

Analisis Sentimen Pada Komentar Youtube Dalam Turnamen MPL *Season 13* Dengan Metode *Ensemble Machine Learning*

Zendhi Yuna Burnama¹, Mochamad Alfian Rosid*², Nuril Lutvi Azizah³

^{1,2,3}Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo
e-mail: ¹zendhiyb10@gmail.com, *² alfianrosid@umsida.ac.id, ³ nurillutviazizah@umsida.ac.id

Abstrak

Analisis sentimen terhadap komentar pada video YouTube yang berhubungan dengan MPL *Season 13* dilakukan menggunakan metode klasifikasi berbasis *Ensemble Learning*. Penelitian ini berfokus pada identifikasi pola sentimen dalam komentar serta penentuan popularitas tim berdasarkan dukungan positif dari penggemar. Metode yang digunakan mencakup perencanaan awal, pengumpulan data melalui teknik *scraping* menggunakan YouTube Data *API v3*, dan proses *preprocessing*. Dari total 6.424 komentar yang dikumpulkan, jumlah komentar yang relevan berkurang menjadi 5.185 setelah tahap *cleaning*, *case folding*, penghapusan *stopword*, konversi *slang*, *stemming*, dan tokenisasi, dengan rincian 3.131 komentar positif dan 2.064 negatif. Berbagai metode klasifikasi diterapkan secara bersamaan dan digabungkan menggunakan teknik *ensemble machine learning* dengan pendekatan *majority voting*. Sebelum klasifikasi, data diproses menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*) untuk mengatasi ketidakseimbangan kelas. Hasil pengujian menunjukkan bahwa akurasi metode *Voting hard* mencapai 86,70 % (data latih 90% dan data uji 10%), sedangkan *voting soft* mencapai 86,17%. Proses pelabelan dilakukan dengan *library flair*, dengan validasi hasil klasifikasi melalui *confusion matrix*. Penerapan metode pelabelan yang menggabungkan pendekatan otomatis dan manual berhasil meningkatkan akurasi klasifikasi serta meminimalkan potensi kesalahan. Selain itu, analisis ini berhasil mengidentifikasi dukungan terbanyak, yaitu 877 pendukung EVOS, diikuti oleh RRQ dan ONIC dengan masing-masing 743 dan 556 pendukung. Penelitian ini diharapkan dapat memberikan kontribusi signifikan bagi pengembangan analisis sentimen dalam konteks *e-sports* dan membuka peluang untuk analisis lebih lanjut di penelitian mendatang.

Kata Kunci: Sentimen, *Ensemble Machine Learning*, *Preprocessing*, Klasifikasi, MPL *Season 13*

Sentiment Analysis on YouTube Comments in MPL Season 13 Tournament Using Ensemble Machine Learning Method

Abstract

Sentiment analysis of comments on YouTube videos related to MPL Season 13 was conducted using an Ensemble Learning-based classification method. This study focuses on identifying sentiment patterns in comments and determining team popularity based on positive fan support. The methods employed include initial planning, data collection through scraping techniques using the YouTube Data API v3, and preprocessing steps. From a total of 6,424 comments collected, the number of relevant comments was reduced to 5,185 after the cleaning, case folding, stopword removal, slang conversion, stemming, and tokenization stages, resulting in 3,131 positive comments and 2,064 negative comments. Various classification methods were applied simultaneously and combined using ensemble machine learning techniques with a majority voting approach. Before classification, the data was processed using SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance. The testing results showed that the hard voting method achieved an accuracy of 86,70% (with 90% training data and 10% testing data), while the soft voting method reached an accuracy of 86,17%. The labeling process was carried out using the Flair library, validated by a confusion matrix. The application of a labeling method that combines both automatic and manual approaches successfully improved classification accuracy and minimized potential

errors. Additionally, this analysis identified the highest supporter count, with 877 supporters for EVOS, followed by RRQ and ONIC with 743 and 556 supporters, respectively. This research is expected to make a significant contribution to the development of sentiment analysis in the context of e-sports and open up opportunities for further analysis in future research.

Keywords: Sentiment, Ensemble Machine Learning, Preprocessing, Classification, MPL Season 13

1. Pendahuluan

Perkembangan teknologi dari tahun ke tahun semakin berkembang pesat, terutama dalam bidang media *platform*. YouTube adalah salah satu media *platform* yang sangat populer, bahkan saat ini YouTube menjadi layanan video *library* paling populer di internet [1]. Media *platform* seperti YouTube menyediakan fitur komentar yang memungkinkan pengguna untuk memberikan pendapat mereka tentang video maupun siaran langsung yang diunggah. Melalui fitur ini, pengguna dapat berkomentar secara terbuka, seperti memberikan masukan, pujian, serta kritik terhadap video dan *live streaming* yang mereka tonton [2].

Dalam beberapa waktu terakhir, video atau siaran langsung bertemakan tentang *game*, seperti yang diunggah oleh *channel* YouTube MPL Indonesia, semakin populer karena menayangkan video dan *live streaming* turnamen *Mobile Legends* yang bergengsi yaitu MPL *Season 13*. *Mobile Legends* adalah salah satu *game online* berbasis *mobile* yang pertama kali diluncurkan pada tahun 2016 oleh Moonton. Perkembangan *game* bergenre *MOBA* seperti *Mobile Legends* dengan beberapa tahun terakhir telah menjadikannya salah satu *game* yang populer dan diminati oleh masyarakat Indonesia, terutama dari kalangan remaja [3].

Pesatnya popularitas *game MOBA*, terutama *Mobile Legends*, telah membuka berbagai peluang dalam industri *e-sports*. *E-sports* adalah kompetisi olahraga yang menggunakan video *game* sebagai medianya dan sangat diminati oleh kalangan anak muda. Dengan tingginya minat terhadap *game Mobile Legends*, pihak pengembang, Moonton, mengadakan turnamen bergengsi seperti MPL (*Mobile Legends Professional League*). MPL pertama kali diadakan pada tahun 2018 dan terus berlanjut hingga tahun 2024, sampai dengan *season* ke-13 ini. Belum lama ini, turnamen MPL *Season 13* di Indonesia telah berakhir di Jakarta pada tanggal 9 juli 2024. Sebanyak sembilan tim berpartisipasi dalam kompetisi bergengsi ini [4].

Keberhasilan turnamen MPL tidak lepas dari dukungan para *investor* yang berperan penting dalam keberlangsungan MPL *Season 13* bagi masing-masing tim *e-sports*. Oleh karena itu, sangat penting bagi tim *e-sports* untuk memiliki dukungan dari *investor* dalam turnamen MPL *Season 13*. Oleh karena itu, dalam sudut pandang *investor*, biasanya seorang *investor* akan melakukan riset sebelum memutuskan untuk melakukan investasi ke tim *e-sports* [5]. Dalam hal ini Popularitas adalah salah satu contohnya popularitas pada masing-masing tim sangat penting untuk mendapatkan dukungan dari *investor*, karena tim yang populer cenderung menarik lebih banyak perhatian dan kepercayaan.

