

ANALISIS SENTIMEN KEUANGAN (DATA FIQA AND FINANCIAL PHRASEBANK) MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE

Rasiban ¹, Julinar Sari Hutagalung ^{2*}

^{1,2*} Program Studi Sistem Informasi, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia.

Email: julinarhutagalung02@gmail.com ^{2*}

Histori Artikel:

Dikirim 26 Juli 2023; *Diterima dalam bentuk revisi* 3 Agustus 2023; *Diterima* 8 Agustus 2023; *Diterbitkan* 10 September 2023. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMIK Indonesia Banda Aceh.

Abstrak

Keuangan merupakan sektor yang sangat vital pada suatu perusahaan dan lembaga karena memiliki peran strategis yang sangat penting dalam menciptakan lingkungan yang kondusif, terutama untuk peningkatan perekonomian nasional. Melalui gabungan kumpulan data teks FiQA dan Financial PhareBank, dilakukan analisis sentimen positif, negatif dan netral terkait di bidang keuangan yang dapat dijadikan pertimbangan untuk membuat suatu kebijakan pada sektor atau konteks keuangan dalam mencapai peran strategis tersebut. Penerapan analisis sentimen menggunakan hyperparameter tuning pada algoritma logistic regression dan support vector machine, dengan pembobotan TF-IDF dan Smote pada data latih. Hasil model terbaik accuracy sebesar 70,70% pada algoritma Support Vector Machine pada saat training model menggunakan data latih yang tidak dilakukan imbalance class Smote.

Kata Kunci: Analisis Sentimen; Logistic Regression; Support Vector Machine; TF-IDF.

Abstract

Finance is a very vital sector in a Company and institution because it has a very important strategic role in creating a conducive environment, especially for the improvement of the national economy. Through a combination of FiQA and Financial PhareBank text datasets, an analysis of positive, negative and neutral sentiments related to finance is carried out that can be taken into consideration to make a policy in the financial sector or context in achieving this strategic role. Application of sentiment analysis using hyperparameter tuning in Logistic Regression and Support Vector Machines algorithms, with TF-IDF and Smote weighting on training data. The best model results of 70.70% accuracy on the Support Vector Machine algorithm during model training using training data that is not done Smote class imbalance.

Kata Kunci: Sentiment Analysis; Logistic Regression; Support Vector Machine; TF-IDF

1. Pendahuluan

Lembaga perbankan memiliki peran strategis yang sangat penting dalam menciptakan lingkungan yang kondusif, terutama untuk peningkatan perekonomian nasional. Akan tetapi dalam prosesnya peningkatan perekonomian nasional dalam konteks keuangan tidak terlepas dari pada komentar masyarakat yang mengemukakan pendapatnya baik melalui forum – forum ataupun situs *web* yang berkaitan dengan keuangan. Di era sekarang pemikiran – pemikiran tersebut perlu dipertimbangkan untuk mengetahui, sehingga dapat menjadi bahan mengevaluasi dan meningkatkan layanan keuangan pada suatu lembaga sehingga peran serta strategis dapat tercapai dengan sangat baik.

Analisis sentimen merupakan suatu proses memahami, mengekstrak dan memproses data berbasis teks secara otomatis. Proses ini dilakukan untuk memperoleh sebuah informasi sentimen yang tersirat dari suatu opini. Dilakukannya analisis sentimen untuk mengetahui kecenderungan penilaian terhadap permasalahan dari sudut pandang orang lain, umumnya apakah mereka mempunyai opini negatif atau positif terhadap suatu permasalahan. Dalam beberapa penelitian, sentimen sering kali dikategorikan sebagai *big data* dimana ukuran dari data teks tersebut semakin banyak dan makna konteksnya semakin beragam. Saat ini pentingnya mengetahui sentimen masyarakat yang berkaitan dalam konteks dan bidang keuangan, adapun tujuannya untuk mengukur kepercayaan masyarakat umum di sektor keuangan.

Terdapat dataset publik [1] yang berisi dua gabungan sumber data, yaitu data FiQA (*Financial Question Answering*) berisi pasangan pertanyaan dan jawaban yang dikumpulkan dari berbagai sumber seperti forum keuangan dan situs web terkait keuangan, dengan data *Financial PhraseBank* yaitu sebuah data yang berisi kalimat – kalimat pendek yang berhubungan dengan bidang keuangan. Umumnya data teks mengandung URL (*Uniform Resource Locators*), *hashtag*, *tager*, singkatan kata, huruf besar semua dan kata yang berawalan huruf besar, tanda baca atau karakter khusus, *single character*, mengandung angka, serta kata – kata yang umum dalam konteks bahasa Inggris. Kemudian yang sering kali terjadi adanya data *imbalance class*.

Penelitian ini mengidentifikasi sejumlah masalah yang perlu diatasi dalam analisis sentimen pada konteks keuangan. Pertama, data latih dan data uji berasal dari gabungan dua kumpulan data, yaitu data FiQA (*Financial Question Answering*) dan data *Financial PhraseBank* yang diambil dari situs Kaggle. Kedua, algoritma *Logistic Regression* dan *Support Vector Machine* (SVM) secara teoritis hanya cocok untuk klasifikasi biner, sedangkan dalam konteks ini klasifikasi multikelas diperlukan. Ketiga, variasi metrik evaluasi untuk model klasifikasi *Machine Learning* menimbulkan kebutuhan akan skala pengukuran yang tepat untuk menilai model terbaik.

Rumusan masalah dalam penelitian ini mencakup tiga pertanyaan penelitian. Pertama, teknik-teknik *Machine Learning* mana yang cocok untuk mengolah data FiQA dan data *Financial PhraseBank* dalam analisis sentimen. Kedua, metode apa yang dapat diterapkan pada *Logistic Regression* dan SVM untuk klasifikasi multikelas dalam analisis sentimen keuangan. Ketiga, bagaimana cara menentukan model klasifikasi *Machine Learning* terbaik untuk analisis sentimen keuangan.

Tujuan penelitian ini adalah tiga bagian. Pertama, menerapkan teknik pra-pemrosesan teks dan data, termasuk TF-IDF dan Smote, untuk mengatasi masalah pada data FiQA dan data *Financial PhraseBank*. Kedua, menerapkan metode *OneVsRestClassifier*, penyetelan *hyperparameter*, dan validasi silang k-fold pada *Logistic Regression* dan SVM untuk klasifikasi multikelas. Ketiga, menentukan model klasifikasi terbaik melalui evaluasi nilai akurasi dan konfirmasi manual menggunakan matriks kebingungan. Kontribusi penelitian ini adalah mengusulkan metode *Machine Learning* yang sesuai untuk dataset yang digunakan, sehingga menghasilkan model terbaik untuk Analisis Sentimen Keuangan dengan menggunakan *Logistic Regression* dan SVM. Penelitian ini juga diharapkan dapat berkontribusi pada perkembangan pengetahuan dalam pengolahan teks berbahasa Inggris.

