Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2607~2615
e-ISSN: 2548-6861

Sentiment Analysis of Economic Policy Comments on YouTube Using Ensemble Machine Learning

Kety Nandini [1]*, Majid Rahardi [2]*
* Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
ketynandini@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-08-25
Revised 2025-09-08
Accepted 2025-09-10

Keyword:

Sentiment Analysis,
Ensemble Learning,
Youtube Comments,
Economic Policy,
Machine Learning.

ABSTRACT

Public sentiment analysis of economic policies is increasingly important in the digital age, as social media platforms have become the main arena for public discussion. This study analyzes YouTube comments related to Tom Lembong's economic policies to address the lack of policy sentiment analysis tools in Indonesian. A dataset containing 1,029 comments was collected and systematically processed using normalization, stop word removal, and stemming techniques tailored to Indonesian. To overcome data scarcity and class imbalance, advanced data augmentation methods—synonym replacement, random insertion, and random deletion—were applied, expanding the dataset to 2,169 samples. Feature extraction used TF-IDF vectorization (unigram, bigram, trigram) and CountVectorizer, followed by an 80:20 split into training and testing sets. Several machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, Random Forest, Gradient Boosting, and Naïve Bayes, were evaluated with hyperparameter tuning through grid search. The results showed that SVM with TF-IDF bigrams achieved the best performance (accuracy: 96.08%, F1-score: 96.03%). Class-level evaluation showed high performance for negative sentiment (F1-score: 0.97) and positive sentiment (F1-score: 0.97), while neutral sentiment was more challenging (F1-score: 0.90) due to ambiguity, sarcasm, and fewer samples. The ensemble model, which combines several optimized SVM variants with soft voting, achieved robust and stable performance (accuracy and F1-score: 95.16%). These findings confirm the effectiveness of the ensemble-based approach for Indonesian sentiment analysis, while providing valuable insights into public perceptions of economic policy in the digital space.

I. PENDAHULUAN

Media sosial saat ini menjadi salah satu sarana utama bagi masyarakat untuk menyampikan pendapat mengenai kebijakan publik, terutama kebijakan ekonomi yang dikeluarkan oleh pemerintah[1]. Fenomena ini mendorong munculnya berbagai penelitian analisis sentimen terkait kebijakan publik di Indonesia. Penelitian oleh Nurbagja dkk. Menganalisis sentimen masyarakat terhadap kebijakan BBM di Twitter, mengungkapkan dominasi sentimen negatif yang mencerminkan ketidakpuasan publik [2]Sementara itu, penelitian oleh Imawan dkk. mengembangkan model klasifikasi sentimen untuk rencana kenaikan PPN 12% yang berhasil mencapai akurasi 83%, menunjukkan bahwa model klasifikasi dapat memberikan wawasan berbasis data untuk memahami respon public terhadap kebijakan pajak secara mendalam [3].

Salah satu platform media sosial yaitu YouTube sebagai salah satu platform berbagi video tidak hanya menjadi sarana hiburan, tetapi juga telah menjadi wadah diskusi bagi masyarakat untuk merespon isu-isu strategis, seperti kebijakan ekonomi. Pemilihan Tom Lembong sebagai objek penelitian tidak terlepas dari keterlibatannya dalam berbagai isu strategis kebijakan ekonomi Indonesia. Sebagai mantan Menteri perdagangan dan Ketua Badan Koordinasi Penanaman Modal (BKPM), keputusannya berperan penting

dalam mendorong iklim investasi dan liberalisasi perdagangan[4].

Analisis terhadap figur pembuat kebijakan seperti ini sejalan dengan penelitian Pristika dkk. yang menganalisis komentar masyarakat terhadap pelantikan Menteri Agraria dan Tata Ruang/Kepala Badan pertahanan Nasional pada akun instragram @agusyudhoyono dengan tujuan memahami respons publik terhadap figur pembuat kebijakan di media sosial secara mendalam [5]. Salah satu isu yang menimbulkan diskusi publik luas adalah kebijakan impor gula, yang oleh sebagian kalangan dipandang lebih mencerminkan orientasi ekonomi kapitalis dibandingkan dengan prinsip demokrasi ekonomi dan sistem ekonomi Pancasila [6]. Kebijakan semacam ini kerap menimbulkan pro dan kontra karena menyangkut kepentingan nasional, arus ivestasi, serta perlindungan terhadap petani lokal. Oleh karena itu, Tom Lembong menjadi figur publik yang relevan untuk dianalisis dalam konteks persepsi masyarakat.

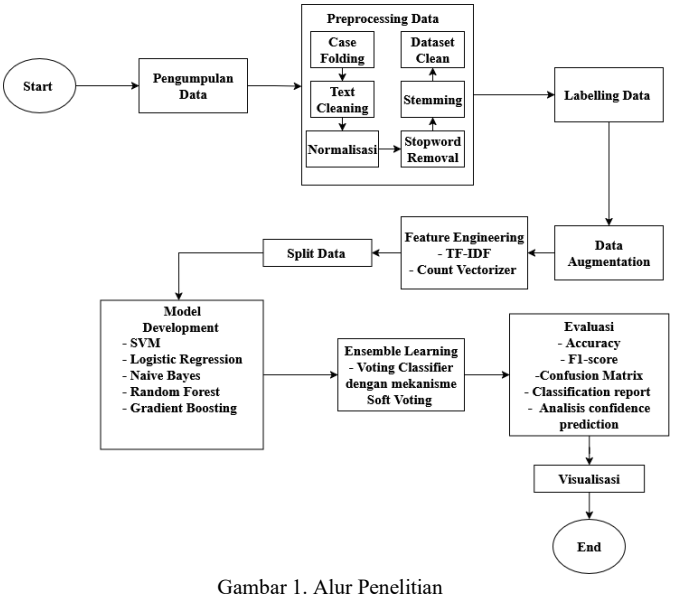
Permasalahan utama yang dihadapi adalah sulitnya menganalisis ribuan komentar YouTube secara manual untuk memahami sentimen publik terhadap kebijakan ekonomi, ditambah dengan kompleksitas bahasa Indonesia informal yang menggunakan banyak singkatan, slang, dan variasi penulisan yang memerlukan preprocessing khusus. Penggunaan algoritma tunggal juga menjadi keterbatasan dalam menangkap kompleksitas dan variasi dalam data teks komentar, sehingga diperlukan pendekatan yang lebih komprehensif untuk mencapai akurasi prediksi yang tinggi dan reliable[7].

