

# ANALISIS SENTIMEN DAN PEMODELAN TOPIK TERHADAP APLIKASI PEMBELAJARAN ONLINE PADA PLATFORM GOOGLE PLAY

Andhika Ihsan Kamil\*<sup>1)</sup>, Oktariani Nurul Pratiwi<sup>2)</sup>, Deden Witarsyah<sup>3)</sup>

1. Telkom University, Indonesia
2. Telkom University, Indonesia
3. Telkom University, Indonesia

## Article Info

**Kata Kunci:** Analisis Sentimen; LDA; Pemodelan Topik; Ruangguru; SVM

**Keywords:** *Sentiment Analysis; LDA; Topic Modeling; Ruangguru; SVM*

## Article history:

Received 29 Oktober 2024

Revised 13 November 2024

Accepted 4 Desember 2024

Available online 15 March 2025

## DOI :

<https://doi.org/10.29100/jupi.v10i2.6023>

\* Corresponding author.

Andhika Ihsan Kamil

E-mail address:

[andhikaihs@student.telkomuniversity.ac.id](mailto:andhikaihs@student.telkomuniversity.ac.id)

## ABSTRAK

Kita hidup di era teknologi di mana setiap aspek kehidupan terhubung dengan teknologi. Maraknya aplikasi belajar daring di Indonesia, seperti Ruangguru, menandai perkembangan di bidang pendidikan. Aplikasi Ruangguru fokus pada jasa pendidikan dan telah melayani lebih dari 22 juta pengguna. Untuk mempertahankan kepuasan pelanggan, diperlukan analisis sentimen dan pemodelan topik terhadap ulasan pengguna. Penelitian ini menggunakan 31.070 dataset ulasan pengguna di *Google Play*, dilanjutkan dengan pelabelan dan *preprocessing* sebelum data akan digunakan. Analisis sentimen memakai algoritma *Support Vector Machine* menunjukkan hasil yang baik dengan akurasi 88,89%, presisi 87,11%, recall 91,22%, dan F1-score 89,11%. Teknik *k-10 fold cross validation* menghasilkan akurasi rata-rata 89,06%. Kemudian model digunakan pada 10.000 ulasan baru, dengan hasil mayoritas ulasan memiliki sentimen positif. Pemodelan topik dengan *Latent Dirichlet Allocation* mengidentifikasi 5 topik utama pada sentimen positif dengan nilai koherensi 0,4779, berfokus pada pengalaman positif dan kegunaan aplikasi dalam membantu belajar. Pada sentimen negatif, ditemukan 4 topik utama dengan nilai koherensi 0,4899, yang banyak mengungkapkan keluhan tentang materi pembelajaran yang kurang lengkap.

## ABSTRACT

*We live in a technological era where every aspect of life is connected to technology. The rise of online learning apps in Indonesia, such as Ruangguru, marks a development in education. Ruangguru app focuses on education services and has served more than 22 million users. To maintain customer satisfaction, sentiment analysis and topic modeling of user reviews are required. This research uses 31,070 user review datasets on Google Play, followed by labeling and preprocessing before the data will be used. Sentiment analysis using Support Vector Machine algorithm shows good results with 88.89% accuracy, 87.11% precision, 91.22% recall, and 89.11% F1-score. The k-10 fold cross validation technique produced an average accuracy of 89.06%. Then the model was used on 10,000 new reviews, with the result that the majority of reviews had positive sentiments. Topic modeling with Latent Dirichlet Allocation identified 5 main topics on positive sentiment with a coherence value of 0.4779, focusing on positive experiences and the usefulness of the application in helping to learn. On negative sentiment, 4 main topics were found with a coherence value of 0.4899, which mostly expressed complaints about incomplete learning materials.*

## I. PENDAHULUAN

**S**AAT ini, kita hidup di zaman teknologi di mana setiap aspek kehidupan terhubung dengan teknologi, oleh karena itu teknologi memainkan peran penting dalam berbagai skenario [1]. Maraknya aplikasi belajar daring di Indonesia menjadi salah satu penanda peran penting teknologi di bidang pendidikan. Salah satunya yaitu Ruangguru, sebuah aplikasi yang dibuat oleh Belva Devara dan Iman Usman pada tahun 2014. Sebagai salah satu aplikasi terkemuka di Indonesia, Ruangguru memiliki fokus pada pelayanan pendidikan dan sudah sukses melayani

lebih dari 22 juta orang. Perusahaan ini juga mengendalikan lebih dari 300.000 guru yang menyediakan layanan dalam lebih dari 100 bidang pelajaran [2]. Banyak penghargaan yang telah diraih, seperti pada tahun 2015 meraih UNICEF *Young Innovation Forum*, dan pada tahun 2016 meraih *Bubu Awards* dalam kategori *Indonesia Best Education Web* [3]. Secara sederhana, konsep Ruangguru adalah menghubungkan guru les dengan muridnya melalui aplikasi daring. Sebagai bagian dari kegiatan pendidikan, Ruangguru terus berinovasi untuk menarik minat pelaku pendidikan. Dengan menyediakan kelancaran akses, kelengkapan materi, dan metode pembelajaran yang memuaskan, Ruangguru telah dengan cepat menjadi populer [4]. Sehingga dalam mempertahankan kepuasan pengguna, sentimen analisis dan pemodelan topik dapat digunakan karena keduanya dapat memberikan wawasan mendalam terkait respon pengguna terhadap aplikasi.

