul secara umum dan kurang relevan serta mengubah kata berimbuhan dari setiap kata yang sudah disaring menjadi kata dasar.
5. Stemming: proses mencari *root* (dasar) kata dari tiap kata hasil *filtering*.

## C. Machine Learning

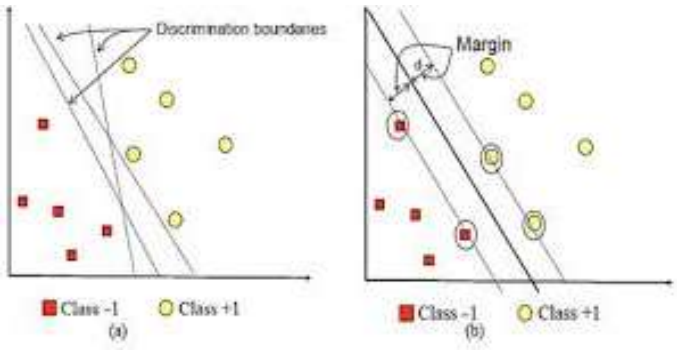
*Machine learning* merupakan sub dari bidang keilmuan kecerdasan buatan (a*rtificial intelligence*). *Machine learning* dapat diartikan sebagai aplikasi komputer dan algoritma matematika yang diadopsi dengan cara pembelajaran yang berasal dari data dan menghasilkan prediksi di masa yang akan datang (Goldberg & Holland, 1988). Proses pembelajaran yang dimaksud adalah suatu usaha dalam memperoleh kecerdasan yang melalui dua tahap antara lain, latihan (*training*) dan pengujian (*testing*). Bidang *machine learning* berkaitan dengan pertanyaan tentang bagaimana membangun program komputer agar meningkat secara otomatis dengan berdasar dari pengalaman. Penelitian terkini mengungkapkan bahwa *machine learning* terbagi menjadi tiga kategori, yaitu

*supervised learning*, *unsupervised learning*, dan *reinforcement learning* (Somvanshi & Chavan, 2016).

Teknik yang digunakan oleh supervised learning adalah metode klasifikasi di mana kumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang tidak dikenal. Sedangkan teknik unsupervised learning dikenal dengan *cluster* dikarenakan tidak ada kebutuhan untuk pemberian label dalam kumpulan data dan hasilnya tidak mengidentifikasi contoh di kelas yang telah ditentukan (Thupae et al., 2018). Teknik reinforcement learning berada antara supervised learning dan unsupervised learning, teknik ini bekerja dalam lingkungan yang dinamis di mana konsepnya harus menyelesaikan tujuan tanpa adanya pemberitahuan dari komputer secara eksplisit jika tujuan tersebut telah tercapai (Das & Nene, 2017). Penelitian analisis sentimen yang dilakukan akan menggunakan teknik supervised learning karena data yang digunakan terlebih dahulu diberi label untuk dapat mengklasifikasikan kelas yang tidak dikenal.

## D. Support Vector Machine

Support Vector Machine (SVM) diperkenalkan oleh Vapnik pada tahun 1992 sebagai teknik klasifikasi yang efisien untuk masalah nonlinier. Support vector machine juga disebut sebagai teknik pembelajaran mesin tercanggih, setelah pembelajaran mesin sebelumnya dikenal sebagai Neural Network (NN). SVM dan NN telah berhasil digunakan untuk pengenalan pola. Pembelajaran terjadi atas pasangan data masukan dan keluaran berupa tujuan yang diinginkan. Konsep SVM secara sederhana dapat dijelaskan sebagai upaya untuk menemukan hyperplane optimal yang bertindak sebagai pemisah antara dua kelas dalam ruang input. SVM berupaya mencari fungsi pemisah (hyperplane) dengan memaksimalkan jarak antar kelas. Dengan cara ini, SVM dapat menjamin kemampuan generalisasi yang tinggi untuk data yang akan datang (Suyanto, 2017).