Pada penelitian terdahulu telah membuktikan efektivitas machine learning seperti SVM, Naïve Bayes, dan Logistic Regression telah digunakan untuk membantu mengklasifikasikan sentimen dalam bahasa Indonesia[8]. Namun, penelitian yang dilakukan oleh Sari mengungkapkan bahwa model tunggal seperti Logistic Resression cenderung kurang stabil ketika mengadapi variasi kosakata informal di media sosial[9]. Sementara itu, hasil penelitian dari Mustofa menunjukkan bahwa pendekatan ensemble learning berbasis voting mampu memberikan peningkatan akurasi dalam klasifikasi sentimen produk jika dibandingkan dengan pendekatan model tunggal[10] [11]. Dalam penelian yang dilakukan oleh Daniati dkk. menyoroti pentingnya menerapkan kombinasi beberapa algoritma untuk mengatasi tantangan linguistik dalam Bahasa Indonesia [12]. Tantangan tersebut mencakup penggunaan singkatan, Bahasa tidak baku, serta variasi haya Bahasa yang umum ditemukan dalam komunikasi digital.

Meskipun berbagai penelitian telah dilakukan dalam bidang analisis sentimen bahasa Indonesia, masih terdapat gap yang signifikan. Pendekatan ensemble diharapkan dapat meningkatkan performa klasifikasi dengan meminimalkan kelemahan masing-masing metode dan mampu menghasilkan prediksi sentimen yang lebih akurat. Inovasi metodologis ini tidak hanya memberikan solusi untuk analisis kebijakan ekonomi, tetapi juga berkontribusi pada pengembangan Teknik NLP untuk teks media sosial berbahasa Indonesia.

Penelitian ini bertujuan mengembangkan sistem analisis sentimen yang akurat dan efisien untuk mengklasifikasi komentar publik terhadap kebijakan ekonomi Tom Lembong di YouTube. Kontribusi utama penelitian ini adalah pengembangan sistem ensemble yang dioptimalkan khusus untuk teks media sosial berbahasa Indonesia, dengan pipeline preprocessing adaptif dan teknik augmentasi data yang disesuaikan dengan karakteristik komentar di platform tersebut. Secara khusus, penelitian ini akan mengembangkan rangkaian preprocessing yang optimal untuk teks komentar YouTube berbahasa Indonesia dengan normalisasi, penghapusan stopword, dan stemming yang disesuaikan dengan karakteristik media sosial. Penelitian ini juga akan menerapkan metode ekstraksi fitur menggunakan TF-IDF untuk menangkap berbagai aspek informasi dari teks komentar. Teknik augmentasi data berupa penggantian sinonim, penyisipan acak, dan penghapusan acak akan diterapkan untuk mengatasi keterbatasan ukuran dataset dan ketidakseimbangan kelas. Beberapa model termasuk SVM, Naïve Bayes, Logistic Regression, Random Forest, dan Gradient Boosting akan dilatih dengan penyetelan parameter yang optimal untuk mengembangkan model ensemble yang unggul dalam mengklasifikasikan sentimen.

II. METODE



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Untuk memastikan relavansi konten dengan objek penelitian, data penelitian diperoleh melalui crawling komentar YouTube menggunakan YouTube Data API v3. Untuk digunakan sebagai sumber data, video yang berkaitan dengan kebijakan ekonomi Tom Lembong dipilih berdasarkan beberapa kriteria untuk memastikan bahwa kontennya relevan dengan subjek penelitian. Kriteria tersebut meliputi bahwa video tersebut membahas kebijakan ekonomi Tom Lembong, memiliki banyak komentar yang signifikan, dirilis dalam periode waktu yang relevan dengan penelitian, dan memiliki komentar yang terbuka untuk umum.

Pengumpulan data dilakukan dengan library *googleapiclient* dengan tahapan sebagai berikut. Pertama-tama, konfigurasi API dilakukan dengan menggunakan YouTube Data API v3, yang memungkinkan akses ke data komentar public melalui *developer key*. Selanjutnya, komentar diekstrasi, yang mencakup komentar utama dan balasan dari setiap video yang dimaksud. Selain itu, metadata tambahan dikumpulkan. Ini termasuk (1) *publishedAt* (tanggal publikasi komentar), (2) *authorDisplayName* (nama pengguna yang berkomentar), (3) *textDisplay* (konten teks komentar), dan (4) *likeCount* (jumlah like pada komentar).

B. Pre-Processing Data

Proses preprocessing dilakukan setelah data dikumpulkan untuk memastikan teks pelatihan model dalam format yang bersih dan siap untuk diproses[13]. Proses ini digunakan untuk memastikan kualitas data sebelum diproses lebih lanjut. Pada tahap ini semua huruf pada komentar diubah menjadi bentuk kecil (case folding) untuk menyamakan format penulisan[14]. Pada tahap ini, komentar dibersihkan dari tautan URL, mention, hashtag, angka, symbol, serta karakter non-alfabet yang tidak relevan. Komentar juga dibersihkan dari spasi ganda atau tanda baca yang berlebihan. Proses normalisasi dilakukan menggunakan kamus konversi kata untuk mengubah kata tidak baku, singkatan, bahasa gaul, dan bentuk penulisan tidak standar menjadi bentuk yang sesuai dengan kaidah bahasa Indonesia. Selanjutnya, stopword removal digunakan untuk menghapus kata-kata umum yang tidak memiliki makna penting dalam analisis dan penggunaan Teknik stemming menggunakan Pustaka sastrawi untuk mengembalikan kata berimbuhan ke bentuk dasar.