Analisis sentimen terhadap komentar YouTube dapat memberikan wawasan mengenai bagaimana popularitas tim *e-sports* mempengaruhi minat dan kepercayaan dari *investor*. Berdasarkan pembahasan di atas penelitian ini melakukan penelitian terkait analisis sentimen dengan data *comments* pengguna YouTube tentang video MPL *Season 13* yang di unggah pada *channel* MPL Indonesia. Data komentar ini diambil menggunakan Teknik *scraping* dengan *API* YouTube untuk mengumpulkan komentar-komentar dari video yang ingin di analisis [6].

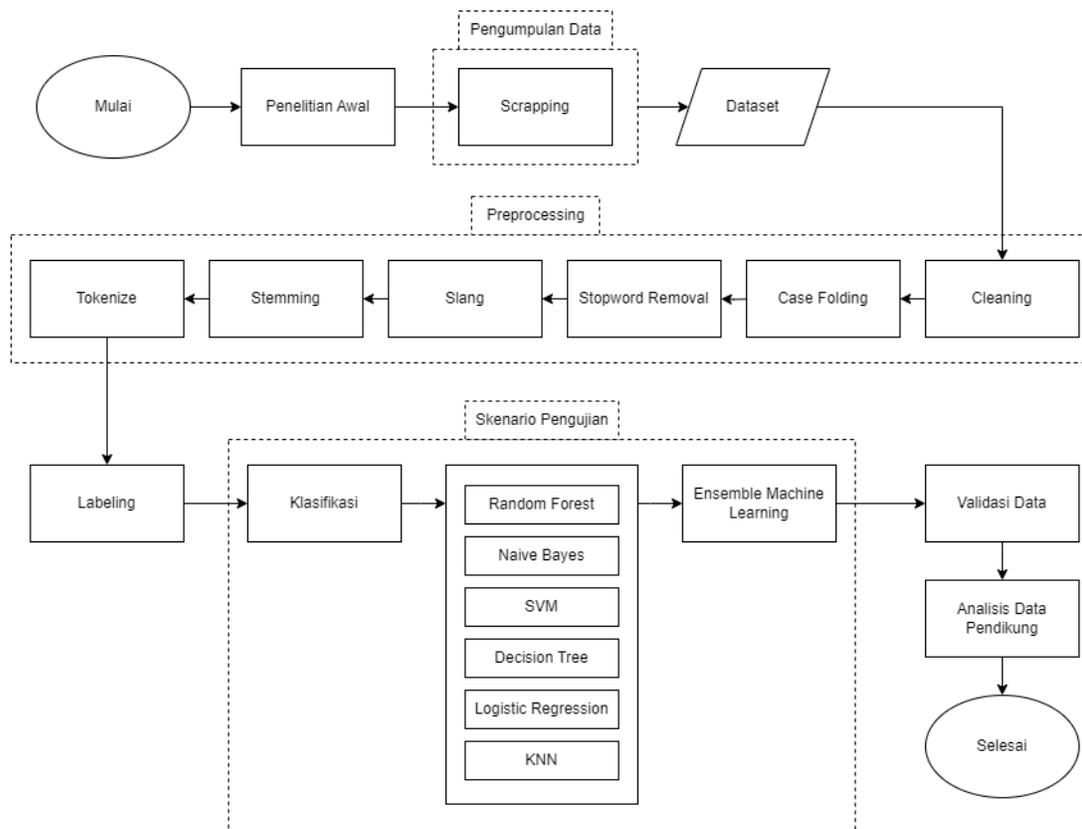
Tujuan dari penelitian ini adalah melakukan analisis sentimen guna mengetahui jumlah popularitas dari masing-masing tim dan pengaruh popularitas terhadap *investor*. Pendekatan yang digunakan adalah *text mining* yang bertujuan untuk menganalisis dan memahami makna dari setiap opini atau komentar pengguna YouTube [7]. Untuk melakukan analisis sentimen tentunya dibutuhkan sebuah metode untuk mengklasifikasi data teks yang ada, penelitian ini menggunakan beberapa metode sekaligus yaitu metode *Random Forest*, *Naïve Bayes*, *SVM*, *Decision Tree*, *Logistic Regression* dan *KNN*. Selain itu, juga diterapkan metode *Ensemble machine learning*, dimana hasil yang diprediksi oleh beberapa model digabungkan untuk mendapatkan kesimpulan akhir. Metodologi ini sangat penting untuk mengklasifikasi komentar YouTube dan

mengevaluasi hasil akurasi paling optimal yang di dapat selama proses klasifikasi [8]. Penelitian ini sangat penting untuk memahami dinamika antara popularitas tim, respon *public*, dan keputusan investasi di *sector e-sports*.

Penelitian ini diharapkan dapat menjelaskan bagaimana *platform* YouTube, khususnya melalui komentar video atau *live streaming*, mempengaruhi popularitas tim *e-sports* dalam MPL *Season 13* dan peran vital *investor* dalam mendukung perkembangan serta keberhasilan tim-tim tersebut. Analisis sentimen dari komentar YouTube diharapkan dapat memberikan wawasan mendalam tentang dinamika interaksi antara *public* dan tim *e-sports*, serta bagaimana hal ini mempengaruhi persepsi dan keputusan *investor* dalam *industry e-sports*.

2. Metode Penelitian

Pada bagian metode penelitian ini, peneliti akan menjelaskan secara rinci setiap tahapan yang dilalui, mulai dari perencanaan awal hingga tahapan akhir. Penjelasan ini mencakup proses pengumpulan data yang digunakan, transformasi data menjadi *dataset* yang siap dianalisis , serta penerapan Teknik *Preprocessing* untuk mempersiapkan data [9]. Selanjutnya, peneliti akan membahas proses *labeling* untuk mengklasifikasikan data, kemudian dilanjutkan dengan penggunaan metode klasifikasi untuk mengidentifikasi pola dalam data. Metode *ensemble learning* diterapkan untuk meningkatkan kinerja model dengan menggabungkan beberapa algoritma. Setelah itu, dilakukan validasi model dengan *confusion matrix* untuk memastikan keakuratannya, dan diakhiri dengan analisis data hasil dari keseluruhan proses [10]. Seluruh Langkah ini diilustrasikan melalui **Gambar 1**.



Gambar 1 Alur Rancangan Penelitian

Penelitian Awal

Dalam tahapan perencanaan awal ini, langkah pertama adalah mengumpulkan referensi. Referensi sangat penting untuk merancang penelitian dan bisa didapat dari berbagai sumber seperti media, buku, dan jurnal. Tujuannya adalah untuk memberikan dasar yang kuat bagi penelitian ini. Fungsi utama referensi adalah untuk mendapatkan pemahaman yang komprehensif tentang analisis sentimen serta meninjau hasil penelitian yang telah dilakukan oleh peneliti.