Sejumlah penelitian terdahulu telah dilakukan dalam analisis sentimen, yang melibatkan berbagai platform media sosial dan metode analisis yang beragam. Beberapa penelitian yang relevan akan diuraikan untuk memberikan gambaran luas tentang perkembangan penelitian ini. Qian, Xie, dan Bruckmann (2022) menjalankan penelitian tentang analisis sensitivitas pada arsitektur neural yang

digunakan dalam BERT dan GPT-2 untuk menganalisis sentimen keuangan [1]. Santoso, Noviriandini, Kurniasih, Wicaksono, dan Nuryanto (2021) menggali aspek analisis sentimen dengan fokus pada klasifikasi persepsi pengguna Twitter terhadap kasus Covid-19, menggunakan metode Regresi Logistik [2]. Ardiansyah, Saepudin, Aryanti, dan Fitriani (2023) menerapkan analisis sentimen pada ulasan di platform media sosial TikTok dengan algoritma K-NN dan SVM [3].

Selanjutnya, Sari dan Irhamah (2020) melakukan studi yang berkaitan dengan analisis sentimen pada layanan perbankan, memanfaatkan metode Regresi Logistik Biner, Naïve Bayes Classifier, dan Support Vector Machine (SVM) [4]. Turjaman dan Budi (2022) memfokuskan analisis sentimen mereka pada aplikasi Dompot Digital, khususnya aplikasi Linkaja di Twitter, dengan mempertimbangkan aspek Marketing Mix [5]. Novantika dan Sugiman (2022) menganalisis sentimen ulasan pengguna aplikasi Video Conference Google Meet dengan menggunakan metode SVM dan Regresi Logistik [6].

Penelitian yang lain seperti yang dilakukan oleh Hanif, Affandi, Hasan, Sinduningrum, dan Halim (2022) juga berkaitan dengan analisis sentimen, khususnya dalam mengamati opini masyarakat terkait penyelenggaraan sistem elektronik dengan metode Regresi Logistik [7]. Putri dan Kharisudin (2022) menjalankan penelitian yang mengevaluasi analisis sentimen pada pengguna aplikasi Marketplace Tokopedia, menggunakan metode Support Vector Machine, Naïve Bayes, dan Regresi Logistik [8].

Di samping itu, Sreya dan Setiawan (2022) memperkaya fitur analisis sentimen Twitter dengan memanfaatkan metode Glove dan mengaplikasikannya bersama Naïve Bayes dan Support Vector Machine [9]. Lidinillah, Rohana, dan Juwita (2023) mengulas analisis sentimen terhadap ulasan Steam, dengan memanfaatkan algoritma Regresi Logistik dan Support Vector Machine [10]. Noviriandini, Hermanto, dan Yudhistira (2022) mengintegrasikan Support Vector Machine dengan Particle Swarm Optimization untuk menganalisis sentimen pengguna aplikasi Pedulilindungi [11]. Santoso (2022) menganalisis sentimen di Twitter berbahasa Indonesia dengan pendekatan Machine Learning [12].

Pratama, Murdiansyah, dan Lhaksmana (2023) mengarahkan perhatian mereka pada analisis sentimen terkait kendaraan listrik di media sosial Twitter, menggunakan algoritma Regresi Logistik dan Principal Component Analysis [13]. Selain itu, Nurhusen, Indra, dan Baihaqi (2023) melakukan penelitian terkait analisis sentimen pengguna Twitter terhadap kenaikan harga bahan bakar minyak dengan metode Regresi Logistik [14]. Rahmawati dan Sukmasetya (2022) menganalisis sentimen opini masyarakat terhadap kebijakan Kominfo yang berkaitan dengan pemblokiran situs non-PSE di media sosial Twitter [15].

Lestandy, Abdurrahim, dan Syafa'ah (2021) memanfaatkan Recurrent Neural Network dan Naïve Bayes untuk menganalisis sentimen terhadap tweet berkaitan dengan vaksin COVID-19 [16]. Di samping itu, Astuti (2022) menjalankan analisis sentimen terkait penerimaan CPNS 2021 di Twitter dengan menggunakan TextBlob [17]. Kelvin, Banjarnahor, Nababan, dan Sinurat (2022) melakukan perbandingan sentimen terkait Covid-19 di Twitter dengan metode Regresi Logistik dan Support Vector Machine [18]. Lengkong, Safitri, Machsus, Putra, Syahadati, dan Nooraeni (2021) menganalisis sentimen terhadap penerapan PSBB di DKI Jakarta dan dampaknya terhadap pergerakan IHSG [19]. Onantya, Indriati, dan Adikara (2019) menjalankan analisis sentimen pada ulasan aplikasi BCA Mobile dengan BM25 dan Improved K-Nearest Neighbor [20].

Selain itu, beberapa penelitian mengkaji konteks yang berbeda seperti literasi keuangan dan dampaknya terhadap UMKM [21], pengelompokan dan prediksi karyawan yang berpotensi PHK [22], aplikasi CV Matcher untuk mencocokkan CV dengan lowongan pekerjaan menggunakan Machine Learning [23], serta pembelajaran bahasa pemrograman C++ [24]. Ragam penelitian ini mencerminkan diversifikasi pendekatan dan fokus penelitian dalam ranah analisis sentimen yang telah dikembangkan dalam literatur terdahulu. Penelitian terdahulu yang telah dipaparkan memberikan landasan yang signifikan dan perspektif dalam pengembangan penelitian ini.

2. Metode Penelitian

Pada penelitian ini penulis menggunakan dataset *public* yang bersumber pada situs *Kaggle* dengan judul dataset *Financial Sentiment Analysis* (<https://www.kaggle.com/sbhatti/financial-sentiment-analysis>) [1]. Dataset yang digunakan berkaitan dengan konteks atau bidang keuangan yang berupa kalimat (teks) finansial berbahasa Inggris dengan label sentimen. Data teks pada dataset yaitu gabungan dari dua kumpulan data yaitu data FiQA (*Financial Question Answering*) dan data *Financial PhraseBank* yang digabungkan menjadi satu file CSV. Dataset ini termasuk ke dalam jenis data kualitatif dan termasuk jenis data sekunder. Dimana data kualitatif merupakan jenis data yang dapat diamati dan tidak dapat dihitung atau non-numerik, sedangkan data sekunder merupakan data yang diperoleh peneliti secara tidak langsung karena data diperoleh melalui perantara (situs *Kaggle*).

Kemudian dataset *Financial PhraseBank* adalah dataset yang berisi kalimat – kalimat pendek yang berhubungan dengan bidang keuangan berfokus pada kalimat – kalimat yang sering digunakan dalam laporan keuangan, analisis saham, atau dalam konteks keuangan lainnya. Tabel 1 dibawah ini merupakan tampilan 10 baris pertama dari dataset yang digunakan.