C. Labelling Data

Penelitian ini menggunakan pendekatan berbasis lexicon untuk pelabelan sentimen. Metode ini menggunakan kamus sentimen berbahasa Indonesia, yang mengandung daftar kata dan bobot sentimen, yang diwakili dengan nilai positif, negatif, atau netral. Skor sentimen setiap komentar dihitung dengan menjumlahkan bobot kata yang sesuai dalam kamus. Pelabelan dilakukan dalam beberapa Langkah. Pertama, teks komentar ditokenisasi menggunakan fungsi nltk.word_tokenize() untuk memecah kalimat menjadi token kata. Selanjutnya formula tertentu dilakukan untuk menghitung skor sentimen total, yang dihitung dengan menjumlahkan bobot sentimen dari setiap kata yang ditemukan dalam komentar. Label sentimen untuk setiap komentar secara otomatis ditetapkan bersdasarkan hasil akhir dari proses ini.

Untuk memastikan akurasi, hasil pelabelan otomatis diverifikasi secara manual oleh annotator manusia. Annotator memeriksa sampel komentar yang telah diberi label oleh sistem, memperbaiki kesalahan, dan menyesuaikan label yang ambigu. Tahap ini penting untuk mengatasi keterbatasan kamus *lexicon*, seperti konteks kalimat yang tidak tertangkap atau kata-kata yang tidak tercantum dalam kamus.

D. Data Augmentation

Untuk mengatasi keterbatasan jumlah data dan masalah *class imbalance*, diterapkan strategi data augmentation[15]. Dalam penelitian ini digunakan teknik *Easy data Augmentation* (EDA), sebuah pendekatan yang efektif untuk meningkatkan kinerja tugas klasifikasi teks. Implementasinya meliputi tiga teknik yaitu synonym replacement (mengganti kata dengan sinonim yang sesuai), random insertion (menambahkan sinonim pada posisi acak di dalam kalimat), serta random deletion (menghapus kata tertentu secara acak). Setiap teks hasil augmentasi dicek agar berbeda dari teks aslinya, diberi label sesuai metode yang digunakan, dan tetap mempertahankan keseimbangan label sentimen. Setiap sample teks yang dihasilkan dari proses augmentasi kemudian melalui tahap validasi untuk menjamin bahwa teks telah termodifikasi dan berbeda dari teks sumber. Kriteria kunci yang dijaga adalah konsistensi label sentimen asli dan keseimbangan jumlah sampel antar kelas. Teknik ini tidak hanya memperbanyak data tetapi juga meningkatkan keragaman linguistik dataset secara artifisial. Dengan cara ini, dataset menjadi lebih beragam dan model diharapkan mampu belajar pola bahasa yang lebih kompleks dengan lebih baik serta memiliki kemampuan generalisasi yang lebih tinggi.

E. Feature Engineering

Pada tahap feature engineering, teks akan diubah menjadi representasi numerik agar dapat diproses oleh algoritma machine learning. Seperti yang dilakukan dalam penelitian sebelumnya, data non numerik seperti jenis kelamin dan tingkat Pendidikan dikonversi menajdi representasi numerik untuk memungkinkan pemprosesan oleh model machine learning[16]. Dalam penelitian ini digunakan dua pendekatan utama, yaitu *Term Frequency-inverse Document Frequency (TF-IDF)* dan *CountVectorizer.* Metode TF-IDF diterapkan pada level unigram, biagram, dan trigram sehingga mampu menangkap makna dari kata tunggal, pasangan kata, hingga rangkaian tiga kata yang berurutan. Di sisi lain, *CountVectorizer* berfungsi untuk mengubah teks menjadi representasi numerik berdasarkan jumlah kemunculan kata. Penggabungan kedua pendekatan ini memungkinkan analisis yang lebih menyeluruh terhadap ciri khas teks dalam dataset.

F. Split Data

Pada tahap ini, dataset dibagi menjadi data latih dan data uji. Proses ini dilakukan agar model dapat dievaluasi secara objektif menggunakan data yang sama sekali baru bagi model. Penelitian ini mengalokasikan 80% data untuk pelatihan dan 20% untuk pengujian. Pembagian dilakukan dengan menerapkan parameter stratifikasi untuk mempertahankan proporsi kelas seimbang pada kedua subset. Nilai random_state ditentukan agar proses pembagian dara dapat direproduksi dengan hasil yang konsisten. Melaui tahapan ini, model dapat diuji kemampuannya dalam mengklasfikasikan sampel baru dengan lebih andal sekaligus menhindari risiko overfitting.

G. Model Development

Pada tahap pengembangan model, penelitian ini menggunakan beberapa algoritma secara umum digunakan dalam klasifikasi teks, yaitu Support Vektor Machine (SVM), Logistic Regression, Naïve Bayes, Random Forest, dan Gradient Boosting. Pemilihan algoritma tersebut didasarkan pada keunggulan masing-masing dalam menangani permasalahan klasifikasi bebasis teks. SVM dipilih karena mampu bekerja optimal pada data berdimensi tinggi dengan memaksimalkan margin pemisah antar kelas, sementara Logistic Regression digunakan sebagai model dasar yang sederhana namun efektif dalam memodelkan probabilitas kelas[17]. Naïve Bayes dipertimbangkan karena kesedehanaan serta kemampuannya dalam mengelola data berbasis frekuensi kata dengan asumsi independensi fitur[18]. Random Forest digunakan sebagai algoritma berbasis ensemble yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang stabil, sedangkan Gradient Boosting dipilih karena mampu meningkatkan akurasi melelui proses boosting yang memperbaiki kesalahan dari model sebelumnya[19].

Pengembangan model dilakukan dengan memanfaatkan representasi fitur hasil *feature engineering*. Proses pelatihan model dilakukan dengan skema *stratifield k-fold cross validation* untuk menjaga distribusi kelas tetap seimbang di setiap lipatan data. Selain itu, dilakukan optimasi parameter menggunakan GridSearchCV untuk menemukan konfigurasi hyperparameter yang optimal pada setiap algoritma. Proses ini bertujuan untuk menigkatkan efektivitas model sehingga mampu menghasilkan prediksi sentimen yang lebih presisi.