Dataset

Pada tahapan ini, data dikumpulkan dari video YouTube yang diunggah oleh *channel*/MPL Indonesia, terutama yang terkait dengan MPL *Season* 13. Untuk mengumpulkan data ini, peneliti menggunakan metode *Scraping* dengan YouTube Data *APIv3* dengan bahasa pemrograman *Python* melalui *google colab* untuk mengakses dan mengunduh data komentar dari video-video MPL *Season* 13 [11]. Kemudian data diubah untuk menjadi file *comment_youtube.csv*, setelah proses *scraping* selesai, data yang berhasil dikumpulkan sebanyak 6.424 komentar [12]. Seperti contoh pada **Table 1**.

Tabel 1 Contoh Dataset

<i>comment</i>	<i>channel_name</i>	<i>date</i>	<i>like</i>	<i>dislike</i>
Padahal bisa itu geek menang😭😭	@chchandraa	2024-06-09 00:21:36+00:00	0	0
Gelooo trending 1 game cokkkk	@jatiatmaja1549	2024-06-09 00:16:35+00:00	0	0
yg katanya dibilang ga bs lolos playoff jadi otw msc	@zeealwys	2024-06-09 00:15:24+00:00	0	0
Ngapain si age ikut ikutan co	@karyamelati	2024-06-09 00:09:12+00:00	0	0
Males bgt. ngeliat muka Caderaa rasanya pen berak	@rriskylocos4242	2024-06-09 00:04:59+00:00	1	0

Preprocessing

Setelah melakukan *scraping* data pada tahapan sebelumnya, *dataset* akan menjalani proses selanjutnya, yaitu *Preprocessing*. Fase ini dilakukan untuk memilih data agar data yang tidak terstruktur menjadi data yang terstruktur. Dibawah ini adalah tahapan yang akan dilakukan pada tahapan *preprocessing* ini [13].

Cleaning adalah proses yang bertujuan untuk menghapus karakter atau *symbol* yang tidak relevan dari data yang diperoleh, terutama dari komentar-komentar yang berkaitan dengan video MPL *Season* 13. *Cleaning* mencakup untuk menghilangkan tanda baca, *symbol*, angka, dan karakter khusus lainnya yang tidak memberikan nilai tambah dalam analisis teks, sebagai contoh, *symbol* seperti "@", "#", tautan *URL*, emotikon, atau tanda khusus lainnya yang sering muncul dalam teks di media sosial atau komentar *online* akan dihilangkan agar data mejadi lebih bersih dan lebih mudah di proses [14]. Berikut adalah contoh dari proses *cleaning* yang dapat dilihat pada **Tabel 2**.

Tabel 2 Contoh Proses *Cleaning*

Sebelum	Sesudah
Gila sih keren banget sumpah... GG EVOS... NT GEEK...	Gila sih keren banget sumpah GG EVOS NT GEEK

Case Folding merupakan proses yang mengubah semua karakter alfabet dalam teks menjadi huruf kecil dari 'a' hingga 'z'. Langkah ini sangat krusial dalam analisis teks karena membantu mengurangi variasi yang disebabkan oleh penggunaan huruf besar dan kecil, sehingga kata "MPL" dan "mpl" dianggap sebagai kata yang sama. Dengan menerapkan *case folding*, sistem menjadi lebih konsisten dalam mengenali kata-kata [15]. Contoh dapat dilihat pada **Tabel 3**.

Tabel 3 Contoh Proses *Case Folding*

Sebelum	Sesudah
Gila sih keren banget sumpah GG EVOS NT GEEK	gila sih keren banget sumpah gg evos nt geek

Stopword Removal adalah proses menghapus kata-kata yang tidak memiliki makna penting dalam analisis teks. Kata-kata seperti "dan," "atau," dan "adalah" biasanya tidak memberikan kontribusi *signifikan* terhadap makna keseluruhan kalimat. Dengan menghilangkan *stopword*, analisis dapat lebih berfokus pada kata-kata yang relevan dan inovatif, sehingga meningkatkan efektivitas dan akurasi dalam pengolahan data [16]. Contoh disajikan pada **Tabel 4**.

Tabel 4 Contoh Proses *Stopword Removal*

Sebelum	Sesudah
gila sih keren banget sumpah gg evos nt geek	gila sih keren banget sumpah gg evos nt geek

Slang digunakan untuk mengubah kosakata informal menjadi kosakata baku yang sesuai dengan Kamus Besar Bahasa Indonesia (KBBI). Misalnya, dalam konteks permainan *Mobile Legends*

, istilah slang "gg" bisa diubah menjadi "good game," sedangkan "nt" dapat diartikan sebagai "nice try." Proses ini bertujuan untuk memastikan bahwa Bahasa yang dipakai dalam analisis atau komunikasi formal memenuhi standar yang berlaku, sehingga dapat dipahami dengan jelas oleh semua pihak, dengan mengganti istilah *slang* dengan bentuk baku, kualitas dan kejelasan informasi yang disampaikan akan meningkat [17]. Berikut contoh dari proses *slang* dapat di lihat pada **Tabel 5**.

Tabel 5 Contoh Proses *Slang*

Sebelum	Sesudah
gila sih keren banget sumpah gg evos nt geek	gila sih keren banget sumpah good game evos nice try geek

Stemming adalah proses yang mengonversi kata majemuk menjadi kata dasar. Dengan melakukan *stemming*, variasi dalam bentuk kata dapat disederhanakan, memungkinkan analisis untuk lebih fokus ada makna inti kata tersebut. Proses ini meningkatkan efektifitas pengolahan teks dengan mengurangi kompleksitas data, sehingga algoritma dapat lebih muda mengenali dan mengelompokkan kata-kata yang memiliki arti serupa. Oleh karena itu, *stemming* berperan penting dalam meningkatkan akurasi dan efisiensi analisis teks [18]. Contoh dapat dilihat di **Tabel 6**.

Tabel 6 Contoh Proses *Stemming*

Sebelum	Sesudah
gila sih keren banget sumpah good game evos nice try geek	gila sih keren banget sumpah good game evos nice try geek

Tokenizing adalah proses yang memecah teks dalam dokumen menjadi istilah atau token berdasarkan spasi. Dengan melakukan *tokenizing*, setiap kata atau frasa diidentifikasi sebagai unit terpisah, yang memudahkan analisis selanjutnya. Proses ini sangat penting dalam pengelolaan Bahasa alami karena memungkinkan algoritma untuk menangani dan menganalisis setiap kata secara individual, sehingga meningkatkan akurasi dan efektifitas dalam pemrosesan data [19]. Dan berikut contoh dari proses *tokenize* dapat dilihat pada **Tabel 7**.