Tabel 1. Dataset

No	Sentence	Sentiment
1	<i>The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology, a Communities Platform, location relevant multimedia content and a new and powerful commercial model.</i>	Positive
2	<i>\$ESI on lows, down \$1.50 to \$2.50 BK a real possibility</i>	Negative
3	<i>For the last quarter of 2010, Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier, while it moved to a zero pre-tax profit from a pre-tax loss of EUR7m.</i>	Positive
4	<i>According to the Finnish-Russian Chamber of Commerce, all the major construction companies of Finland are operating in Russia.</i>	Neutral
5	<i>The Swedish buyout firm has sold its remaining 22.4 percent stake, almost eighteen months after taking the company public in Finland.</i>	Neutral
6	<i>\$SPY wouldn't be surprised to see a green close</i>	Positive
7	<i>Shell's \$70 Billion BG Deal Meets Shareholder Skepticism</i>	Negative
8	<i>SSH COMMUNICATIONS SECURITY CORP STOCK EXCHANGE RELEASE OCTOBER 14, 2008 AT 2:45 PM The Company updates its full year outlook and estimates its results to remain at loss for the full year.</i>	Negative
9	<i>Kone 's net sales rose by some 14 % year-on-year in the first nine months of 2008.</i>	Positive
10	<i>The Stockmann department store will have a total floor space of over 8,000 square metres and Stockmann 's investment in the project will have a price tag of about EUR 12 million.</i>	Neutral

3. Hasil dan Pembahasan

3.1 Alat Penelitian

1) Spesifikasi *Software*

Tabel 2 dibawah ini merupakan spesifikasi dari *software* yang digunakan oleh penulis dalam membuat perancangan dan eksperimen penelitian. Sistem operasi *Windows 10* merupakan jenis sistem operasi yang digunakan pada laptop penulis. Kemudian menggunakan *Google Colaboratory* pada *Google Drive* dengan bahasa pemrograman *Python v3.10* untuk membuat program dari eksperimen penelitian, dan menggunakan *browser Google Chrome* untuk menjalankan program yang ada pada *Google Colaboratory*. Selanjutnya penulis menggunakan *Microsoft Word Office 2016* untuk membuat laporan dari penelitian ini dan menggunakan *tools draw io* (<https://app.diagrams.net/>) yang diakses secara *online* untuk membuat diagram pada penelitian ini.

Tabel 2. Spesifikasi *Software*

Jenis <i>Software</i>	Spesifikasi
Sistem Operasi	Windows 10
<i>Browser</i>	<i>Google Chrome</i>
<i>Code Editor</i>	<i>Google Colaboratory</i>
Bahasa Pemrograman	<i>Python v3.10</i>
Ms. Word	Office 2016
Draw io	Draw io

2) Spesifikasi *Hardware*

Dalam penelitian ini penulis menggunakan laptop **Asus seri A441U** dengan spesifikasi *hardware* sebagai berikut Tabel 3 :

Tabel 3. Spesifikasi *Hardware*

Jenis <i>Hardware</i>	Spesifikasi
<i>Processor</i>	Intel Core i5-8250U
<i>Memory</i>	DDR4 8 GB
<i>Storage</i>	HDD 1 TB
<i>Display</i>	14" FHD (1920 x 1080)

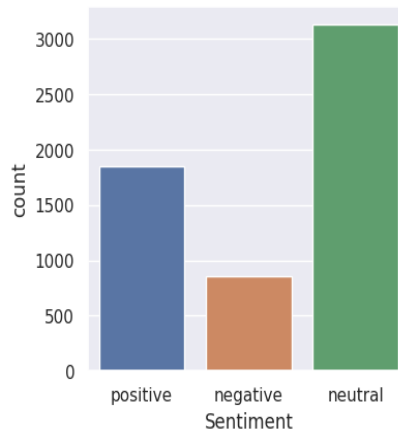
3.2 Hasil Data *Exploration*

Dengan memanfaatkan *library pandas, numpy, matplotlib, dan seaborn* untuk melakukan eksplorasi data. Hasil eksplorasi data yang telah dilakukan bahwa terdapat jumlah data sebanyak 5.842 data (baris) dan dua atribut yaitu atribut "*Sentence*" yang berisi kalimat teks berbahasa Inggris yang berkaitan dengan konteks keuangan dan atribut "*Sentiment*" yang berisi label sentimen untuk analisis sentimen terkait konteks keuangan. Dimana kedua atribut tersebut yaitu atribut '*Sentence*' dan atribut '*Sentiment*' bertipe data *object*.

Setelah dilakukan pengecekan data yang kosong dan data yang duplikat bahwa dinyatakan tidak adanya indikasi dari data yang kosong, akan tetapi terindikasi adanya data yang duplikat sebanyak 6 baris data. Unik data dari label atribut '*Sentiment*' yang menjadi target klasifikasi model yaitu terdiri dari label *positive*, label *negative* dan label *neutral*. Tabel 4 jumlah distribusi data berdasarkan label yaitu sebanyak 1.852 data untuk label *positive*, 3.130 data untuk label *neutral*, dan 860 data untuk label *negative*. Banyaknya data tersebut seperti divisualisasikan pada diagram batang Gambar 3.1.

Tabel 4. Jumlah Data Berdasarkan Label

Label	Jumlah
<i>Positive</i>	1.852
<i>Neutral</i>	3.130
<i>negative</i>	860



Gambar 1. Visualisasi Jumlah Data Berdasarkan Label

Selanjutnya setelah dilakukan pengecekan dan pengamatan dari pada data teks keuangan (atribut ‘Sentence’), telah terindikasi adanya data teks yang mengandung URL (*Uniform Resource Locators*), *hashtag* (@), *tager* (#), kosakata yang disingkat, kalimat dengan huruf besar semua dan kosakata yang berawalan huruf besar, tanda baca, *single character*, kata – kata yang umum dalam konteks bahasa Inggris dan mengandung angka. Hasil dari pada data teks yang mengandung tersebut, 10 baris data yang secara acak (tidak berurutan) disajikan pada Tabel 3.4 dibawah ini.

Tabel 5. Data Teks

No	Sentence	Sentiment
1	<i>The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology, a Communities Platform, location relevant multimedia content and a new and powerful commercial model.</i>	<i>positive</i>
2	<i>\$MSFT SQL Server revenue grew double-digit with SQL Server Premium revenue growing over 30% http:// stks.co/ ir2F</i>	<i>Positive</i>
3	<i>HUHTAMAKI OYJ STOCK EXCHANGE RELEASE, 16.9.2008 AT 13.32 Huhtamaki 's Capital Markets Day for institutional investors and analysts is held in Espoo, September 16, 2008 starting at 13.30 pm Finnish time.</i>	<i>Neutral</i>
4	<i>RT @andrennyquist Chartology: Yahoo: http:// stks.co/ 1itr \$YHOO Crazy looking chart, but nearing breakout resistance line...</i>	<i>Positive</i>
5	<i>Finnish IT solutions provider Affecto Oyj HEL: AFE1V said today its slipped to a net loss of EUR 115,000 USD 152,000 in the second quarter of 2010 from a profit of EUR 845,000 in the corresponding period a year earlier.</i>	<i>Neutral</i>
6	<i>66Netflix Grows Efforts to Create Loyal Customers \$NFLX https:// t.co/ 5QXBkoBpbe</i>	<i>Positive</i>
7	<i>Facebook, near a buy point last week, faces a different technical test today https:// t.co/ c72LLMpiNM \$FB \$AAPL \$NFLX https:// t.co/ jPFbYTYpUy</i>	<i>Positive</i>
8	<i>Coke dividend - 3%. Google dividend - 0%. For passive income Coke takes it. \$GOOG \$KO https:// t.co/ QBsoWjxOST</i>	<i>Neutral</i>
9	<i>\$CRM Sep 40 calls are +35% since entry #BANG http:// stks.co/ deDm</i>	<i>Positive</i>
10	<i>Tesla is recalling 2,700 Model X cars: https:// t.co/ 8Z7BkVsTl9 \$TSLA</i>	<i>Negative</i>