Sumber: (Pisner & Schnyer, 2019)

**Gambar 2. Ilustrasi SVM menemukan *hyperplane* terbaik yang memisahkan dua kelas -1 dan +1**

Gambar 2 mengilustrasikan dua kelas data yang terpisah. Garis merah tebal merupakan hyperplane yang memisahkan kedua titik tersebut, sehingga titik data pada salah satu sisinya diberi label kelas negatif -1 dan label kelas positif +1. Titik data (vektor) yang paling dekat dengan hyperplane, disebut vektor pendukung, pada garis merah kecil yang memotong hyperplane, mempunyai pengaruh paling besar. Jarak antara hyperplane dan support vector disebut margin. Beberapa *kernel* yang umum dipakai pada SVM adalah:

1. *Polynomial. Kernel trick polynomial* cocok digunakan untuk menyelesaikan masalah klasifikasi, dimana dataset pelatihan sudah normal.
2. *Radial Basis Function* (RBF) atau *Gaussian*. *Kernel trick radial basis function* atau *gaussian* merupakan kernel yang paling banyak digunakan untuk menyelesaikan masalah klasifikasi untuk dataset yang tidak terpisah secara linear, dikarenakan akurasi pelatihan dan akurasi prediksi yang sangat baik pada kernel ini.

Meskipun waktu pelatihan SVM sebagian besar lambat, model ini sangat akurat karena dapat menangani model nonlinier yang kompleks. SVM memiliki kecenderungan overlifting yang lebih rendah dibandingkan model lainnya. SVM sangat cepat dan efektif dalam menyelesaikan permasalahan data teks. Data teks cocok untuk klasifikasi menggunakan algoritma SVM. Hal ini disebabkan karena sifat-sifat dasar teks cenderung berdimensi besar, dengan beberapa fitur yang tidak berkaitan namun cenderung berkorelasi satu sama lain, dan umumnya dikelompokkan ke dalam kategori-kategori yang terpisah secara linier (Aggarwal & Zhai, 2012).

## E. Pembobotan Kata TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan salah satu metode yang biasa dipakai dalam pembobotan sebuah kata di dalam sistem pencarian informasi (Aizawa, 2003). Term adalah kata, frasa, atau unit pengindeksan lainnya dalam suatu dokumen yang dapat digunakan untuk menentukan konteks dokumen tersebut. Karena setiap kata dalam dokumen memiliki arti yang berbeda, setiap kata diberi indeks, atau bobot kata (Zafikri, 2010). Menurut Zafikri (2008), pembobotan kata sangat dipengaruhi oleh faktor-faktor berikut:

1. *Term Frequency* (TF). *Term Frequency* (TF) merupakan faktor yang menentukan bobot suatu kata dalam suatu dokumen berdasarkan seberapa sering kata tersebut

muncul dalam dokumen tersebut. Saat membobotkan kata, nilai frekuensi kata (term frequency) diperhitungkan. Semakin sering suatu kata muncul dalam suatu dokumen (semakin tinggi tf), maka semakin besar bobotnya dalam dokumen tersebut dan semakin tinggi pula nilai relevansinya.

2. *Inverse Document Frequency* (IDF). *Inverse Document Frequency* (IDF) mengurangi dominasi kata yang sering muncul dalam dokumen berbeda. Hal ini diperlukan karena kata-kata yang sering muncul dalam dokumen yang berbeda dianggap kata-kata umum dan nilainya tidak penting. Sebaliknya, pembobotan harus mempertimbangkan faktor jarangnya kata tersebut muncul dalam kumpulan dokumen. Menurut Mandala (Witten, 1999), kata-kata yang muncul dalam lebih sedikit dokumen harus dianggap lebih penting dibandingkan kata-kata yang muncul dalam banyak dokumen. Pembobotannya memperhitungkan kebalikan dari frekuensi dokumen yang mengandung kata tersebut (inverse document frequency).

**F. Evaluasi Hasil**

Metode yang digunakan untuk evaluasi adalah *confusion matrix*. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya.

**Tabel 1. Confusion matrix 3x3**

| | | Kelas Prediksi | | |
|---|---|---|---|---|
| | | Positive | Negative | Neutral |
| Kelas Sebenarnya | Positive | True Positive (TP) | True Negative1 (FNg1) | False Neutral1 (FNt1) |
| | Negative | False Positive1 (FP1) | True Negative (TNg) | False Neutral2 (FNt2) |
| | Neutral | False Positive2 (FP2) | False Negative2 (FNg2) | True Neutral (TNt) |