Tabel 7 Contoh Proses *Tokenizing*

Sebelum	Sesudah
gila sih keren banget sumpah good game evos nice try geek	['gila', 'sih', 'keren', 'banget', 'sumpah', 'good', 'game', 'evos', 'nice', 'try', 'geek']

Labeling

Langkah berikutnya adalah pelabelan data untuk menentukan apakah komentar tersebut bersifat positif dan negatif. Penelitian ini menggunakan dua tahapan *labeling* yang pertama menggunakan *library Flair*, sebuah perpustakaan yang diunduh melalui *Python Package Index (PyPi)*, untuk melabeli data secara otomatis dan akurat. *Flair* dapat menganalisis sentimen teks dengan membedakan apakah teks bersifat positif dan negatif. Setelah pelabelan dengan *library Flair* selesai, tahap berikutnya adalah pelabelan manual untuk memverifikasi hasil dari proses sebelumnya. Pelabelan manual ini bertujuan untuk memastikan hasil labeling menjadi lebih akurat.

Skenario Pengujian

Setelah menyelesaikan Langkah *preprocessing* dan *labeling data*, langkah berikutnya adalah menerapkan skenario pengujian untuk mengevaluasi kinerja berbagai algoritma *klasifikasi* dan teknik *Ensemble Machine Learning* pada data komentar yang telah diproses. Data yang telah *diversampling* menggunakan SMOTE kemudian dibagi menjadi data latih dan data uji dengan ukuran yang berbeda-beda. Algoritma yang diterapkan dalam penelitian ini mencakup *Random Forest*, *SVM*, *Naïve Bayes*, *Decision Tree*, *Logistic Regression*, dan *KNN*. Setiap algoritma akan dilatih dan dievaluasi dengan memperhatikan akurasi prediksi serta matrik kebingungan guna membandingkan kinerja masing-masing. Berikut adalah penjelasan tentang setiap metode yang digunakan dalam penelitian ini.

Pertama, *Random Forest* merupakan metode *ensemble* yang menggabungkan beberapa pohon keputusan dengan menggunakan mekanisme suara terbanyak. Pendekatan ini sangat efektif dalam mengatasi *overfitting*, sehingga dapat memberikan hasil prediksi yang lebih stabil dan akurat [20].

Support Vector Machine (SVM) berfungsi untuk menemukan *hyperplane* optimal yang memisahkan kelas-kelas dalam data. SVM sangat cocok untuk data berdimensi tinggi dan dapat mengatasi data *non-linear* dengan memanfaatkan *kernel*, serta memiliki ketahanan terhadap *outlier* [21].

Naïve Bayes adalah metode yang berbasis pada probabilitas dengan menggunakan *Teorema Bayes* dan mengasumsikan independensi antar fitur. Meskipun asumsi ini terlihat selalu realistis, metode ini terbukti efektif untuk klasifikasi teks dan *dataset* besar karena kecepatannya dan kemudahan implementasinya [22].

Decision Tree membangun model dengan mengikuti serangkaian aturan berdasarkan fitur. Setiap *node* dalam pohon mewakili fitur tertentu, sementara cabang menunjukkan keputusan. Meskipun mudah dimengerti, metode ini beresiko mengalami *overfitting* pada data yang kompleks, sehingga sering kali perlu menggunakan teknik *pruning* untuk meningkatkan kemampuan generalisasi [23].

Logistic Regression berfungsi untuk menghitung probabilitas klasifikasi menggunakan fungsi *logit*. Meskipun lebih cocok untuk klasifikasi biner, algoritma ini membantu dalam memahami hubungan linear antara fitur kelas yang diprediksi [24].

K-Nearest Neighbors (KNN) mengklasifikasikan data berdasarkan tetangga terdekat. Meskipun algoritma ini cukup sederhana, algoritma ini memiliki biaya Komputasi yang tinggi pada *dataset* besar karena setiap prediksi membutuhkan pencarian tetangga terdekat [25].

Selain itu, teknik *Ensemble Machine Learning* digunakan dengan cara menggabungkan prediksi dari berbagai algoritma yang telah dijelaskan sebelumnya untuk meningkatkan akurasi secara keseluruhan. Metode yang digunakan adalah *voting classifier*, yang terdiri dari dua pendekatan yaitu *hard voting* dan *soft voting*. Pada *hard voting*, prediksi dari setiap algoritma digabungkan, dan hasil dengan suara terbanyak dipilih sebagai prediksi akhir, sementara pada *soft voting*, probabilitas dari masing-masing prediksi dihitung dan hasil dengan probabilitas tertinggi dipilih. Algoritma yang diterapkan dalam *ensemble* ini mencakup *Random Forest*, *SVM*, *Naïve Bayes*, *Decision Tree*, *Logistic Regression* dan, *KNN*. Pengujian ini dilakukan untuk menemukan algoritma paling efektif dalam memodelkan data, sekaligus membandingkan kinerja *ensemble* dengan kinerja algoritma individu.

Validasi Data

Pentingnya dilakukan Validasi Data guna untuk pencocokan data dari hasil klasifikasi dari beberapa metode algoritma sebelumnya. Dengan hasil dari beberapa metode tersebut, kemudian disatukan dengan menggunakan Teknik *Ensemble Machine Learning* kemudian hasilnya di validasi dengan menggunakan *Confusion Matrix*. *Confusion Matrix* pada penelitian ini digunakan untuk mengverifikasi hasil klasifikasi dari algoritma RF, SVM, NB, DT, LR dan, KNN. Perhitungan *F1-Score* dan akurasi digunakan sebagai metode evaluasi untuk model yang diterapkan dalam penelitian ini. Indikator yang digunakan untuk mengevaluasi kinerja algoritma meliputi *True Positif* (TP), *True Negatif* (TN), *True Netral* (TNetral), *False Positif* (FP), *False Negatif* (FN), *False Netral* (FNetral), seperti pada tabel berikut [26].

Tabel 8 Contoh Tabel *Confusion Matrix*

		Prediksi		
		Netral	Negatif	Positif
True	Netral	TNetral	FN	FP
	Negatif	FNetral	TN	FP
	Positif	FNetral	FN	TP

Analisis Data Pendukung

Langkah ini berfokus pada analisis popularitas serta jumlah pendukung dari tim-tim yang berpartisipasi di MPL *Season 13*. Prosesnya dimulai dengan memuat *dataset* yang telah melalui tahap *preprocessing* dan klasifikasi sebelumnya. Pada analisis ini, hanya komentar dengan label positif yang dipilih, karena tujuan utama adalah mengevaluasi dukungan positif yang diterima oleh setiap tim. *Dataset* tersebut kemudian dianalisis untuk mendeteksi penyebutan nama tim dalam setiap komentar.

Setelah nama tim diidentifikasi, jumlah komentar positif yang berkaitan dengan setiap tim dihitung. Hasil penghitungan ini memberikan wawasan tentang tingkat popularitas masing-masing tim berdasarkan dukungan positif yang diterima dari para penggemar. Dengan membandingkan jumlah komentar positif di antara tim-tim tersebut, dapat ditemukan tim mana yang memiliki pendukung terbanyak, serta bagaimana sentimen positif dari komunitas berpengaruh terhadap reputasi mereka selama turnamen MPL *Season 13*.