H. Ensemble Learning

Selain menggunakan model tunggal, penelitian ini juga menerapkan pendekatan ensemble learning untuk meningkatkan klasifikasi. Pendekatan ini mengkombinasikan berbagai algoritma klasifikasi untuk menghasilkan prediksi yang lebih stabil dan akurat dibandingkan penggunaan model tunggal. Secara spesifik, diterapkan Voting Classifier dengan mekanisme soft voting yang mengintegrasikan prediksi dari beberapa base learners. Soft voting menentukan kelas kahir berdasarkan rata-rata probabilitas prediksi tertinggi dari seluruh model.

Pada tahap ensemble, penelitian ini menggabungkan empat model SVM terbaik yang diperoleh dari hasil vektorisasi berbeda, yaitu TF-IDF unigram, bigram, trigram, dan CountVectorizer. Setiap model SVM dilatih menggunakan *Grid Search* dengan beberapa parameter yang disesuaikan meliputi C (0.1, 1, 10), kernel (linear, rbf), dan gamma (scale, auto). Dengan ini, ensemble terbentuk dari kumpulan model SVM yang telah dioptimalkan, sehingga dapat menangkap variasi pola pada data sekaligus meningkatkan konsistensi hasil prediksi.

I. Evaluasi

Tahap evaluasi model dilakukan untuk mengukur kinerja algoritma klasifikasi yang digunakan dalam penelitian. Evaluasi ini menggunakan data uji yang telah dipisahkan sebelumnya dari dataset agar hasil pengujian lebih objektif dan tidak bias. Beberapa matrik evaluasi yang digunakan untuk menilai kinerja model klasifikasi sentimen. Akurasi (accuracy) digunakan sebagai ukuran dasar untuk mengevaluasi presentase prediksi yang benar secara keseluruhan[20]. Selain itu, precision dan recall dihitung untuk masing-masing kelas, memberikan informasi tentang kempuan model dalam menghasilkan prediksi yang banar (precison) dan menangkap semua sampel relevan dari kelas tertentu (recall) .

Selain itu, F1-score dengan pembobotan diterapkan untuk menangani ketidakseimbangan kelas, menggabungkan precision dan recall dalam satu metrik yang seimbang. Analisis lebih mendalam dilakukan melalui confusion matrix yang memvisualisasi true positive, true negative, false positive, dan false negative, memberikan gambaran jelas tentang keunggulan dan kelemahan model dalam mengklasifikasikan setiap kelas sentiment. Classification report juga disertakan untuk menyajikan precision, recall, dan F1-score per kelas beserta jumlah sampel (support), menungkinkan evaluasi performa model secara lebih rinci pada setiap kategori sentimen.

Selain menggunakan metrik evaluasi konvensional (Accuracy, Precision, Recall, dan F1-Score), penelitian ini juga melakukan uji signifikansi statistic untuk membandingkan performa model ensemble dengan baseline. Uji yang digunakan adalah McNemar-s test, yang umum dipakai untuk menguji perbedaan proporsi kesalahan klasifikasi antara dua model pada data yang sama. Dengan tingkat signifikansi $\alpha = 0.05$, uji ini membantu memastikan apakah perbedaan performa yang diperoleh signifikan secara statistik atau hanya disebabkan oleh kebetulan.

III. HASIL DAN PEMBAHASAN

A. Pemgumpulan Data

Dataset yang digunakan dalam penelitian ini dikumpulkan dari komentar teks yang diperoleh melalui platform YouTube terkait isu kebijakan ekonomi Tom Lembong. Komentar-komentar tersebut dipilih karena dianggap representative dalam menggambarkan opini publik yang beragam. Proses pengumpulan dilakukan dengan memanfaatkan metode scraping menggunakan library python, sehingga data dapat diekstraksi secara otomatis dan tersimpan dalam format CSV. Jumlah dataset yang terkumpul adalah 1029 komentar.

TABEL I
CONTOH DATASET HASIL CRAWLING

publishedAt	authorDisplay Name	textDisplay	likeCount
2025-08-05T06:38:31Z	@MasD7	Pak tom.. ingin perpanjang	0

		perkara. Sudah jelas menguntungkan pihak tertentu, merugikan negara	
2025-08-04T08:35:48Z	@ariadwipangga-h7q	Salam Bahagia 😊 Yang bermasalah adalah hak abolisi dipakai setelah putusan amar hakim ditetapkan di persidangan setelah hakim memutuskan hukuman yaitu 4,5 tahun 😳😳😳😳	0

B. Preprocessing Data

Proses preprocessing data dilakukan untuk menyiapkan data agar lebih bersih dan konsisten sebelum digunakan dalam pemodelan. Tahapan ini meliputi case folding, penghapusan URL, mention, hasthtag, dan karakter khusus, normalisasi kata tidak baku atau slang menggunakan kamus normalisasi, stopword removal untuk membuang kata-kata yang tida memiliki makna signifikan, serta stemming dengan sastrawi untuk mengubah kata ke bentuk dasarnya.