Untuk pengukuran performa klasifikasi cara yang digunakan adalah menghitung akurasi, *precission*, *recall* dan *f1-Score*. Akurasi merupakan persentase dari total sentimen yang benar dikenali. Perhitungan akurasi dilakukan dengan cara membagi jumlah data sentimen yang benar dengan total data dan data uji. Untuk menghitung nilai akurasi digunakan Persamaan 1. *Precision* merupakan perbandingan jumlah data relevan yang ditemukan terhadap jumlah data yang ditemukan. Untuk menghitung nilai *precision* digunakan Persamaan 2. *Recall* merupakan perbandingan jumlah materi relevan yang ditemukan terhadap jumlah materi yang relevan. Untuk menghitung nilai *recall*

digunakan Persamaan 3. *F1-Score* merupakan parameter tunggal ukuran keberhasilan *retrieval* yang menggabungkan *recall* dan *precision*. Untuk menghitung nilai *F1-Score* digunakan Persamaan 4.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} x\ 100\% \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = 2x\frac{PrecisionxRecall}{Precision+Recall} \quad (4)$$

## G. Penelitian Terdahulu

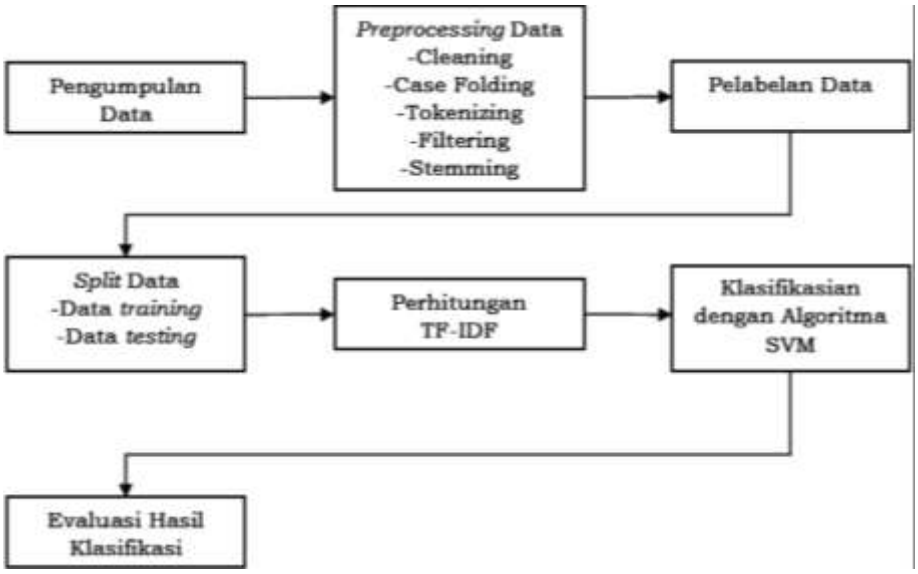
Penelitian terdahulu dijadikan sebagai acuan dalam melakukan penelitian ini. Fokus penelitian yaitu bagaimana sebuah algoritma support vector machine dapat dengan baik dan menghasilkan nilai akurasi yang tinggi dalam mengklasifikasikan sentimen analisis twitter dengan kata kunci "MyPertamina" ke dalam kelas positif, netral dan negatif yang belum pernah dilakukan oleh peneliti sebelumnya. Penelitian yang berjudul Penerapan Metode *Support Vector Machine* Untuk Analisis Sentimen Pengguna Twitter melakukan analisa terhadap tanggapan pengguna Bukalapak di media sosial Twitter dan mengkategorikan opini-opini dengan menggunakan metode support vector machine. Dari hasil klasifikasi opini tersebut diperoleh tingkat akurasi tertinggi sebesar 93% (Alhaq et al., 2021). Penelitian lainnya dengan judul Sentiment Analysis on Twitter Posts: An analysis of Positive or Negative Opinion on GoJek mengusulkan sistem yang dapat mendeteksi sentimen publik berdasarkan tweet pengguna Twitter tentang layanan transportasi online terutama GoJek menggunakan metode support vector machine. Hasil dari pengujian yang dilakukan ditemukan bahwa metode support vector machine yang dikolaborasikan dengan ekstraksi fitur TF-IDF menghasilkan akurasi sebesar 86% lebih unggul dibandingkan dengan metode Naïve Bayes yang juga dikolaborasikan dengan TF-IDF ketika melakukan analisis opini masyarakat mengenai Gojek pada Twitter yang dibagi menjadi dua kelas yaitu positif dan negatif (Windasari et al., 2017).