3.3 Hasil Teks Preprocessing

Tabel 6 dibawah ini merupakan 10 baris data yang disajikan secara acak (tidak berurutan) dari data teks awal pada dataset atau data teks yang belum dilakukan normalisasi teks, dan hasil dari data teks yang telah dilakukan normalisasi teks. Berdasarkan hasil dari eksperimen seperti menghapus *url*, *hashtag*, *mention*, tanda baca, karakter huruf, spasi yang ganda pada teks dan menghapus *tab* atau *newline*,

menggunakan ekspresi reguler (*regex*) dengan *Python*.

Tabel 6. Hasil Normalisasi Teks

No	Sebelum	Sesudah Normalisasi Teks
1	<i>The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology, a Communities Platform, location relevant multimedia content and a new and powerful commercial model.</i>	<i>the geosolutions technology will leverage benefon gps solutions by providing location-based search technology communities platform location relevant multimedia content and new and powerful commercial model</i>
2	<i>\$MSFT SQL Server revenue grew double-digit with SQL Server Premium revenue growing over 30% http://stks.co/ir2F</i>	<i>msft sql server revenue grew doubledigit with sql server premium revenue growing over 30</i>
3	<i>HUHTAMAKI OYJ STOCK EXCHANGE RELEASE, 16.9.2008 AT 13.32 Huhtamaki 's Capital Markets Day for institutional investors and analysts is held in Espoo, September 16, 2008 starting at 13.30 pm Finnish time.</i>	<i>huhtamaki oyj stock exchange release 1692008 at 1332 huhtamaki capital markets day for institutional investors and analysts is held in espoo september 16 2008 starting at 1330 pm finnish time</i>
4	<i>RT @andrewnyquist Chartology: Yahoo: http://stks.co/1itr \$YHOO Crazy looking chart, but nearing breakout resistance line...</i>	<i>rt chartology yahoo yboo crazy looking chart but nearing breakout resistance line</i>
5	<i>Finnish IT solutions provider Affecto Oyj HEL: AFE1V said today its slipped to a net loss of EUR 115,000 USD 152,000 in the second quarter of 2010 from a profit of EUR 845,000 in the corresponding period a year earlier.</i>	<i>finnish its solutions provider affecto oyj hel afe1v said today its slipped to net loss of eur 115000 usd 152000 in the second quarter of 2010 from profit of eur 845000 in the corresponding period year earlier</i>
6	<i>66Netflix Grows Efforts to Create Loyal Customers \$NFLX https://t.co/5QXBkoBpbe</i>	<i>netflix grows efforts to create loyal customers nflx</i>
7	<i>Facebook, near a buy point last week, faces a different technical test today https://t.co/c72LLMpiNM \$FB \$AAPL \$NFLX https://t.co/jPFbYTYPuY</i>	<i>facebook near buy point last week faces different technical test today fb aapl nflx</i>
8	<i>Coke dividend – 3%. Google dividend – 0%. For passive income Coke takes it. \$GOOG \$KO https://t.co/QBsoWjxOST</i>	<i>coke dividend 3 google dividend 0 for passive income coke takes it goog ko</i>
9	<i>\$CRM Sep 40 calls are +35% since entry #BANG http://stks.co/deDm</i>	<i>crm sep 40 calls are 35 since entry</i>
10	<i>Tesla is recalling 2,700 Model X cars: https://t.co/8Z7BkVstI9 \$TSLA</i>	<i>tesla is recalling 2700 model cars tsla</i>

Kemudian Tabel 7 dibawah ini merupakan 10 baris data yang disajikan secara acak (tidak berurutan) dari hasil data teks yang telah dilakukan normalisasi teks, untuk selanjutnya dari hasil data teks yang telah dilakukan *tokenize* menggunakan *word_tokenize* pada NLTK (*Natural Language Tool Kit*) dan penghapusan *stopwords* menggunakan *library stopwords* berbahasa Inggris pada NLTK dengan *Python*.

Tabel 7. Hasil *Tokenize* dan Menghapus *Stopwords*

No	Sesudah Normalisasi Teks	Sesudah <i>Tokenize</i> dan Menghapus <i>Stopwords</i>
1	<i>the geosolutions technology will leverage benefon gps solutions by providing location-based search technology communities platform location relevant multimedia content and new and powerful commercial model</i>	<i>['geosolutions', 'technology', 'leverage', 'benefon', 'gps', 'solutions', 'providing', 'location', 'based', 'search', 'technology', 'communities', 'platform', 'location', 'relevant', 'multimedia', 'content', 'new', 'powerful', 'commercial', 'model']</i>
2	<i>msft sql server revenue grew doubledigit with sql server premium revenue growing over 30</i>	<i>['msft', 'sql', 'server', 'revenue', 'grew', 'doubledigit', 'sql', 'server', 'premium', 'revenue', 'growing', '30']</i>
3	<i>hubtamaki ojj stock exchange release 1692008 at 1332 hubtamaki capital markets day for institutional investors and analysts is held in espoo september 16 2008 starting at 1330 pm finnish time</i>	<i>['hubtamaki', 'ojj', 'stock', 'exchange', 'release', '1692008', '1332', 'hubtamaki', 'capital', 'markets', 'day', 'institutional', 'investors', 'analysts', 'held', ' espoo', 'september', '16', '2008', 'starting', '1330', 'pm', 'finnish', 'time']</i>
4	<i>rt chartology yahoo yhoo crazy looking chart but nearing breakout resistance line</i>	<i>['rt', 'chartology', 'yahoo', 'yhoo', 'crazy', 'looking', 'chart', 'nearing', 'breakout', 'resistance', 'line']</i>
5	<i>finnish it solutions provider affecto ojj bel afe1v said today its slipped to net loss of eur 115000 usd 152000 in the second quarter of 2010 from profit of eur 845000 in the corresponding period year earlier</i>	<i>['finnish', 'solutions', 'provider', 'affecto', 'ojj', 'bel', 'afe1v', 'said', 'today', 'slipped', 'net', 'loss', 'eur', '115000', 'usd', '152000', 'second', 'quarter', '2010', 'profit', 'eur', '845000', 'corresponding', 'period', 'year', 'earlier']</i>
6	<i>netflix grows efforts to create loyal customers nflx</i>	<i>['netflix', 'grows', 'efforts', 'create', 'loyal', 'customers', 'nflx']</i>
7	<i>facebook near buy point last week faces different technical test today fb aapl nflx</i>	<i>['facebook', 'near', 'buy', 'point', 'last', 'week', 'faces', 'different', 'technical', 'test', 'today', 'fb', 'aapl', 'nflx']</i>
8	<i>coke dividend 3 google dividend 0 for passive income coke takes it goog ko</i>	<i>['coke', 'dividend', '3', 'google', 'dividend', '0', 'passive', 'income', 'coke', 'takes', 'goog', 'ko']</i>
9	<i>crm sep 40 calls are 35 since entry</i>	<i>['crm', 'sep', '40', 'calls', '35', 'since', 'entry']</i>
10	<i>tesla is recalling 2700 model cars tsla</i>	<i>['tesla', 'recalling', '2700', 'model', 'cars', 'tsla']</i>