TABEL 2
DATASET HASIL PREPROCESSING

Sebelum preprocessing	Setelah preprocessing
Hakim 💩💩💩 rudal aja rumah nya Pake SEPITENK. makan duit Haram tuh hakim ama Jaksa. ASN FEODAL G0BLK	hakim rudal rumah sepiteng makan duit haram hakim jaksa asn feodal goblok
Setelah melihat vonis Bapak Tom Lembong, makin yakin kalau hukum di Negeri ini sudah Tidak Baik baik saja. RIP Hukum RI😳	Lihat vonis tom lembong yakin hukum negeri tidak baik rip hukum ri
capek g sih ges kalian hidup di negara yang pemerintahannya kaya gini... 	Capek hidup negara pemerintah kaya gini

C. Labelling Data

Pelabelan sentimen dilakukan dengan menggunakan pendekatan berbasis lexicon dan diverifikasi oleh annotator, diperoleh distribusi komentar ke dalam tiga kategori sentiment, yaitu positif, negatif, dan netral. Hasil analisis menunjukkan bahwa sebagain besar komentar cenderung bernuansa negatif, disusul oleh komentar positif, sementara komentar netral muncul dalam jumlah yang lebih sedikit. Komentar negatif umumnya terkait dengan kritik terhadap kebijakan maupun ketidakpuasan terhadap kondisi pemerintahan, seperti penggunaan kata "korupsi", "jahat", atau "tidak adil". Sebaliknya, komentar positif banyak memuat apresiasi dan dukungan, misalnya dengan kata "bagus", "baik", atau "adil". Adapun komentar netral biasanya berupa penyampaian informasi atau opini tanpa ekspresi emosi yang kuat. Verifikasi manual juga menunjukkan bahwa tingkat kesesuaian hasil pelabelan otomatis cukup tinggi, meskipun beberapa koreksi masih diperlukan terutama pada komentar yang mengandung konteks sarkastik atau kata-kata yang tidak tercantum dalam kamus lexicon. Hasil pelabelan ini memberikan gambaran yang jelas mengenai kecenderungan opini publik dalam dataset yang dianalisis.

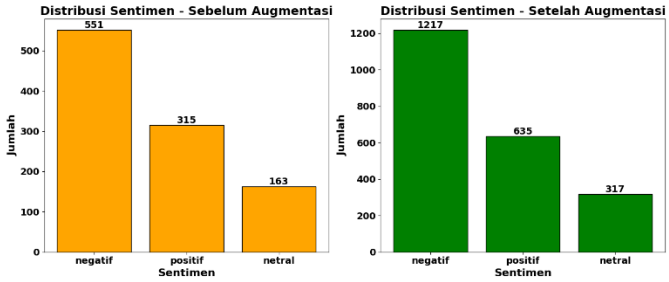
TABEL 3
DISTRIBUSI SENTIMEN

Label	Jumlah
Negatif	551
Positif	315
Netral	163
Total	1029

D. Data Augmentation

Untuk mengatasi permasalahan ketidakseimbangan kelas pada dataset, penelitian ini menerapkan teknik augmentasi data berbasis teks yang terdiri atas tiga metode utama, yaitu *synonym replacement, random insertion,* dan *random deletion.* Metode *synonym replacement* dilakukan dengan mengganti kata tertentu menggunakan sinonimnya, *random insertion* dilakukan dengan menyisipkan sinonim ke dalam teks secara acak, sedangkan *random deletion* dilakukan dengan menghapus kata-kata pada posisi tertentu dengan probabilitas tertentu. Setiap komentar berpotensi menghasilkan antara satu hingga empat variasi baru, tergantung Panjang teks, ketersediaan sinonim dalam kamus WordNet, serta hasil dari proses penghapusan kata secara acak. Dengan demikian, jumlah data hasil augmentasi tidak bersifat tetap, tetapi umumnya lebih besar dari dataset awal. Berbeda dengan metode *oversampling* yang hanya menggandakan data lama, pendekatan ini menghasilkan variasi baru dari komentar asli, sehingga meningkatkan jumlah data sekaligus memperkaya pola bahasa yang dapat dipelajari oleh model.

Sebelum proses augmentasi, distribusi kelas tidak seimbang dengan 551 komentar negative, 315 komentar positif, dan 163 komentar netral. Setelah proses augmentasi, jumlah data meningkat dari meningkat dari 1029 komentar menjadi 2169 komentar, dengan distribusi yang lebih proporsional yaitu 1217 komentar negatif, 635 komentar positif, dan 317 komentar netral. Gambar 2 memperlihatkan persebaran data sebelum dan setelah proses augmentasi.



Gambar 2. Distribusi Data Augmentasi

TABEL 4
DISTRIBUSI DATA DENGAN TIGAS METODE AUGMENTASI

Augmentation	Jumlah
Original	1029
Deletion	635
Synonym	389
Insertion	116

E. Feature Engineering

Data penelitian ini, ekstraksi fitur teks dilakukan menggunakan empat teknik vektorisasi, yaitu TF-IDF Unigram dengan maksimal 5.000 fitur, TF-IDF Bigram dengan maksimal 8.000 fitur, TF-IDF Trigram dengan maksimal 10.000 fitur, serta Count Vectorizer dengan maksimal 5.000 fitur. Setiap vectorizer dikonfigurasi dengan parameter min_df =2, sehingga hanya mempertahankan kata yang muncul minimal dua kali, serta max_df=0.8 untuk menghapus kata yang terlalu sering muncul pada lebih dari 80% dokumen. Pengaturan ini bertujuan mengurangi keberadaan kata yang terlalu jarang maupun terlalu umum, sehingga fitur yang dipilih lebih representatif dan relevan dalam mendukung analisis sentimen.

F. Split Data

Sebelum dilakukan proses data augmentasi, dataset terdiri atas 1029 komentar dengan distribusi kelas yang tidak seimbang. Untuk memperkaya variasi data dan mengurangi ketidakseimbangan kelas, dilakukan augmentasi menggunakan beberapa teknik berbasis modifikasi teks.

TABEL 5
DISTRIBUSI DATASET TRAINING

Label	Jumlah
Negatif	973
Positif	508
Netral	254

Setelah proses ini, jumlah data meningkat menjadi 2.169 komentar. Selanjutnya, dataset dibagi menjadi training set dan test set dengan perbandingan 80:20. Dari total data sebanyak 1.735 komentar digunakan sebagai training set dan 434 komentar sebagai test set.