Penelitian dengan judul Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film melakukan pengujian analisis sentimen review film dengan menggunakan data yang berasal dari situs IMDb yang terdiri dari 1000 review film, berisi 500 review positif dan 500 review negatif. Dari penelitian yang dilakukan dengan menggunakan algoritma klasifikasi *Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *Artificial Neural Network* (ANN) dapat diperoleh kesimpulan bahwa *Support Vector Machine* (SVM) memiliki hasil terbaik dengan nilai

akurasi sebesar 81,10 % dan nilai AUC = 0,904. SVM menghasilkan nilai *accuracy* dan AUC terbaik dibanding ANN dan NB (Chandani & Wahono, 2015). Dan penelitian lainnya yang berjudul Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi *Support Vector Machine* (SVM), sentimen analisis terhadap respon pengguna aplikasi Gojek di media sosial Twitter. Dari 6.917 tweets yang didapat, dipilih 1.500 tweets terbaru yang digunakan pada penelitian ini. Model klasifikasi support vector machine menggunakan fungsi kernel linear dan RBF, serta menggunakan evaluasi model 10-Fold Cross Validation. Hasil klasifikasi sentimen model kernel linear dan kernel RBF dari hasil pelabelan data secara manual dan sentiment scoring memiliki akurasi keseluruhan tertinggi yang sama yaitu 79,19% (Fitriyah et al., 2020).

## 3. METODE PENELITIAN

Alur kerja atau langkah-langkah penelitian ditunjukkan pada Gambar 3.




**Gambar 3. Alur kerja penelitian**

Tahapan penelitian yang akan dilakukan yaitu:

1. Pengumpulan data. Data yang digunakan dalam penelitian ini adalah tanggapan masyarakat berupa tweet dari pengguna media sosial Twitter mengenai kebijakan penerapan penggunaan aplikasi MyPertamina. Data teks Twitter dikumpulkan sebagai data berformat Json menggunakan *library snscrape* dan mengonversinya menjadi file csv. Data yang dikumpulkan dengan menggunakan kata kunci "MyPertamina" pada kategori tweets berbahasa Indonesia adalah sebanyak 451 tweets.
2. *Preprocessing* data atau tahap persiapan data sebelum mengolah data yang dibagi menjadi lima tahap:

- *Cleaning*, yaitu proses menambahkan data yang hilang dan mengoreksi, memperbaiki, atau menghapus data yang salah atau tidak relevan dari dataset. Termasuk juga menghapus username pada *tweets.*
- *Case folding*, di mana semua huruf besar diubah menjadi huruf kecil.
- *Tokenizing*, tahapan ini memecah string berdasarkan setiap kata yang menyusun teks tersebut, menghapus URL, @mention, hashtag, serta menghilangkan delimiter titik (.), koma (,), spasi, angka, dan tanda baca dari dalam dokumen tweet.
- *Filtering* (disebut juga dengan *stopword removing*), yaitu menghapus kata-kata yang tidak bermakna dan tidak berdampak pada analisis sentimen.
- *Stemming*, memotong imbuhan kata sehingga kata hasil filtering menjadi kata dasar.

3. Pelabelan pada dataset dilakukan manual dengan bantuan ahli bahasa. Pemberian label dibagi menjadi 3 kelas yaitu kelas positif, negatif, dan netral.
4. Split data, yaitu pemisahan dataset menjadi 2 bagian (dataset *training* dan dataset *testing*).
5. Perhitungan TF-IDF untuk menentukan kata atau *term* yang sering muncul.
6. Pengklasifikasian data dengan model support vector machine. Setelah melewati proses pembobotan dengan TF-IDF, dilakukan pelatihan pada data *training* untuk tiap kernel SVM. Pelatihan data *training* tersebut akan menghasilkan model pembelajaran, lalu model tersebut akan diuji dengan data *testing*.
7. Pengukuran tingkat keakuratan dari hasil pengklasifikasian data. Proses evaluasi dilakukan dengan menggunakan confusion matrix untuk menguji seberapa baik performa model algoritma klasifikasi yang dibangun. Confusion matrix memberikan informasi komperatif antara hasil klasifikasi yang sebenarnya dengan prediksi. Pada penelitian ini diterapkan klasifikasi tiga kelas yang terdiri dari negatif, netral, dan positif. Gunakan Persamaan 1, 2, 3, dan 4 untuk mencari nilai akurasi, presisi, recall, dan f1-Score untuk tiap kelas tersebut.