3.4 Hasil *Feature Extraction*

Gambar 2 *Feature extraction* yang digunakan untuk mengkonversi token data teks menjadi numerik menggunakan *tfidfvectorizer* dari *scikit learn feature extraction text* dengan *max_features* 1000. Data latih (X_train) dilakukan *fit_transform*, dan data uji (X_test) hanya dilakukan *transform*. Data X merupakan data dari *input* untuk model yang diambil dari atribut 'Sentence', dimana X_train merupakan dari pembagian data yaitu data latih, dan X_test dari data uji.

```
matrix = TfidfVectorizer(max_features=1000)
X_train = matrix.fit_transform(Xtrain).toarray()
X_test = matrix.transform(Xtest).toarray()
```

Gambar 2. *Feature Extraction* TF-IDF

3.5 Hasil Pembuatan Model

Hasil dari tahapan pembuatan model seperti dilakukannya pembagian data, setelah melakukan teknik *imbalance class* Smote, *training* model dengan *hyperparameter tuning* disajikan secara jelas dan mendetail sebagai berikut:

1) Hasil Pembagian Data

Setelah data dilakukan tahapan data *exploration*, *text preprocessing*, dan data *preprocessing*, yang selanjutnya yaitu dilakukan pembagian data menjadi data latih sebesar 80% untuk melakukan *training* model klasifikasi dengan algoritma *Logistic Regression* dan *Support Vector Machine*. Data uji sebesar 20% untuk melakukan evaluasi model klasifikasi *Machine Learning* tersebut. Pembagian data menggunakan

train_test_split dari *scikit learn* dengan nilai *random_state* 42. Data latih kemudian dibagi lagi menjadi *X_train* yang diambil dari data atribut ‘Sentence’ dan *y_train* dari data atribut ‘Sentiment’. Begitu juga data uji, *X_test* diambil dari data atribut ‘Sentence’, dan *y_test* dari data atribut ‘Sentiment’. Dihasilkan untuk jumlah data latih sebanyak 4.656 baris data dan data uji sebanyak 1.164 baris data. Tabel 8 merupakan jumlah data dari pembagian data latih dan data uji.

Tabel 8. Jumlah Pembagian Data

Data	Persentase	Jumlah
Data Latih	80%	4.656
Data Uji	20%	1.164

2) Hasil Smote

Tabel 9 merupakan perbandingan jumlah data latih sebelum dilakukannya teknik Smote dan setelah dilakukannya teknik Smote. Teknik *Imbalance class* Smote hanya dilakukan pada data latih menggunakan Smote dengan *Imblearn Oversampling* dari *Scikit Learn*. *X_train* merupakan data *input* dari model klasifikasi menggunakan data teks dari atribut ‘*Sentence*’ yang telah di *feature extraction*, dan *y_train* merupakan data label target klasifikasi model dari atribut ‘*Sentiment*’ yang telah di *label encoder*. Kedua data tersebut merupakan bagian dari data latih.

Tabel 9. Hasil Smote

Data	Jumlah Sebelum Smote	Jumlah Sesudah Smote
2 (<i>positive</i>)	1.488	2.478
1 (<i>neutral</i>)	2.478	2.478
0 (<i>negative</i>)	690	2.478
<i>X_train</i>	(4.656)	(7.434, 1000)
<i>y_train</i>	(4.656)	(7.434)

3) Hasil Training Model dengan *Hyperparameter Tunning*

Training model klasifikasi untuk algoritma *Logistic Regression* dengan menggunakan *GridSearchCV* untuk membantu menemukan nilai dari *hyperparameter* terbaik, dan menggunakan *max_iter* 1000, *k-Fold Cross Validation* 10, serta menggunakan *OneVsRestClassifier* dari *scikit learn* untuk melatih model dengan *hyperparameter tuning* ‘*penalty*’ dan ‘*C*’. Model dilatih dengan menggunakan data latih yang tidak dilakukan teknik *imbalance class* Smote dan dengan data latih yang telah dilakukan teknik *imbalance class* Smote.

Hasil dari *training* model klasifikasi algoritma *Logistic Regression* yang tidak menggunakan teknik Smote, menghasilkan nilai *hyperparameter* terbaik pada ‘*penalty*’ nilai l2, dan ‘*C*’ nilai 1, serta dengan nilai *score training* terbaik sebesar 68,51%. Kemudian model klasifikasi algoritma *Logistic Regression* ketika menggunakan teknik Smote, menghasilkan nilai *hyperparameter* terbaik pada ‘*penalty*’ nilai l2, dan ‘*C*’ nilai 100, serta dengan nilai *score training* terbaik sebesar 73,79%. Kedua hasil tersebut disajikan pada Tabel 10 dibawah ini.

Tabel 10. Hasil Training Model dengan *Hyperparameter Tunning Logistic Regression*

Model	Hyperparameter	Nilai	Best Score for Training
Tanpa Smote	Penalty	l2	68,51%
	C	1	
Dengan Smote	Penalty	l2	73,79%
	C	100	

Training model klasifikasi untuk algoritma *Support Vector Machine* (SVM) dengan ‘kernel’ rbf, poly dan linear. Masing – masing dari ‘kernel’ algoritma SVM tersebut dilatih menggunakan *GridSearchCV*

untuk membantu menemukan nilai dari *hyperparameter* terbaik, dan menggunakan *k-Fold Cross Validation* 10, serta menggunakan *OneVsRestClassifier* dari *scikit learn* untuk melatih model dengan *hyperparameter tuning* ‘gamma’ dan ‘C’. Tabel diatas merupakan hasil *training* model klasifikasi dengan algoritma *Support Vector Machine* pada masing – masing ‘kernel’ yang digunakan. Setiap ‘kernel’ pada algoritma *Support Vector Machine* di *training* dengan menggunakan *hyperparameter tuning* ‘gamma’ dan ‘C’, baik itu ketika model di *training* menggunakan data latih yang tidak dilakukan teknik Smote dan ketika model di *training* menggunakan data latih yang dilakukan teknik Smote. Ketika tidak menggunakan Smote, model terbaik dihasilkan pada ‘kernel’ linier dengan nilai ‘gamma’ 1 dan ‘C’ 1 pada score training terbaik yaitu 68,39%. Kemudian ketika menggunakan Smote, model terbaik dihasilkan pada ‘kernel’ rbf dengan nilai ‘gamma’ 1 dan ‘C’ 10 pada score training terbaik yaitu 79,37%.