G. Evaluasi Model

Evaluasi model dilakukan menggunakan 434 sampel pada test set (20% dari 2169 total sampel). Hasil eksprimen menunjukkan bahwa SVM dengan konfigurasi TF-IDF Bigram memperoleh performa terbaik dengan akurasi 96.08% dan F1-score 96.03% diikuti oleh SVM TF-IDF Unigram dengan akurasi 96.08% dan F1-score 96.02%. SVM TF-IDF Trigram menempati posisi ketiga dengan akurasi 95.85% dan F1-score 95.80%. Secara konsisten, SVM memberikan hasil paling stabil pada berbagai teknik representasi fitur, khususnya dengan parameter optimal (C=10, kernel=rbf, gamma=scale). Logistic Regression menunjukkan hasil konpetitif dengan performa tertinggi pada CountVectorizer (akurasi 94.01%), sementara Random Forest, Gradient Boosting, dan Naïve Bayes menunjukkan kinerja yang lebih rendah dengan rentang akurasi 86.87% - 93.55%.

TABEL 6
MODEL COMPARISON

Model	Vectorizer	Accuracy	Pricision	Recall	F1-Score
SVM	TF-IDF_bigram	0.9608	0.9608	0.9608	0.9603
SVM	TF-IDF_unigram	0.9608	0.9609	0.9608	0.9602
SVM	TF-IDF_trigram	0.9585	0.9584	0.9585	0.9580
Voting Ensemble	Multi-Vectorizer	0.9516	0.9517	0.9516	0.9516
Logistic Regression	count_vec	0.9401	0.9405	0.9401	0.9402
Logistic Regression	TF-IDF_bigram	0.9401	0.9407	0.9401	0.9393
SVM	count_vec	0.9401	0.9409	0.9401	0.9392
Random Forest	TF-IDF_unigram	0.9355	0.9358	0.9355	0.9355
Logistic Regression	TF-IDF_trigram	0.9355	0.9364	0.9355	0.9347
Logistic Regression	TF-IDF_unigram	0.9332	0.9327	0.9332	0.9323
Random Forest	TF-IDF_bigram	0.9309	0.9324	0.9309	0.9312
Random Forest	count_vec	0.9263	0.9263	0.9263	0.9261
Gradient Boosting	count_vec	0.9263	0.9317	0.9263	0.9241
Random Forest	TF-IDF_trigram	0.9240	0.9249	0.9240	0.9237

Gradient Boosting	TF-IDF_unigram	0.9124	0.9133	0.9124	0.9119
Gradient Boosting	TF-IDF_trigram	0.9032	0.9031	0.9032	0.9022
Gradient Boosting	TF-IDF_bigram	0.8986	0.8999	0.8986	0.8991
Naive Bayes	TF-IDF_bigram	0.8825	0.8856	0.8825	0.8768
Naive Bayes	count_vec	0.8687	0.8676	0.8687	0.8662
Naive Bayes	TF-IDF_unigram	0.8710	0.8789	0.8710	0.8642
Naive Bayes	TF-IDF_trigram	0.8687	0.8737	0.8687	0.8626

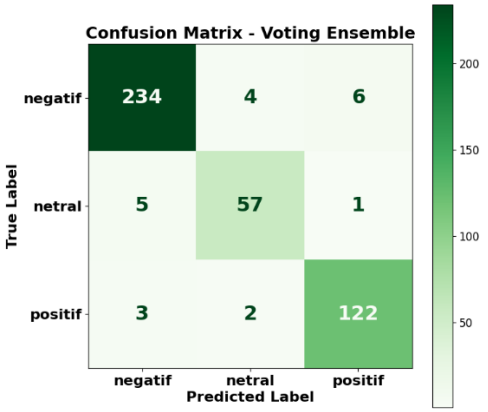
Untuk meningkatkan performa lebih lanjut, dibangun sebuah model ensemble yang menggabungkan beberapa model SVM terbaik dari berbagai skema vektorisasi (unigram, bigram, trigram, dan count vectorizer) menggunakan strategi soft voting. Model ini berhasil meningkatkan performa akurasi 95.16% dan F1-score 95.16%. Model ensemble menunjukkan karakteristik yang berbeda dibandingkan dengan model individual terbaik. Meskipun akurasi ensemble (95.16%) sedikit lebih rendah dibandingkan SVM TF-IDF Bigram terbaik (96.08%), ensemble memberikan keunggulan dalam hal stabilitas dan konsistensi prediksi.

Analisis detail per kelas menunjukkan, sentimen negatif pada model ensemble mencapai F1-score 0.96 (precision: 0.97, recall: 0.96), sedikit lebih rendah dari individual best (F1-score 0.97), namun menunjukkan *balance* yang lebih baik antara precision dan recall. Pada sentimen positif model ensemble mencapai F1-score 0.95 (precision: 0.95, recall: 0.96). Sedangkan pada sentiment netral, model ensemble mencapai F1-score 0.90 (precison: 0.90, recall:0.90), hal ini menunjukkan *perfect balance* antara precision dan recall, yang merupakan keunggulan signifikan dibandingkan model individual yang cenderung memiliki recall rendah untuk kelas netral. Hasil ini mengindikasikan bahwa sentiment netral merupakan kelas yang paling *challenging* untuk diklasifikasikan, diduga karena jumlah sampelnya yang lebih sedikit dan sifatnya yang ambigu. Selain itu, gaya bahasa sarkastik yang umum di media social dapat menimbulkan kesalahan interpretasi. Analisis kinerja klasifikasi per kelas menungungkapkan bahwa model ensemble menunjukkan performa yang seimbang pada ketika kelas sentimen.