## 4. HASIL DAN PEMBAHASAN

### A. Pengumpulan Data

Tahap awal yang dilakukan adalah mengumpulkan data tweet berbahasa Indonesia dengan pencarian kata kunci “MyPertamina”. Terlebih dahulu dilakukan *Import Libraries* yang disediakan oleh Python. *Library* yang digunakan dalam pengumpulan data yang digunakan adalah *library snscrape*. Setelah mengimport *library*

*snscrape* selanjutnya dilakukan pengambilan data dengan menggunakan *query TwitterSearchScraper*. Dengan memasukkan kata kunci 'Mypertamina' dengan rentang waktu pengambilan dari 01-01-2023 sampai 31-05-2023. Data yang sudah terkumpul tersebut dimasukkan ke dalam dataframe lalu disimpan ke dalam format xlsx, diperoleh data sebanyak 451 tweet. Data yang tersimpan terdiri dari tanggal pembuatan tweet, id tweet, username twitter yang memposting tweet serta tweet yang diposting.

**B. Preprocessing Data**

Tahapan *preprocessing* data perlu dilakukan karena beberapa kalimat tweet yang didapatkan tidak sepenuhnya menggunakan kata baku dan menggunakan bahasa indonesia yang baik. *Preprocessing* dilakukan menggunakan bantuan *library* pada bahasa pemrograman Python. Preprocessing data dilakukan dengan tahap *cleaning, case folding, tokenizing, filtering,* dan *stemming* sehingga menghasilkan data bersih dan siap untuk lanjut pada proses berikutnya.

1. Cleaning

   Tahapan awal dari *preprocessing* ini untuk melakukan penghapusan *username* dan data yang duplikat menggunakan fungsi *drop_duplicates*. Dari 451 data, tersisa 382 data dari proses penghapusan data duplikat. Pada kolom teks yang berisi *tweet*, juga dilakukan penghapusan atribut dan simbol pada kolom tersebut dengan fungsi dari *library re*.

2. Case Folding

   Pada tahapan ini dilakukan proses mengubah data tweet menjadi lowercase menggunakan fungsi lower() yang sudah tersedia pada python. Selain itu, juga dilakukan normalisasi data unicode, menghapus angka, tanda baca, dan white space dengan fungsi re.sub() dari library re yang sebelumnya sudah diimport untuk melakukan tahapan Regular Expression (regex) atau deretan karakter yang digunakan untuk pencarian teks dengan menggunakan pola (*pattern*). Dengan menggunakan library regex dapat memudahkan dalam mencari string tertentu dari teks yang banyak.

3. Tokenizing

   Tokenizing dalam penelitian ini merupakan tahapan dalam memecah string atau input terhadap suatu teks yang telah melewati tahap case folding berdasarkan tiap kata yang menyusunnya.

4. Filtering

Proses membuang kata yang tidak memiliki arti. Proses *filtering* disebut dengan *Stopword Removal*. Pada tahap ini menggunakan nlp. NLP (*Natural Language Processing*) adalah *Library* yang disediakan oleh oleh *Python* untuk membangun program analisis teks. Pada library *nlp_id* ini terdapat *stopword* indonesia. Selain *list stopword* indonesia yang disediakan oleh *library nlp_id*, ditambahkan *list* kata yang tidak dibutuhkan dalam analisis sentimen dengan cara menambahkan secara langsung kata pada *remove_stopwords2* agar dapat dihapus oleh sistem.

5. Stemming

   Stemming adalah tahap mencari *root* (dasar) kata dari tiap kata hasil *filtering* dengan menghapus kata imbuhan di depan maupun imbuhan di belakang kata. Tahap stemming dilakukan dengan menggunakan bantuan *library* pada bahasa pemrograman *Python3* yang bernama *Sastrawi.* Hasil Proses *stemming* akan menghapus kata yang memiliki imbuhan pada awalan, akhiran maupun sisipan kata menjadi bentuk kata dasar.

6. Buang kata kurang dari 2 huruf

   Karena masih ada beberapa data yang mengandung huruf tunggal yang mana ini tidak memiliki arti, maka dilakukan penghapusan huruf tersebut.

**C. Pemberian Label**

Dari data twitter yang telah dikumpulkan yaitu sejumlah 382 data, kemudian dikelompokkan menjadi tiga kelas yaitu positif, negatif, dan netral. Pengelompokan menjadi tiga kelas tersebut dilakukan dengan bantuan ahli Bahasa Indonesia. Contoh hasil pelabelan dapat dilihat pada Tabel 2. Hasil pelabelan tersebut adalah 151 data tweet positif, 137 data tweet negatif dan 94 sebagai data tweet netral.