3. Hasil

Bagian ini menyajikan hasil penelitian yang meliputi jumlah data, hasil pelabelan, visualisasi dengan *wordcloud*, hasil klasifikasi, validasi data, serta analisis tambahan yang memperkuat temuan dan interpretasi penelitian ini.

Hasil Pemrosesan Data Teks

Pada tahapan ini, dilakukan evaluasi terhadap Langkah-langkah *preprocessing* yang diperlukan untuk mempersiapkan data sebelum analisis sentimen. Proses ini meliputi penghapusan data duplikat, penanganan nilai yang hilang, serta normalisasi dan transformasi data. Setiap langkah memiliki peran penting dalam menjamin kualitas data yang optimal, sehingga analisis yang dihasilkan lebih akurat dan dapat dipercaya, mendukung pengambilan keputusan yang tepat.

Langkah pertama adalah *cleaning*, yang membersihkan teks dari elemen-elemen yang tidak relevan seperti tanda baca, angka, dan karakter khusus. Proses ini bertujuan untuk menghapus elemen yang dapat mengganggu analisis, sehingga data lebih terfokus dan siap diolah. Hasil dapat di lihat pada **Tabel 9**.

Tabel 9 Hasil *Cleaning* Data Komentar

No	Comment	Cleaning
1	Gila sih keren banget sumpah... GG EVOS... NT GEEK... dari #9 sampe sekarang #2 😊 jujur terharu banget,	Gila sih keren banget sumpah GG EVOS NT GEEK dari sampe sekarang jujur terharu banget ngelewatin
2	ngelewatin caci maki smua orang + di remehin semua orang. we love you all ❤️	caci maki smua orang di remehin semua orang we love you all
3	RRQ.EVOS.ONIC itu itu doang yg juara kasian yg lain	RRQ EVOS ONIC itu itu doang yg juara

Langkah berikutnya adalah *case folding*, yang dilakukan untuk menyeragamkan teks dengan mengonversi semua huruf kecil. Hal ini bertujuan untuk menghindari perbedaan dalam pengolahan kata yang sebetulnya sama, tetapi ditulis dengan huruf kapital yang berbeda, seperti "Teknologi" dan "teknologi," sehingga dianggap sebagai satu kata yang sama. Setelah itu, dilakukan *stopword removal*, yakni penghapusan kata-kata umum yang sering muncul tetapi tidak memberikan kontribusi penting dalam analisis sentimen, seperti "yang," "di," dan "dan." Tujuannya adalah agar algoritma dapat lebih fokus pada kata-kata yang lebih relevan dalam menentukan sentimen. Setelah proses *stopword removal*, langkah

selanjutnya adalah *slang replacement*, di mana kata-kata *slang* atau informal diganti dengan kata-kata formal sesuai. Proses ini menggunakan file *JSON* berisi daftar kata *slang* dan padanannya, misalnya "gg" diubah menjadi "good game" dan "nt" menjadi "nice try". Tujuan dari tahapan ini adalah untuk meningkatkan konsistensi bahasa dalam data, sehingga analisis sentimen dapat dilakukan. Berikut adalah hasil yang dapat dilihat melalui **Tabel 10**, **Tabel 11** dan, **Tabel 12**.

Tabel 10 Hasil *Case Folding* Data Komentar

No	<i>Cleaning</i>	<i>Case Folding</i>
1	Gila sih keren banget sumpah GG EVOS NT GEEK	gila sih keren banget sumpah gg evos nt geek
2	dari sampe sekarang jujur terharu banget ngelewatn caci maki smua orang di remehin semua orang we love you all	dari sampe sekarang jujur terharu banget ngelewatn caci maki smua orang di remehin semua orang we love you all
3	RRQ EVOS ONIC itu itu doang yg juara	rrq evos onic itu itu doang yg juara

Tabel 11 Hasil *Stopword Removal* Data Komentar

No	<i>Case Folding</i>	<i>Stopword Removal</i>
1	gila sih keren banget sumpah gg evos nt geek	gila sih keren banget sumpah gg evos nt geek
2	dari sampe sekarang jujur terharu banget ngelewatn caci maki smua orang di remehin semua orang we love you all	sampe jujur terharu banget ngelewatn caci maki smua orang remehin orang we love you all
3	rrq evos onic itu itu doang yg juara	rrq evos onic doang yg juara kasian yg

Tabel 12 Hasil *Slang* Data Komentar

No	<i>Stopword Removal</i>	<i>Slang</i>
1	gila sih keren banget sumpah gg evos nt geek	gila sih keren banget sumpah good game evos nice try geek
2	sampe jujur terharu banget ngelewatn caci maki smua orang remehin orang we love you all	sampai jujur terharu banget ngelewatn caci maki smua orang remehin orang we love you all
3	rrq evos onic doang yg juara kasian yg	rrq evos onic doang yang juara kasian yang

Pada tahapan *stemming*, proses dilakukan dengan memanfaatkan *library sastrawi* untuk mengubah kata-kata ke bentuk dasarnya. Contohnya, kata "bermain" akan diubah menjadi "main." Langkah ini krusial karena memungkinkan variasi kata yang memiliki akar yang sama untuk dikenali sebagai satu entitas, sehingga meningkatkan akurasi dalam analisis. Selanjutnya, pada tahapan tokenisasi, teks dipecahkan menjadi unit-unit kecil, atau token, seperti kata-kata individu. Proses ini mempermudah analisis lebih lanjut, karena setiap karakter pisah dan dapat dianalisis secara mandiri, memberikan pemahaman yang lebih mendalam terhadap konten teks. Berikut hasil dapat dilihat pada **Tabel 13** dan **Tabel 14**

Tabel 13 Hasil *Stemming* Data Komentar

No	<i>Slang</i>	<i>Stemming</i>
1	gila sih keren banget sumpah good game evos nice try geek	gila sih keren banget sumpah good game evos nice try geek
2	sampai jujur terharu banget ngelewatn caci maki smua orang remehin orang we love you all	sampai jujur haru banget ngelewatn caci maki smua orang remehin orang we love you all
3	rrq evos onic doang yang juara kasian yang	rrq evos onic doang yang juara kasi yang