📌 Model: Voting Ensemble | Vectorizer: Multi-Vectorizer

	precision	recall	f1-score	support
negatif	0.97	0.96	0.96	244
netral	0.90	0.90	0.90	63
positif	0.95	0.96	0.95	127
accuracy			0.95	434
macro avg	0.94	0.94	0.94	434
weighted avg	0.95	0.95	0.95	434

Gambar 3. Classification Report Model Ensemble



Gambar 4. Confusion Matrix Model Ensemble

📌 SVM | Vectorizer: tfidf_unigram

	precision	recall	f1-score	support
negatif	0.95	0.99	0.97	244
netral	0.95	0.84	0.89	63
positif	0.98	0.96	0.97	127
accuracy			0.96	434
macro avg	0.96	0.93	0.94	434
weighted avg	0.96	0.96	0.96	434

📌 Naive Bayes | Vectorizer: tfidf_bigram

	precision	recall	f1-score	support
negatif	0.86	0.97	0.91	244
netral	0.90	0.57	0.70	63
positif	0.93	0.87	0.90	127
accuracy			0.88	434
macro avg	0.90	0.80	0.84	434
weighted avg	0.89	0.88	0.88	434

📌 Logistic Regression | Vectorizer: tfidf_bigram

	precision	recall	f1-score	support
negatif	0.94	0.97	0.95	244
netral	0.96	0.81	0.88	63
positif	0.94	0.95	0.95	127
accuracy			0.94	434
macro avg	0.95	0.91	0.93	434
weighted avg	0.94	0.94	0.94	434

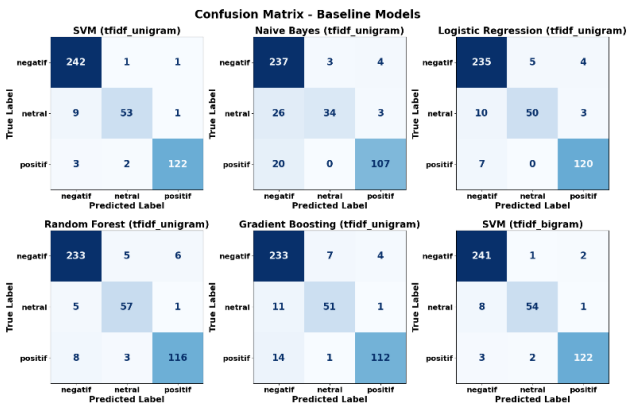
📌 Random Forest | Vectorizer: tfidf_unigram

	precision	recall	f1-score	support
negatif	0.95	0.95	0.95	244
netral	0.88	0.90	0.89	63
positif	0.94	0.91	0.93	127
accuracy			0.94	434
macro avg	0.92	0.92	0.92	434
weighted avg	0.94	0.94	0.94	434

📌 Gradient Boosting | Vectorizer: count_vec

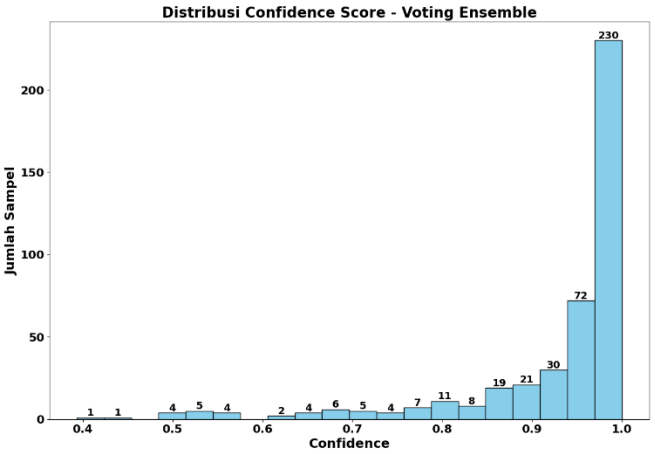
	precision	recall	f1-score	support
negatif	0.90	0.99	0.94	244
netral	1.00	0.71	0.83	63
positif	0.97	0.91	0.94	127
accuracy			0.93	434
macro avg	0.95	0.87	0.90	434
weighted avg	0.93	0.93	0.92	434

Gambar 5. Classification Report Model Ensemble



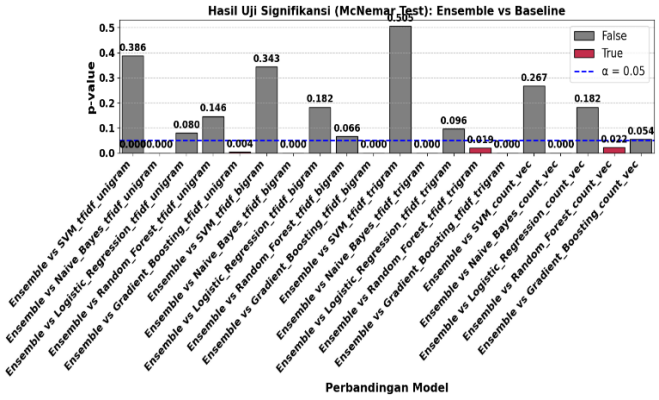
Gambar 7. Baseline Model Confusion Matrix

Distribusi confidence score menunjukkan bahwa sebagian besar prediksi (85.5%) memiliki tingkat kepercayaan di atas 0.8, dengan hanya 4.7% prediksi yang memiliki confidence di bawah 0.6.



Gambar 5. Analisis Confidance Prediksi

Hasil uji McNemar menunjukkan bahwa ensemble memiliki perbedaan performa yang signifikan ($p < 0.05$) dibandingkan dengan beberapa baseline (misalnya Naïve Bayes dan Gradient Boosting), namun tidak signifikan ($p > 0.05$) jika dibandingkan dengan SVM dan Logistic Regression. Hal ini menandakan bahwa keunggulan ensemble tidak seragam terhadap semua baseline, melainkan bergantung pada algortima pembandingnya.



Gambar 8. Hasil uji McNemar

IV. KESIMPULAN

Dengan memanfaatkan ensemble learning, penelitian ini dapat mengevaluasi tanggapan masyarakat terhadap kebijakan ekonomi Tom Lembong yang terefleksi dalam komentar YouTube. Melalui penerapan teknik preprocessing yang disesuaikan dengan karakteristik bahasa Indonesia, augmentasi data,serta ekstraksi fitur TF-IDF, kualitias dataset berhasil ditingkatkan secara signifikan dari 1029 menjadi 2169 komentar. Hasil evaluasi menunjukkan bahwa metode bahwa Support Vector Machine (SVM) dengan pendekatan TF-IDF Bigram menghasilkan kinerja tertinggi dengan

akurasi 96.08% dan F1-score 96.03%, mengungguli algoritma lain seperti Logistic Regression, Random Forest, Gradient Boosting, dan Naïve Bayes. Model ensemble berbasis soft voting yang menggabungkan beberapa variasi SVM mencapai akurasi 95.16% dan F1-score 95.16%, mendemonstrasikan bahwa meskipun tidka selalu melampaui model individual terbaik, pendekatan ensemble tetap memberikan stabilitas dan konsistensi prediksi yang baik. Penelitian ini memberikan kontribusi dalam pengembangan model NLP untuk bahasa Indonesia hal ini dapat dimanfaatkan sebagai alat bantu untuk memehami respons publik terhadap analisis sentimen, sehingga mendukung proses pengambilan keputusan yang lebih reposif dan bebasis data empiris.