**Tabel 2. Contoh hasil pelabelan data tweet**

| Data Positif | Data Negatif | Data Netral |
|---|---|---|
| terbukti banget ya gais dengan mypertamina semuanya jadi tepat sasaran | intinya aplikasi mypertamina hanya gertak udah susah daftarnya udah bisa digunakan eh malah gak berguna akhirnya pertamak turun biar stoknya gk menguap | min saya sudah mengisi data di mypertamina saya sudah menunggu selama hari tapi kok blm ada balasan verifikasinya |
| beli bensin pke mypertamina dapet untung mahh enak sihh jdi lebih hemat jdinya | bikin kebijakan pembatasan pembelian solar tapi belinya harus bikin barcode dulu di mypertamina dan prosesnya panjang wong tani yang hpan cuma bisa sms telpon ya ndramus kabeh | apakah benar mypertamina melakukan survey melalui rekanan survey dengan menelpon ke pelanggan |

| | | |
|---|---|---|
| nahh bner nihh jdi klo diterapin gini yg tajir bisa auto ketar ketir krna mypertamina bisa jdi pengendali bbm subsidi tepat sasaran | ini mypertamina gangguan lama banget dari tadi pagi ga bisa kebukak appnya mypertamina pertamina spbu bumn | mampukah aturan qr code mypertamina tekan kasus penimbunan minyak indonesia |

**Pembobotan TF-IDF**

Sebelum masuk ke tahap pembobotan TF-IDF, perlu dilakukannya pembagian data menjadi data *training* dan data *testing*. Namun, sebelum itu data yang masih berbentuk token harus diubah terlebih dulu menjadi kalimat biasa. Selanjutnya membagi data dengan terlebih dahulu mengimport *train_test_split* dari *library scikit-learn*, data dibagi sebesar 80% untuk data *training* dan 20% untuk data *testing.* Selanjutnya dilakukan *import TfidfVectorizer* dari *library scikit-learn* yang di dalamnya sudah terdapat *CountVectorize*r dan *TfidfTransformer*. TF dihitung pada *class CountVectorizer*, IDF dihitung pada *class TfidfTransformer.* Untuk menghitung TF-IDF, maka *TfidfVectorizer* disimpan pada variabel *Tfidf_vect* lalu dilakukan proses perhitungan untuk masing-masing data *training* dan data *testing* lalu hasilnya disimpan pada variabel *Train_X_Tfidf* dan *Test_X_Tfidf.*

**D. Klasifikasi dengan Algoritma Support Vector Machine**

Setelah melalui proses pembobotan kata selanjutnya dibuat model yang akan digunakan untuk melakukan klasifikasi pada data *training.* Pemrosesan menggunakan support vector machine ini berfokus pada penggunaan tiga fungsi kernel yang umum digunakan dalam SVM, yaitu kernel linear, kernel RBF, dan kernel *Polynomial*. Proses ini dilakukan dengan menggunakan bantuan *library* pada bahasa pemograman Python3 yang bernama scikit-learn untuk proses klasifikasi.

Pada library *scikit-learn* diimport SVC untuk melakukan proses klasifikasi pada *data training*. Proses klasifikasi dilakukan dengan perhitungan probabilitas antar kalimat terhadap setiap kelas agar dapat menghasilkan dengan jelas prediksi data yang dimasukan. Data *Training* yang digunakan adalah sebanyak 305 data. Klasifikasi SVM ini menggunakan kernel linear, RBF, dan *polynomial*. Selanjutnya model hasil klasifikasi dari *data training* pada tiap kernel tersebut akan disimpan pada variabel *predictedsvm1*, *predictedsvm2*, dan *predictedsvm3* untuk selanjutnya dilakukan tes dengan data *testing.* Data *Testing* yang digunakan adalah sebanyak 77 data.

**E. Pengujian Model Klasifikasi Support Vector Machine**

Untuk mengetahui performa dari Algoritma support vector machine ini, maka dilakukan pengujian terhadap model yang telah dibuat. Hasil klasifikasi akan

divisualisasi dalam bentuk *confusion matrix. Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Pada bagian ini, dilakukan optimasi parameter SVM menggunakan Grid Search. Dari hasil pencarian *grid* dengan kombinasi dari parameter C, degree, dan tiga kernel didapatkan hasil kernel RBF dengan C = 2 menghasilkan f1_score paling maksimal yaitu sebesar 80,51% dan ini lebih besar jika dibandingkan dengan menggunakan kernel linear yang memiliki tingkat akurasi sebesar 79,22%. Maka dari itu, pada penelitian ini selanjutnya model klasifikasi yang digunakan adalah model klasifikasi SVM kernel RBF dengan parameter C = 2.