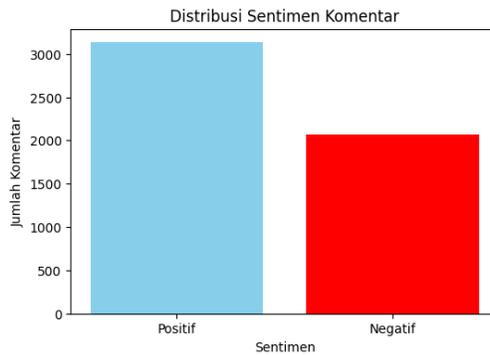
Tabel 14 Hasil *Tokenize* Data Komentar

No	<i>Stemming</i>	<i>Tokenize</i>
1	gila sih keren banget sumpah good game evos nice try geek	['gila', 'sih', 'keren', 'banget', 'sumpah', 'good', 'game', 'evos', 'nt', 'geek']
2	sampai jujur haru banget ngelewatn caci maki smua orang remehin orang we love you all	['sampai', 'jujur', 'haru', 'banget', 'ngelewatn', 'caci', 'maki', 'smua', 'orang', 'remehin', 'orang', 'we', 'love', 'you', 'all']
3	rrq evos onic doang yang juara kasi yang	['rrq', 'evos', 'onic', 'doang', 'yang', 'juara', 'kasi', 'yang']

Hasil Labeling *Dataset*

Tahapan ini akan menjelaskan hasil pelabelan yang dilakukan dengan memanfaatkan *library Flair*, yang merupakan Langkah krusial dalam analisis sentimen. Proses pelabelan diawali dengan penerapan metode otomatis dari *Flair*, kemudian diikuti oleh pelabelan manual untuk meningkatkan tingkat akurasi hasil akhir. Melalui kombinasi kedua pendekatan ini, diharapkan hasil pelabelan dapat menjadi lebih akurat dan dapat diandalkan. Dari total 6.424 data komentar yang belum di olah, jumlah komentar yang relevan berkurang menjadi 5.185 setelah melalui proses *preprocessing* dan pelabelan.

Hasil pelabelan mengindikasikan bahwa terdapat 3.131 komentar yang termasuk dalam kategori positif, sementara 2.064 komentar diklasifikasikan sebagai negatif. Penggabungan antara pelabelan otomatis dan manual ini bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai sentimen dalam data komentar, sekaligus mengurangi kemungkinan kesalahan yang mungkin terjadi jika hanya menggunakan satu metode pelabelan. Hasil dapat dilihat dalam bentuk grafik pada **Gambar 2**.



Gambar 2 Grafik Positif dan Negatif

WordCloud

Wordcloud dalam penelitian ini berfungsi sebagai alat visualisasi yang menampilkan frekuensi kemunculan kata dalam teks, yang sangat bermanfaat untuk analisis sentimen. Dengan visualisasi ini, peneliti dapat dengan cepat mengidentifikasi tema dan pola sentimen yang terdapat dalam komentar, di mana kata-kata yang lebih besar menunjukkan frekuensi kemunculan yang tinggi. Untuk memberikan gambaran yang lebih jelas, penggunaan *wordcloud* dalam penelitian ini menggambarkan kata-kata yang sering muncul dalam komentar positif dan negatif, sehingga meningkatkan pemahaman tentang sentimen yang ada. Sebagai contoh, berikut ini hasil dari *wordcloud* dapat dilihat di **Gambar 3**.



Gambar 3. a) *WordCloud* Sentimen Positif dan, b) *WordCloud* Sentimen Negatif.

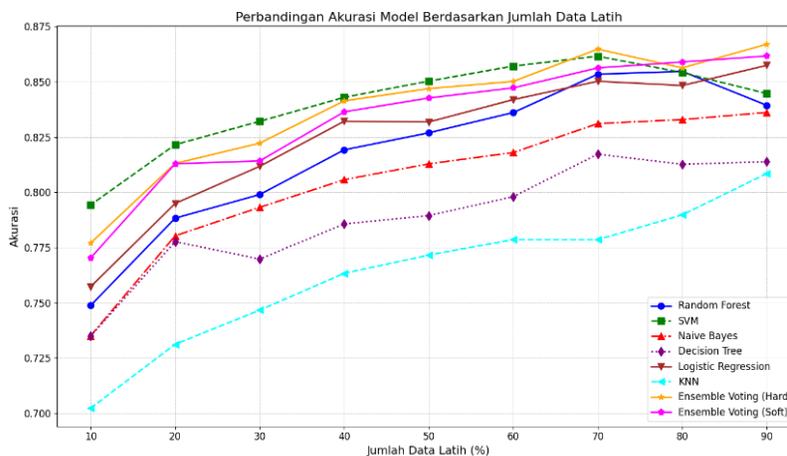
Hasil Metode

Tahapan ini melibatkan penerapan berbagai metode klasifikasi secara simultan, yang kemudian digabungkan menggunakan teknik *ensemble machine learning* dengan pendekatan *majority voting*. Proses ini terdiri dari dua metode, yaitu *hard voting* dan *soft voting*. Sebelum klasifikasi dilakukan, data diproses dengan menggunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menangani masalah ketidakseimbangan kelas. Setelah itu, dataset dibagi menjadi data latih dan data uji, dengan pembagian yang bervariasi, data latih memiliki ukuran 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, sedangkan data uji memiliki ukuran 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%. Hasil dari pengujian menunjukkan bahwa performa pada masing-masing metode yang diuji tidak melebihi teknik *ensemble machine learning*, karena hasil dari semua metode tersebut digabungkan. Data hasil pengujian menunjukkan variasi akurasi untuk setiap metode, seperti yang tertera dalam **Tabel 15** berikut.

Tabel 15 Hasil Akurasi Metode

Data train	Data test	Random forest	Support Vector Machine	Naïve bayes	Decision tree	Logistic regression	K-Nearest Neighbors	Voting hard	Voting soft
90	10	83.94%	84.47%	83.62%	81.38%	85.74%	80.85%	86.70%	86.17%
80	20	85.47%	85.42%	83.29%	81.27%	84.83%	78.98%	85.63%	85.90%
70	30	85.34%	86.16%	83.11%	81.72%	85.02%	77.86%	86.48%	85.63%
60	40	83.61%	85.71%	81.80%	79.80%	84.19%	77.86%	85.02%	84.73%
50	50	82.69%	85.03%	81.29%	78.94%	83.18%	77.16%	84.69%	84.27%
40	60	81.92%	84.30%	80.57%	78.57%	83.22%	76.33%	84.14%	83.64%
30	70	79.90%	83.21%	79.32%	76.98%	81.17%	74.67%	82.22%	81.42%
20	80	78.83%	82.16%	78.03%	77.76%	79.49%	73.12%	81.30%	81.29%
10	90	74.88%	79.43%	73.50%	73.49%	75.73%	70.23%	77.70%	77.03%

Berikut ini disajikan hasil dalam format grafik, yang memberikan visualisasi yang lebih jelas tentang performa setiap metode. Grafik pada **Gambar 4** menunjukkan akurasi yang diperoleh dari masing-masing pendekatan, sehingga mempermudah perbandingan efektivitas metode klasifikasi yang digunakan.