DAFTAR PUSTAKA

[1] Y. Waskithoaji, A. Darmawan, J. Manajemen, F. Bisnis, and D. Ekonomika, "Peran Teknologi dalam Penggunaan Media Sosial dan Dampaknya terhadap UMKM," 2022. [Online]. Available: https://journal.uii.ac.id/selma/index

[2] K. Nurbagja, N. Saputra, A. Riyadi, and M. N. Tentua, "Sentiment Analysis of the Increase in Fuel Prices Using Random Forest Classifier Method," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 5, no. 1, pp. 132–144, Mar. 2023, doi: 10.12928/biste.v5i1.7414.

[3] Ferdian Imawan, Diqy Fakhrun Shiddieq, and Fikri Fahru Roji, "Analisis Sentimen Publik di X Terhadap Rencana Kenaikan PPN 12% Menggunakan Bert," *CESS (Journal of Computer Engineering, System and Science)*, vol. 10, no. 1, pp. 136–148, Jan. 2025, doi: 10.24114/cess.v10i1.65884.

[4] BBC NEWS INDONESIA, "Kebijakan impor gula enam mendag era Jokowi – Apa yang terjadi saat Tom Lembong menjabat?," BBC News Indonesia.

[5] F. Juma Pristika and F. Rozi, "Komunika: Jurnal Ilmu Komunikasi Sentimen Komentar Netizen dalam Postingan Pelantikan Menteri ATR/BPN pada Akun Instagram @agusyudhoyono," vol. 11, 2024, doi: 10.22236/komunika.v11i2.15145.

[6] Tokoh.co.id, "Tom Lembong: Arsitek Kebijakan dan Reformasi Ekonomi Indonesia," Tokoh.co.id. Accessed: Aug. 15, 2025. [Online]. Available: https://tokoh.co.id/tom-lembong-arsitek-kebijakan-ekonomi-indonesia/

[7] A. Y. Setiawan, I. Gede, M. Darmawiguna, and G. A. Pradnyana, "Sentiment Summarization Evaluasi Pembelajaran Menggunakan Algoritma Lstm (Long Short Term Memory)," *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, vol. 11, no. 2, 2022.

[8] A. Kartika Sari, Akhmad Irsyad, Dinda Nur Aini, Islamiyah, and Stephanie Elfriede Ginting, "Analisis Sentimen Twitter Menggunakan Machine Learning untuk Identifikasi Konten Negatif," *Adopsi Teknologi dan Sistem Informasi (ATASI)*, vol. 3, no. 1, pp. 64–73, Jun. 2024, doi: 10.30872/atasi.v3i1.1373.

[9] O. N. Cahyani and F. Budiman, "Performa Logistic Regression dan Naive Bayes dalam Klasifikasi Berita Hoax di Indonesia," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 60–68, Apr. 2025, doi: 10.29408/edumatic.v9i1.28987.

[10] Y. A. Mustofa, I. Surya, and K. Idris, "Pendekatan Ensemble pada Analisis Sentimen Ulasan Aplikasi Google Play Store Ensemble Approach to Sentiment Analysis of Google Play Store App Reviews," *Jambura Journal of Electrical and Electronics Engineering*, vol. 6 nomor 3, Jul. 2024.

[11] A. Mohammed and R. Kora, "An effective ensemble deep learning framework for text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8825–8837, Nov. 2022, doi: 10.1016/j.jksuci.2021.11.001.

[12] E. Daniati and H. Utama, "Analisis Sentimen Dengan Pendekatan Ensemble Learning Dan Word Embedding Pada Twitter," 2023.

[13] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *SMATIKA JURNAL*, vol. 10, no. 02, pp. 71–76, Dec. 2020, doi: 10.32664/smatika.v10i02.455.

[14] M. U. Albab, Y. K. P., and M. N. Fawaiq, "Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic," *Jurnal Transformatika*, vol. 20, no. 2, pp. 1–12, Jan. 2023, doi: 10.26623/transformatika.v20i2.5374.

[15] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst Appl*, vol. 244, p. 122778, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.

[16] H.-N. Huang *et al.*, "Employing feature engineering strategies to improve the performance of machine learning algorithms on echocardiogram dataset," *Digit Health*, vol. 9, Jan. 2023, doi: 10.1177/20552076231207589.

[17] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, p. 12, Dec. 2020, doi: 10.1007/s41133-020-00032-0.

[18] D. Pradana and E. Sugiharti, "Implementation Data Mining with Naive Bayes Classifier Method and Laplace Smoothing to Predict Students Learning Results," *Recursive Journal of Informatics*, vol. 1, no. 1, pp. 1–8, Mar. 2023, doi: 10.15294/rji.v1i1.63964.

[19] R. I. Arumnisaa and A. W. Wijayanto, "SISTEMASI: Jurnal Sistem Informasi Perbandingan Metode Ensemble Learning: Random Forest, Support Vector Machine, AdaBoost pada Klasifikasi Indeks Pembangunan Manusia (IPM) Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI)." [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[20] I. G. T. Isa and F. Elfaladonna, "Penilaian Kinerja Akurasi Metode Klasifikasi dalam Dataset Penerimaan Mahasiswa Baru Universitas XYZ," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 8, no. 2, p. 292, Aug. 2022, doi: 10.26418/jp.v8i2.54316.