## F. Evaluasi Model

Setelah pengujian model selesai dilakukan langkah selanjutnya yang dilakukan adalah melakukan evaluasi model. Evaluasi model bertujuan untuk mengevaluasi performa model *machine learning* yang telah dibuat. Evaluasi model mencakup perhitungan akurasi, presisi, *recall*, dan *f1 score.*

**Tabel 3. Perbandingan nilai akurasi, presisi, recall, f1-score evaluasi model**

| Kernel | Akurasi | Presisi | Recall | F1-Score |
|---|---|---|---|---|
| Linear | 0,79 | 0,79 | 0,79 | 0,78 |
| RBF | 0,75 | 0,77 | 0,75 | 0,73 |
| Polynomial | 0,70 | 0,71 | 0,70 | 0,66 |
| RBF C = 2 | 0,80 | 0,80 | 0,80 | 0,80 |

Tabel 3 merupakan perbandingan nilai presisi, recall dan *F1-Score* pada setiap kernel yang digunakan yaitu kernel linear, RBF, dan polynomial, dan RBF C=2. Berdasarkan tabel perbandingan hasil klasifikasi sentimen menggunakan algoritma *Support Vector Machine* tersebut, kernel RBF dengan parameter C = 2 menghasilkan nilai akurasi tertinggi sebesar 80%, presisi sebesar 80%, *recall* sebesar 80%, dan *F1 Score* sebesar 80%.

**Tabel 4. Confusion matrix kernel RBF C=2**

| | | Kelas Prediksi | | |
|---|---|---|---|---|
| | | Negatif | Netral | Positif |
| Kelas Sebenarnya | Negatif | 24 | 3 | 2 |
| | Netral | 7 | 10 | 1 |
| | Positif | 2 | 0 | 28 |

Tabel 15 menunjukkan c*onfusion matrix* SVM RBF C=2 berupa matriks berukuran 3×3 yang mewakili kelas klasifikasi positif, netral, dan negatif. C*onfusion matrix* menjelaskan bahwa model dengan tepat mengklasifikasikan 24 data sebagai negatif, 10 sebagai netral, dan 28 sebagai positif. Tingkat keberhasilan sistem dalam

menemukan sebuah informasi dalam penelitian ini sebesar 80.51% ketika menggunakan SVM kernel RBF C=2.

**Tabel 5. Nilai presisi, recall, f1-score tiap kelas, kernel RBF C=2**

| Kelas | Presisi | Recall | F1-Score |
|---|---|---|---|
| Negatif | 72,72% | 82,75% | 77,41% |
| Netral | 76,92% | 55,55% | 64,51% |
| Positif | 90,32% | 93,33% | 91,80% |

Hasil dari evaluasi model pada kernel RBF C=2 dapat dilihat bahwa nilai presisi pada ketiga kelas memiliki tingkat kemampuan yang tinggi dalam mencari ketepatan antara informasi yang diminta. Nilai presisi untuk kelas negatif sebesar 73%, untuk netral 77%, dan untuk positif 90%. Angka ini berarti proporsi label prediksi yang benar terhadap total prediksi sangat tinggi untuk semua kelas. Sedangkan tingkat keberhasilan recall dalam pencarian informasi sebesar 83% untuk kelas negatif, 93% untuk kelas positif, dan 56% untuk kelas netral. Artinya sistem yang memunculkan informasi positif dan negatif dalam sebuah dokumen memiliki tingkat keberhasilan yang lebih tinggi dibandingkan ketika sistem memunculkan informasi yang bernilai netral. Nilai rata-rata presisi sebesar 81%, nilai recall sebesar 81%, dan nilai f1-score sebesar 80%.

## 5. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian yang dilakukan terhadap algoritma support vector machine, diperoleh beberapa kesimpulan. Pada penelitian ini, algoritma support vector machine menghasilkan nilai akurasi sebesar 80,51% melalui RBF kernel C=2, terbukti merupakan algoritma yang akurat. Hasil analisis sentimen Twitter dengan menggunakan kata kunci MyPertamina pada penelitian ini memiliki nilai presisi 81%, recall 81%, dan nilai f1-score 80%.

Untuk pekerjaan penelitian selanjutnya, dapat membandingkan hasil uji beberapa model klasifikasi dengan data teks bahasa indonesia untuk mencari algoritma klasifikasi terbaik. Untuk kemudian dibuatkan antarmuka dari proses pengujian model, visualisasi dari performa model, dan visualisasi hasil analisis sentimen sehingga memudahkan publik untuk melihat kecenderungan hasil analisa sentimen dari sekumpulan data opinion tersebut.