Tabel 11. Hasil *Training* Model dengan *Hyperparameter Tunning Support Vector Machine*

Model	Kernel	Gamma	C	Best Score for Training
Tanpa Smote	rbf	0,1	10	67,12%
	poly	1	10	59,56%
	linear	1	1	68,39%
Dengan Smote	rbf	1	10	79,37%
	poly	1	1	77,44%
	linear	1	100	73,97%

Berdasarkan dari hasil *training* model klasifikasi kedua algoritma tersebut, dihasilkan model *training* terbaik pada algoritma *Support Vector Machine* dengan ‘kernel’ rbf pada nilai ‘gamma’ sebesar 1 dan nilai ‘C’ sebesar 10 dengan *score training* sebesar 79,37% pada data latih ketika sudah dilakukan teknik *imbalance class* Smote.

3.6 Hasil Evaluasi Model

Evaluasi model dilakukan dengan menggunakan data uji untuk melihat seberapa baik model akan melakukan prediksi dari data yang benar – benar tidak terdapat pada data *training*. Tabel 12 merupakan hasil evaluasi model klasifikasi algoritma *Logistic Regression*, menghasilkan evaluasi model terbaik pada saat menggunakan *training* model ketika tidak melakukan *imbalance class* Smote, menghasilkan nilai *accuracy* sebesar 69,59% pada *hyperparameter* terbaik dengan nilai ‘penalty’ l2 dan nilai ‘C’ 1 ketika *training* model.

Tabel 12. Hasil Evaluasi Model *Logistic Regression*

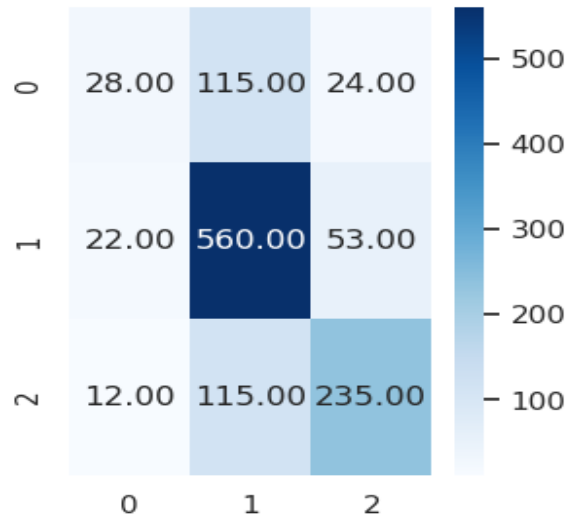
Model	Accuracy
Tanpa Smote	69,59%
Dengan Smote	59,19%

Tabel 13 merupakan hasil evaluasi model klasifikasi algoritma *Support Vector Machine*, menghasilkan evaluasi model terbaik pada saat menggunakan *training* model ketika tidak melakukan *imbalance class* Smote, menghasilkan nilai *accuracy* sebesar 70,70% pada ‘kernel’ linier, dengan *hyperparameter* terbaik pada nilai ‘gamma’ 1 dan ‘C’ 1 ketika *training* model.

Tabel 13. Hasil Evaluasi Model *Support Vector Machine*

Model	Kernel	Accuracy
Tanpa Smote	Rbf	68,90%
	Poly	61,60%
	linear	70,70%
Dengan Smote	Rbf	65,12%
	Poly	62,29%
	linear	56,79%

Berdasarkan hasil dari evaluasi model klasifikasi kedua algoritma tersebut, dihasilkan model terbaik pada algoritma *Support Vector Machine* dengan ‘kernel’ linear menggunakan *hyperparameter* terbaik dengan nilai ‘gamma’ 1 dan nilai ‘C’ 1 pada saat *training* model menggunakan data latih yang tidak dilakukan *imbalance class* Smote, kemudian di evaluasi menggunakan data uji yaitu menghasilkan *accuracy* sebesar 70,70%. Gambar 3 dibawah ini merupakan hasil dari tabel *confusion matrix* pada model terbaik dari kedua algoritma tersebut, yaitu model dengan algoritma *Support Vector Machine* dengan ‘kernel’ linear, nilai ‘gamma’ 1, dan nilai ‘C’ 1.



Gambar 3. *Confusion Matrix*

Sumber: Hasil Penelitian Penulis

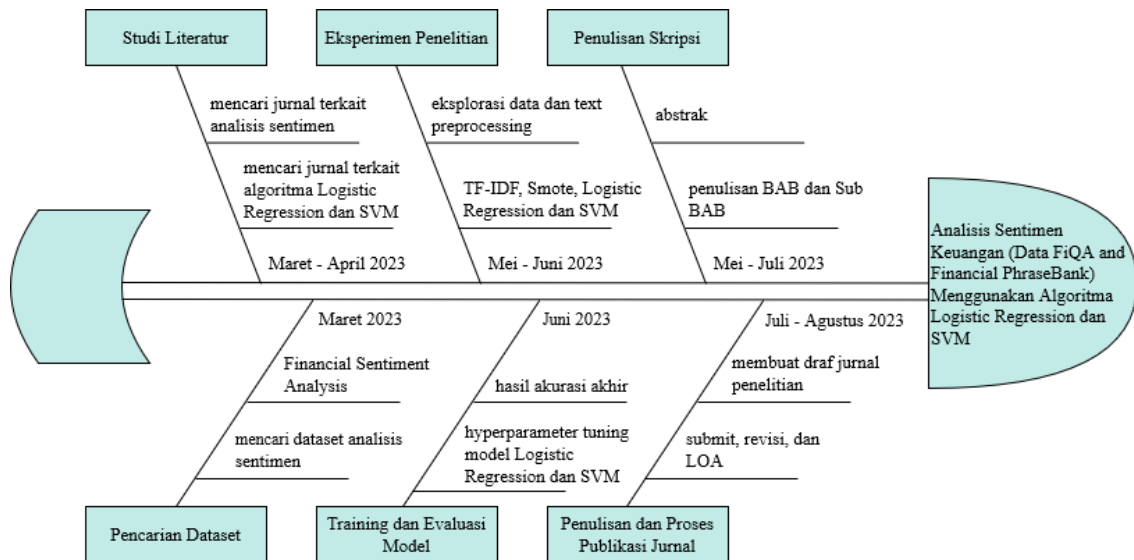
Berdasarkan hasil nilai pada tabel *confusion matrix*, maka perhitungan manual untuk mencari nilai *accuracy* model yaitu sebagai berikut:

$$accuracy = \frac{\# \text{ prediksi yang benar}}{\# \text{ total data point}}$$

$$accuracy = \frac{28 + 560 + 235}{28 + 115 + 24 + 22 + 560 + 53 + 12 + 115 + 235}$$

$$accuracy = \frac{823}{1.164} = 0,7070446735 = 70,70\%$$

Berdasarkan dari hasil perhitungan nilai *accuracy* secara manual dan dari nilai *accuracy* yang dihasilkan pada program, dapat dinyatakan kedua hasil tersebut sama – sama menghasilkan nilai dengan *accuracy* sebesar 70,70%.