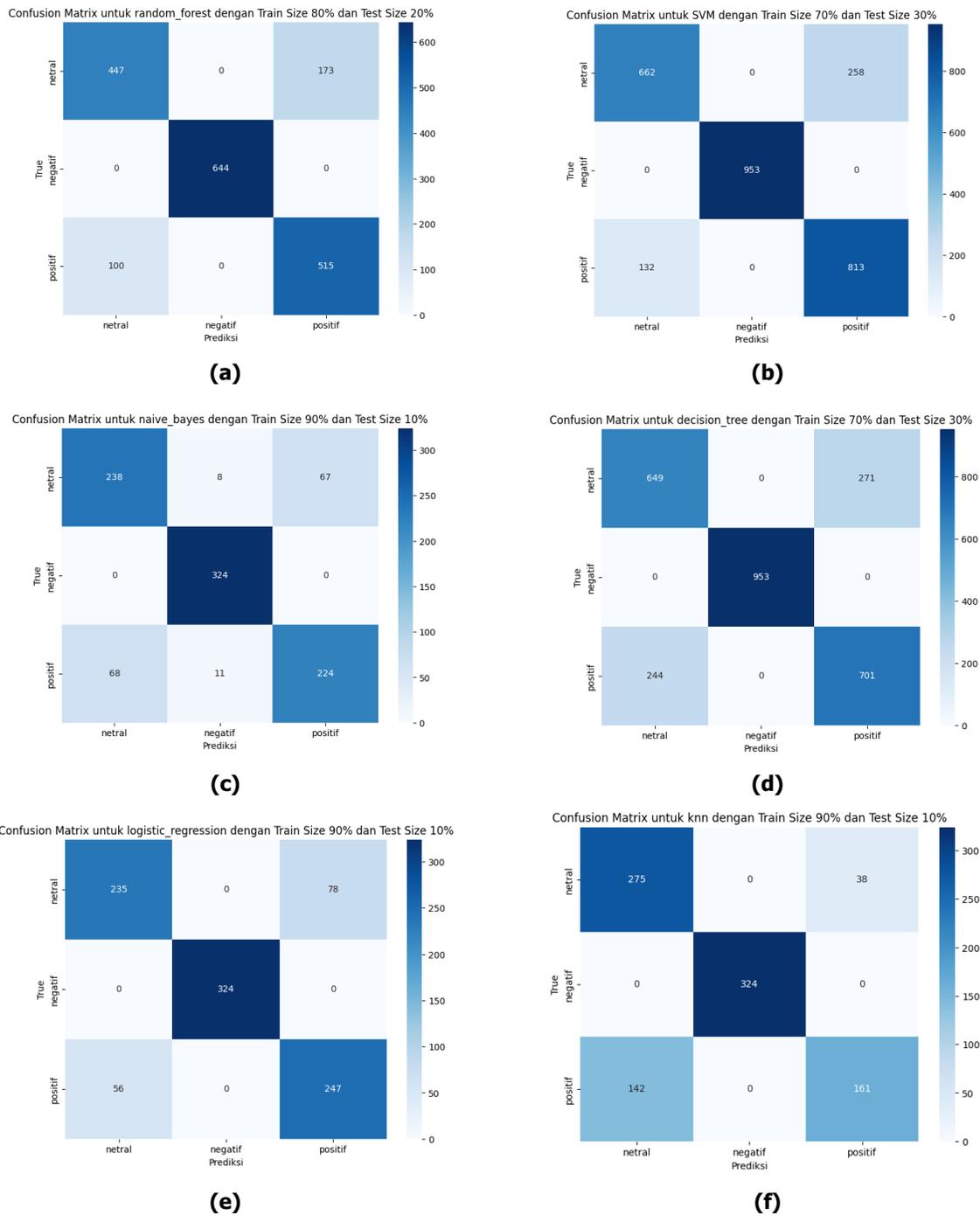


Gambar 4 Grafik Hasil Klasifikasi Metode

Validasi Hasil

Hasil prediksi dari tiap metode klasifikasi dianalisis menggunakan *confusion matrix*. Matriks ini memberikan detail tentang jumlah prediksi yang benar dan salah untuk setiap kelas. Di bawah ini, disajikan visualisasi *confusion matrix* untuk berbagai metode klasifikasi yang telah dievaluasi. Gambar di bawah ini

memberikan gambaran yang lebih jelas mengenai distribusi prediksi yang benar dan salah untuk setiap kelas, sehingga memudahkan dalam menganalisis kinerja masing-masing model secara lebih mendalam. Dan berikut hasil dari penggunaan Confusion Matrix ke beberapa metode dapat dilihat pada **Gambar 5**.



Gambar 5. a) *Confusion Matrix* Random Forest, b) *Confusion Matrix* SVM, c) *Confusion Matrix* Naïve Bayes, d) *Confusion Matrix* Decision Tree, e) *Confusion Matrix* Logistic Regression, f) *Confusion Matrix* KNN.

Hasil Analisis Jumlah Pendukung

Fokus dari tahapan analisis ini bertujuan untuk menilai popularitas dan jumlah pendukung dari berbagai tim yang berpartisipasi dalam turnamen MPL *Season 13*. Dengan memanfaatkan *dataset* yang telah dipilih dan diolah, penelitian ini berhasil mengidentifikasi jumlah pendukung setiap tim. Hasilnya menunjukkan bahwa tim EVOS unggul dengan total pendukung mencapai 877. Tim RRQ dan ONIC juga menunjukkan performa yang baik, dengan jumlah dukungan 743 dan 556. Sementara itu, tim-tim lain seperti AE, BTR, GEEK, RBL, AURA, dan DEWA memiliki jumlah pendukung yang lebih sedikit, yakni 231, 193, 159, 82, 122, dan 111. Data ini memberikan gambaran yang jelas mengenai popularitas tim dalam kompetisi ini. Hasil dapat di lihat pada **Tabel 16**.

Tabel 16 Jumlah Pendukung Tim

Tim	EVOS	AE	BTR	ONIC	GEEK	RBL	RRQ	AURA	DEWA
Total	887	231	193	556	159	82	743	122	111



Gambar 6 Grafik Pendukung Tim

4. Kesimpulan

Penggunaan *library Flair* dalam pelabelan meningkatkan akurasi klasifikasi sentimen pada komentar MPL *Season 13*. Dari 6.424 komentar, jumlah yang relevan berkurang menjadi 5.185 setelah *preprocessing*, dengan rincian 3.131 komentar positif dan 2.064 komentar negatif. Kombinasi pelabelan otomatis dan manual membantu memahami sentimen lebih baik dan mengurangi kesalahan, sehingga memperkuat analisis pola respon pengguna dan meningkatkan validitas data.

Penerapan teknik *ensemble machine learning* dengan pendekatan *majority voting* menunjukkan hasil yang lebih unggul dibandingkan dengan metode klasifikasi individual. Metode *ensemble voting (hard)* mencatat akurasi tertinggi sebesar 86,70% untuk data latih 90% dan data uji 10%, sementara *ensemble voting (soft)* mencapai akurasi 86,17%. meskipun terdapat variasi dalam hasil, teknik *ensemble* terbukti efektif dalam meningkatkan akurasi klasifikasi secara keseluruhan. Analisis ini juga mengungkap bahwa EVOS memiliki jumlah pendukung terbanyak 877, diikuti oleh RRQ 743, dan ONIC 556. Temuan ini memberikan peluang untuk penelitian lebih lanjut.