## 6. UCAPAN TERIMA KASIH

Artikel ini ditulis berdasarkan dari hasil penelitian skripsi oleh Salsabila Dwi Fitri, Program Studi Sistem Informasi Universitas Jambi. Melalui penerbitan artikel ini, diharapkan dapat membantu pembaca lainnya dalam pemahaman tentang tahapan pekerjaan *text mining* dan model *support vector machine*.

## 7. DAFTAR REFERENSI

Abdillah, G., Putra, F. A., Renaldi, F., & ... (2016). Penerapan data mining pemakaian air pelanggan untuk menentukan klasifikasi potensi pemakaian air pelanggan baru di PDAM Tirta Raharja. Seminar Nasional Informatika, 2016(Sentika), 18–19. https://fti.uajy.ac.id/sentika/publikasi/makalah/2016/43.pdf

Aizawa, A. (2003). An information-theoretic perspective of TF-IDF measures. Information Retrieval, 39, 45–65.

Alhaq, Z., Mustopa, A., Mulyatun, S., & Santoso, J. D. (2021). Penerapan metode support vector machine untuk analisis sentimen pengguna Twitter. Journal of Information System Management (JOISM), 3(2), 44–49. https://doi.org/10.24076/joism.2021v3i2.558

Anonim. (2022). Kamus besar bahasa Indonesia. https://kbbi.kemdikbud.go.id/, 9 November 2022

Azhar, Y. (2018). Metode lexicon-learning based untuk identifikasi tweet opini berbahasa Indonesia. Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), 6(3), 237. https://doi.org/10.23887/janapati.v6i3.11739

Budi Santosa, B. (2007). Data mining: Teknik pemanfaatan data untuk keperluan bisnis. Garah Ilmu.

Buntoro, G. A. (2017). Analisis sentimen calon gubernur DKI Jakarta 2017 di Twitter. Jurnal Informatika, 2(1), 32–41.

Chandani, V., & Wahono, R. S. (2015). Komparasi algoritma klasifikasi machine learning dan feature selection pada analisis sentimen review film. Journal of Intelligent Systems, 1(1), 55–59.

Das, S., & Nene, M. J. (2017). A survey on types of machine learning techniques in intrusion prevention systems. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2296–2299). https://doi.org/10.1109/WiSPNET.2017.8300169

Feldman, R., & Sanger, J. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. Cambridge University Press.

Fitriyah, N., Warsito, B., & Maruddani, D. A. I. (2020). Analisis sentimen Gojek pada media sosial Twitter dengan klasifikasi support vector machine (SVM). Jurnal Gaussian, 9(3), 376–390. https://doi.org/10.14710/j.gauss.v9i3.28932

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. Machine Learning, 3(2), 95–99.

Gunadi, G., & Sensuse, D. I. (2012). Penerapan metode data mining market basket analysis terhadap data penjualan produk buku dengan menggunakan algoritma apriori dan frequent pattern growth (FP-Growth). Telematika, 4(1), 118–132.

Koto, F., & Rahmaningtyas, G. Y. (2018). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017 (pp. 391–394). https://doi.org/10.1109/IALP.2017.8300625

Lesmeister, C. (2015). Mastering machine learning with R: Master machine learning techniques with R to deliver insights for complex projects. Packt Publishing.

Mahbubah, L. D., & Zuliarso, E. (2019). Analisa sentimen Twitter pada pilpres 2019 menggunakan algoritma Naive Bayes. Sintak, 194–195. https://www.unisbank.ac.id/ojs/index.php/sintak/article/view/7585

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

Mooney, R., & Wong, Y. W. (2006). Learning for semantic parsing with statistical machine translation. In Proceedings of The Human Language Technology Conference of the NAACL (pp. 439–446).

Muhammadi, R. H., Laksana, T. G., & Arifa, A. B. (2022). Combination of support vector machine and lexicon-based algorithm in Twitter sentiment analysis. Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika, 8(1), 59–71. https://doi.org/10.23917/khif.v8i1.15213

Nandini, R. A., Sari, Y. A., & Adikara, P. P. (2019). Analisis sentimen impor beras 2018 pada Twitter menggunakan metode support vector machine dan pembobotan jumlah retweet. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 3(4), 3396–3406.

Pisner, D. A., & Schnyer, D. M. (2019). Support vector machine. In Machine Learning: Methods and Applications to Brain Disorders (pp.