Gambar 4. Roadmap Penelitian

Berdasarkan pada Gambar 4 diatas, penelitian pada skripsi ini untuk mencapai pada hasil akhir dari pembuatan model *machine learning* tentang analisis sentimen menggunakan data teks keuangan (*Data FiQA and Financial PhraseBank*) dengan algoritma *Logistic Regression* dan *Support Vector Machine* yaitu dimulai dari melakukan pencarian studi literatur, pencarian dataset, kemudian melakukan eksperimen penelitian, *training* dan evaluasi model, penulisan laporan skripsi dan penulisan serta proses publikasi jurnal penelitian skripsi ini.

3.7 Hasil Data *Preprocessing*

Setelah melalui tahapan data *exploration* pada dataset dan *text preprocessing* pada data teks keuangan (atribut '*Sentence*') dari total data sebanyak 5.842 data (baris), yang selanjutnya dilakukan pengecekan kembali untuk mengetahui berapa banyak mengandung data yang kosong dan data yang duplikat setelah dilakukan proses *text preprocessing*. Berdasarkan hasil pengecekan tersebut telah terindikasi mengandung data yang kosong berjumlah 1 baris data pada atribut '*Sentence*' dari yang sebelumnya tidak ada, dan mengandung data yang duplikat sebanyak 21 baris data dari yang sebelumnya hanya mengandung 6 baris data duplikat pada dataset awal. Tabel 14 dibawah ini merupakan perbandingan dari jumlah data yang kosong dan data yang duplikat, sebelum dilakukannya proses *text preprocessing* (dataset awal) dan setelah dilakukannya tahapan *text preprocessing*.

Tabel 14. Jumlah Data yang Kosong dan Duplikat

Pengecekan	Sebelum <i>Text Preprocessing</i>	Sesudah <i>Text Preprocessing</i>
Data yang Kosong	0	1
Data yang Duplikat	6	21

Berdasarkan hasil pengecekan tersebut, pertama dilakukan penghapusan baris data untuk data yang kosong sehingga jumlah data setelah penghapusan baris data yang kosong menjadi sebanyak 5.841 baris data dari yang sebelumnya berjumlah sebanyak 5.842 baris data. Selanjutnya dilakukan penghapusan data yang duplikat sehingga banyaknya data setelah penghapusan data duplikat menjadi 5.820 baris data. Jumlah data tersebut disajikan pada tabel 15.

Tabel 15. Jumlah Data Setelah Penghapusan

Proses	Sebelum	Sesudah
Penghapusan Data yang Kosong	5.842	5.841
Penghapusan Data yang Duplikat	5.841	5.820

Selanjutnya Tabel 16 merupakan hasil dari proses data *preprocessing* yaitu mengubah label target klasifikasi model dari yang sebelumnya berbentuk kategorikal menjadi label target berbentuk data numerik menggunakan *label encoder* dari *scikit learn*. Label ‘*negative*’ menjadi 0, label ‘*neutral*’ menjadi 1, dan label ‘*positive*’ menjadi 2.

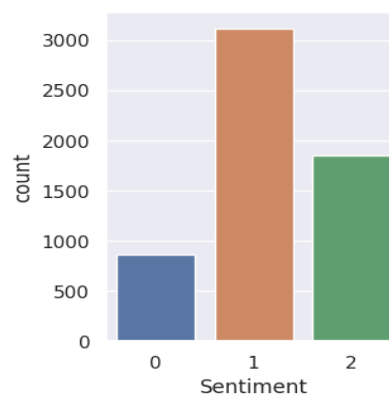
Tabel 16. Hasil *Label Encoder*

Sebelum	Sesudah
<i>Negative</i>	0
<i>Neutral</i>	1
<i>Positive</i>	2

Setelah dilakukan proses data *preprocessing*, dilakukan kembali pengecekan jumlah distribusi data berdasarkan label target klasifikasi model *Machine Learning* dengan tujuan untuk mengetahui kembali jumlah data dari masing – masing label setelah penghapusan data. Tabel 17 dibawah ini disajikan jumlah data berdasarkan label, yaitu jumlah data untuk label ‘*positive*’ sebanyak 1.850 data, label ‘*neutral*’ sebanyak 3.113 data, dan label ‘*negative*’ sebanyak 857 data.

Tabel 17. Jumlah Data Berdasarkan Label

Label	Jumlah
2 (<i>positive</i>)	1.850
1 (<i>neutral</i>)	3.113
0 (<i>negative</i>)	857



Gambar 3.6 Visualisasi Jumlah Data Berdasarkan Label

4. Kesimpulan

Berdasarkan penelitian dan hasil penerapan metode – metode *Machine Learning* serta algoritma *Logistic Regression* dan *Support Vector Machine* dalam mengklasifikasikan sentimen data teks terkait konteks keuangan, dapat diambil kesimpulan sebagai berikut:

- 1) Dengan menerapkan teknik - teknik *text preprocessing* yang tepat dan sesuai kebutuhan pada data teks keuangan berbahasa Inggris, pada penelitian ini telah berhasil menjadikan data teks menjadi lebih terstruktur, dan penggunaan pembobotan TF-IDF sangat membantu dalam meningkatkan hasil akurasi. Akan tetapi penggunaan teknik Smote untuk mengatasi *imbalance class* pada data menjadi tidak cocok pada kasus ini, hal ini dibuktikan dari hasil evaluasi model dimana ketika menggunakan teknik Smote malah menurunkan tingkat akurasi.
- 2) Dengan menerapkan *OneVsRestClassifier* pada algoritma *Logistic Regression* dan *Support Vector Machine* telah berhasil memaksimalkan *training* model klasifikasi *machine learning* dalam kasus data *multiclass*. Dan penggunaan *hyperparameter tuning* dan *k-fold cross validation* dapat membantu kinerja *training* model menjadi lebih baik.

- 3) Hasil model terbaik pada algoritma *Support Vector Machine* dengan ‘kernel’ linear menggunakan *hyperparameter* terbaik dengan nilai ‘gamma’ 1 dan nilai ‘C’ 1 pada saat *training* model menggunakan data latih yang tidak dilakukan *imbalance class* Smote, kemudian di evaluasi menggunakan data uji yaitu menghasilkan *accuracy* sebesar 70,70%. Hal ini berhasil dibuktikan dengan perhitungan manual yang sama – sama menghasilkan nilai *accuracy* sebesar 70,70%.