5. Daftar Pustaka

- [1] Y. Anisa, "Peran Channel Youtube Sebagai Media Alternatif untuk Membantu Proses Pembelajaran Matematika dan Media Informasi pada Tingkat Perguruan Tinggi," 2022. [Online]. Available: <https://ejournal.unib.ac.id/index.php/jpmr>
- [2] G. Sanjaya and K. Muslim Lhaksmana, "Analisis Sentimen Komentar YouTube tentang Terpilihnya Menteri Kabinet Indonesia Maju Menggunakan Lexicon Based," *e-Proceeding Eng.*, vol. 7, no. 3, pp. 9698–9710, 2020, [Online]. Available: <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/10703>
- [3] R. Yusrinawati, "Analisis Perilaku Konsumtif Pemain Game Mobile Legends Bang Bang di ESports Indonesia Jember يليب," *Nucl. Phys.*, vol. 13, no. 1, pp. 104–116, 2023.
- [4] D. Wahyuni, "TANTANGAN DAN PELUANG ESPORTS DALAM KEOLAHRAGAAN NASIONAL CHALLENGES AND OPPORTUNITIES OF ESPORTS IN NATIONAL SPORTS," *Kajian*, vol. 25, no. 4, pp. 341–353, 2020, [Online]. Available: <https://beritagar.id/artikel/arena/>
- [5] J. Budiman, R. Limgestu, and I. Tri Sagiato, "Perilaku Keputusan Investasi Investor Pasar Saham Indonesia," *J. Ilm. Akunt. dan Keuang.*, vol. 5, no. 9, pp. 3518–3526, 2023, [Online]. Available: <https://journal.ikopin.ac.id/index.php/fairvalue>
- [6] F. Gunardi, "PENGARUH E-SPORTS MARKETING DAN KEPUASAN PELANGGAN TERHADAP BRAND LOYALTY RAZER INC," 2020.
- [7] D. Puspita Nilamsari and I. Parma Dewi, "Jurnal Vocational Teknik Elektronika dan Informatika," *J. Vocat. Tek. Elektron. dan Inform.*, vol. 11, no. 1, pp. 96–102, 2023, [Online]. Available: <http://ejournal.unp.ac.id/index.php/voteknika/index>
- [8] M. Yasir and R. Suraji, "Perbandingan Metode Klasifikasi Naive Bayes, Decision, Tree, Random Forest Terhadap Analisis Sentimen Kenaikan Biaya Haji 2023 pada Media Sosial Youtube," *J. Cahaya Mandalika*, vol. 3, no. 2, pp. 180–192, 2023.
- [9] A. E. Budiman and A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 475–488, 2020, doi: 10.28932/jutisi.v6i3.2892.
- [10] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI)*, vol. 5, no. 2, pp. 697–711, 2021.
- [11] D. Dwi Kurnianto and S. Waluyo, "3 rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) 30 Agustus 2023-Jakarta," 2023.
- [12] R. A. Fauzan and M. Mufti, "Analisis Sentimen Komentar Youtube Program Kampus Merdeka Berbasis Web Menggunakan Algoritma Multinomial Naïve Bayes," *Semin. Nas. Mhs. Fak. Teknol. Inf.*, vol. 2, no. 2, pp. 864–871, 2023, [Online]. Available: <https://senafti.budiluhur.ac.id/index.php/senafti/article/view/929/563>
- [13] M. A. Rosid, A. S. Fitriani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.
- [14] Chely Aulia Misrun, E. Haerani, M. Fikry, and E. Budianita, "Analisis sentimen komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 4, no. 1, pp. 207–215, Apr. 2023, doi: 10.37859/coscitech.v4i1.4790.

- [15] S. M. Harahap and R. Kurniawan, "Analisis Sentimen Komentar Youtube Terhadap Food Vlogger Dengan Menggunakan Metode Naïve Bayes," *MEANS (Media Inf. Anal. dan Sist.*, vol. 9, no. 1, pp. 87–96, 2024, doi: 10.54367/means.v9i1.3912.
- [16] D. Firmansyah, F. Sutrisno, and S. Waluyo, "ANALISIS SENTIMEN MASYARAKAT MENGENAI KASUS KEKUASAAN NARKOBA PADA KOMENTAR YOUTUBE MENGGUNAKAN METODE K-NEAREST NEIGHBOURS," 2024.
- [17] S. K. Dirjen, P. Riset, D. Pengembangan, R. Dikti, S. Khomsah, and A. S. Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *J. Resti*, vol. 1, no. 3, pp. 648–654, 2021.
- [18] B. Franko, N. Wilyanto, H. Irsyad, U. Multi, and D. Palembang, "Analisis Sentimen Terhadap Naturalisasi Pemain pada Youtube Menggunakan Decision Tree dan Naive Bayes Sentiment Analysis of Player Naturalization on Youtube Using Decision Trees and Naive Bayes," vol. 03, no. September, pp. 8–16, 2024, doi: 10.57203/session.v3i1.2024.8-16.
- [19] B. I. Jimmy Alga, Cindi Wulandari, "Analisis Sentimen Aplikasi Youtube di Google Play Store Menggunakan Machine Learning," *RESOLUSI Rekayasa Tek. Inform. dan Inf. - ISSN 2745-7966*, vol. 6, no. 1, pp. 197–207, 2024, [Online]. Available: <https://djournals.com/resolusi>
- [20] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4305–4313, 2022.
- [21] N. Hendrastuty, A. Rahman Isnain, and A. Yanti Rahmadhani, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, pp. 150–155, 2021.
- [22] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 10, no. 1, pp. 34–40, 2022, doi: 10.23960/jitet.v10i1.2262.
- [23] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiyari, "Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE," *Aiti*, vol. 18, no. 2, pp. 173–184, 2021, doi: 10.24246/aiti.v18i2.173-184.
- [24] M. R. Nurhusen, J. Indra, and K. A. Baihaqi, "Analisis Sentimen Pengguna Twitter Terhadap Kenaikan Harga Bahan Bakar Minyak (BBM) Menggunakan Metode Logistic Regression," *J. Media Inform. Budidarma*, vol. 7, no. 1, p. 276, 2023, doi: 10.30865/mib.v7i1.5491.
- [25] R. S. Amardita, A. Adiwijaya, and M. D. Purbolaksono, "Analisis Sentimen terhadap Ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung Menggunakan Algoritma KNN," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 1, p. 62, 2022, doi: 10.30865/jurikom.v9i1.3793.
- [26] N. Agustina and C. N. Ihsan, "Pendekatan Ensemble untuk Analisis Sentimen Covid19 Menggunakan Pengklasifikasi Soft Voting," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, p. 263, 2023, doi: 10.25126/jtiik.20231026215.