5. Daftar Pustaka

- [1] Qian, T., Xie, A., & Bruckmann, C. (2022). Sensitivity Analysis on Transferred Neural Architectures of BERT and GPT-2 for Financial Sentiment Analysis. *arXiv preprint arXiv:2207.03037*. DOI: [10.48550/arXiv.2207.03037](https://doi.org/10.48550/arXiv.2207.03037).
- [2] Santoso, A. K., Noviriandini, A., Kurniasih, A., Wicaksono, B. D., & Nuryanto, A. (2021). Klasifikasi Persepsi Pengguna Twitter Terhadap Kasus Covid-19 Menggunakan Metode Logistic Regression. *Jurnal Informatika Kaputama (JIK)*, 5(2), 234-241.
- [3] Ardiansyah, D., Saepudin, A., Aryanti, R., & Fitriani, E. (2023). ANALISIS SENTIMEN REVIEW PADA APLIKASI MEDIA SOSIAL TIKTOK MENGGUNAKAN ALGORITMA K-NN DAN SVM BERBASIS PSO. *Jurnal Informatika Kaputama (JIK)*, 7(2), 233-241. DOI: <https://doi.org/10.59697/jik.v7i2.148>.
- [4] Sari, E. D. N., & Irhamah, I. (2020). Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, Naïve Bayes Classifier (NBC), dan Support Vector Machine (SVM). *Jurnal Sains dan Seni ITS*, 8(2), D177-D184. DOI: [10.12962/j23373520.v8i2.44565](https://doi.org/10.12962/j23373520.v8i2.44565).
- [5] Turjaman, R. M., & Budi, I. (2022). Analisis Sentimen Berbasis Aspek Marketing Mix Terhadap Ulasan Aplikasi Dompot Digital (Studi Kasus: Aplikasi Linkaja Pada Twitter). *Jurnal Darma Agung*, 30(2), 266-275.
- [6] Novantika, A., & Sugiman, S. (2022, February). Analisis Sentimen Ulasan Pengguna Aplikasi Video Conference Google Meet menggunakan Metode SVM dan Logistic Regression. In *PRISMA, Prosiding Seminar Nasional Matematika* (Vol. 5, pp. 808-813).
- [7] Hanif, I. F., Affandi, I. R., Hasan, F. N., Sinduningrum, E., & Halim, Z. (2022). Analisis Sentimen Opini Masyarakat Terkait Penyelenggaraan Sistem Elektronik Menggunakan Metode Logistic Regression. *Jurnal Linguistik Komputasional*, 5(2), 77-84. DOI: <https://doi.org/10.26418/jlk.v5i2.103>
- [8] Putri, M. I., & Kharisudin, I. (2022, February). Analisis Sentimen Pengguna Aplikasi Marketplace Tokopedia Pada Situs Google Play Menggunakan Metode Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression. In *PRISMA, Prosiding Seminar Nasional Matematika* (Vol. 5, pp. 759-766).
- [9] Sreya, M. D. D., & Setiawan, E. B. (2022). Penggunaan Metode Glove Untuk Ekspansi Fitur Pada Analisis Sentimen Twitter Dengan Naïve Bayes Dan Support Vector Machine. *eProceedings of Engineering*, 9(3).

- [10] Lidinillah, E. R., Rohana, T., & Juwita, A. R. (2023). Analisis sentimen twitter terhadap steam menggunakan algoritma logistic regression dan support vector machine. *TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika*, 10(2), 154-164. DOI: <https://doi.org/10.37373/tekno.v10i2.440>
- [11] Noviriandini, A., Hermanto, H., & Yudhistira, Y. (2022). Klasifikasi Support Vector Machine Berbasis Particle Swarm Optimization Untuk Analisa Sentimen Pengguna Aplikasi Pedulilindungi. *JIKA (Jurnal Informatika)*, 6(1), 50-56.
- [12] Santoso, A. K. (2022). ANALISIS SENTIMEN TWITTER BAHASA INDONESIA MENGGUNAKAN PENDEKATAN MACHINE LEARNING. *Jurnal Informatika Kaputama (JIK)*, 6(2), 129-136.
- [13] Pratama, Y., Murdiansyah, D. T., & Lhaksana, K. M. (2023). Analisis Sentimen Kendaraan Listrik Pada Media Sosial Twitter Menggunakan Algoritma Logistic Regression dan Principal Component Analysis. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7(1), 529-535. DOI: <http://dx.doi.org/10.30865/mib.v7i1.5575>.
- [14] Nurhusen, M. R., Indra, J., & Baihaqi, K. A. (2023). Analisis Sentimen Pengguna Twitter Terhadap Kenaikan Harga Bahan Bakar Minyak (BBM) Menggunakan Metode Logistic Regression. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7(1), 276-282.
- [15] Rahmawati, C., & Sukmasetya, P. (2022). Sentimen Analisis Opini Masyarakat Terhadap Kebijakan Kominfo atas Pemblokiran Situs non-PSE pada Media Sosial Twitter. *JURIKOM (Jurnal Riset Komputer)*, 9(5), 1393-1400. DOI: <http://dx.doi.org/10.30865/jurikom.v9i5.4950>.
- [16] Lestandy, M., Abdurrahim, A., & Syafa'ah, L. (2021). Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(4), 802-808. DOI: <https://doi.org/10.29207/resti.v5i4.3308>.
- [17] Astuti, W. (2022). ANALYSIS SENTIMENT ON THE ACCEPTANCE OF CPNS 2021 ON TWITTER SOCIAL MEDIA USING TEXTBLOB. *Jurnal Techno Nusa Mandiri*, 19(1), 15-20.
- [18] Kelvin, K., Banjarnahor, J., Nababan, M. N., & Sinurat, S. H. (2022). Analisis perbandingan sentimen Corona Virus Disease-2019 (Covid19) pada Twitter Menggunakan Metode Logistic Regression Dan Support Vector Machine (SVM). *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, 5(2), 47-52. DOI: <https://doi.org/10.34012/journalsisteminformasidanilmukomputer.v5i2.2365>.
- [19] Lengkong, N. C., Safitri, O., Machsus, S., Putra, Y. R., Syahadati, A., & Nooraeni, R. (2021). Analisis Sentimen Penerapan Psbb Di Dki Jakarta Dan Dampaknya Terhadap Pergerakan Ihsq. *Jurnal Teknoinfo*, 15(1), 20-25.
- [20] Onantya, I. D., Indriati, I., & Adikara, P. P. (2019). Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan BM25 Dan Improved K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(3), 2575-2580.
- [21] Naufal, M. I., & Purwanto, E. (2022). DAMPAK LITERASI KEUANGAN TERHADAP KINERJA DAN KEBERLANJUTAN UMKM (Studi Kasus Industri F&B Kecamatan Sumpersari Jember). *Profit: Jurnal Administrasi Bisnis*, 16(2), 209-215. DOI: <https://doi.org/10.21776/ub.profit.2022.-16.02.5>.



- [22] Purba, W., Siawin, W., Nababan, M. N., Dharshinni, N. P., & Aisyah, S. (2019). Implementasi Data Mining Untuk Pengelompokan Dan Prediksi Karyawan Yang Berpotensi Phk Dengan Algoritma K-Means Clustering. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, 2(2), 85-90. DOI: <https://doi.org/10.34012/jusikom.v2i2.429>.
- [23] Herdiana, Y., Rusdianto, D., & Geraldine, W. A. (2023). APLIKASI CV MATCHER UNTUK MELIHAT KECOCOKAN DAFTAR RIWAYAT HIDUP DENGAN LOWONGAN PEKERJAAN MENGGUNAKAN MACHINE LEARNING DAN METODE COSINE SIMILARITY BERBASIS WEB. *COMPUTING | Jurnal Informatika*, 10(01), 26-30.
- [24] Anam, S., Yanti, I., Fitriah, Z., & Habibah, U. (2021). *Cara Mudah Belajar Bahasa Pemrograman C++*. Universitas Brawijaya Press.