Respon kepuasan pengguna terhadap Ruangguru pada *Google Play* merupakan sumber yang dapat dimanfaatkan oleh Ruangguru untuk mengevaluasi penentuan tujuan perusahaan serta memantau kinerja produknya. Kepuasan pelanggan merupakan konsep umum yang menyatakan kepuasan mereka dalam mengkonsumsi suatu jasa, sedangkan kualitas jasa berkaitan dengan komponen-komponen tertentu [5]. Sedangkan menurut Khadka & Maharjan, kepuasan pelanggan bisa diartikan sebagai evaluasi komprehensif terhadap suatu produk atau layanan, yang didasarkan pada pengalaman membeli dan menggunakan produk tersebut seiring berjalannya waktu [6]. Sampai tanggal 27 Oktober 2023, data dari situs *Google Play* menunjukkan bahwa aplikasi Ruangguru telah mencapai lebih dari 10 juta unduhan dan telah dikomentari lebih dari 640 ribu ulasan. Ulasan aplikasi tersebut dapat diolah menjadi sumber informasi yang bernilai. Teknik untuk menyimpulkan faktor-faktor yang sering muncul dalam ulasan dilakukan dengan mengklasifikasikan opini ke dalam dua kategori, yaitu kelas positif dan negatif [7]. Namun untuk mendapatkan dan mengolah data ulasan menjadi informasi tidaklah mudah, sehingga dibutuhkan suatu metode yang salah satunya analisis sentimen. Menganalisis sentimen ulasan pengguna memungkinkan identifikasi apakah pengguna merasa puas atau tidak puas terhadap fitur yang ada. Sehingga pengembang dapat merespons terhadap fitur yang dikeluhkan oleh pengguna untuk meningkatkan kepuasan dan pengalaman pengguna secara keseluruhan.

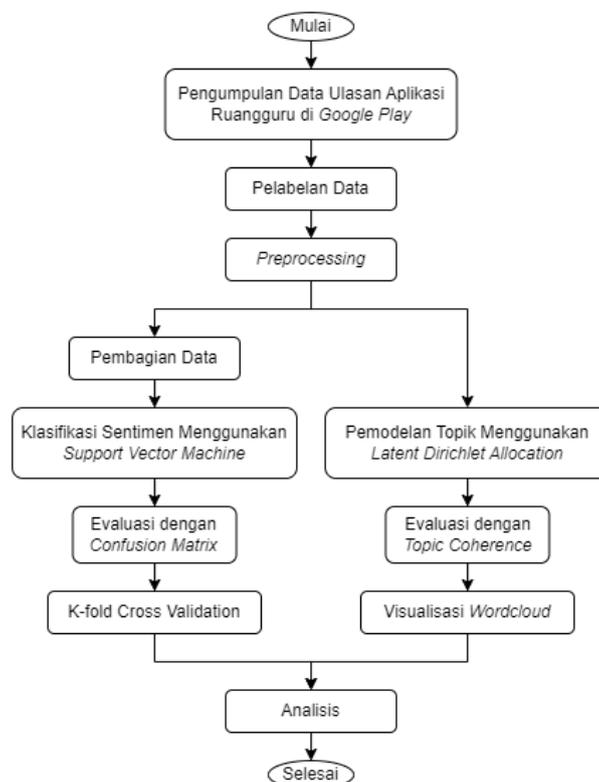
Analisis sentimen merupakan metode yang dipakai untuk memproses dan memahami teks tidak terstruktur dengan maksud untuk mengekstrak dan menilai sentimen yang terkandung dalam sebuah kalimat, pendapat, atau opini [8]. Analisis sentimen dilakukan untuk mengevaluasi opini dan kecenderungan sebuah pendapat terhadap suatu topik, baik secara positif maupun negatif [9]. Dalam praktik analisis sentimen melalui metode *text mining*, beberapa teknik yang umum digunakan antara lain adalah *Naive Bayes*, *K-NN*, *Random Forest*, dan *Support Vector Machine* (SVM). Dalam studi yang dilakukan oleh Lukmana dan rekan-rekannya, mereka memanfaatkan metode *text mining* untuk menganalisis sentimen politik dari data *Twitter* yang berkaitan dengan calon presiden Jokowi dan Prabowo. Mereka menggunakan algoritma SVM untuk analisis tersebut dan berhasil mencapai tingkat akurasi sebesar 86% dalam hasilnya [10]. Kemudian dalam studi yang dilakukan oleh Annur dan rekan-rekannya, mereka telah berhasil dalam menganalisis sentimen menggunakan algoritma SVM dari data ulasan aplikasi Quipper pada *Google Play*, dengan hasil evaluasi kinerja SVM menggunakan *kernel linier* menghasilkan akurasi sebesar 90,25% [11]. Berdasarkan ulasan literatur dari penelitian-penelitian sebelumnya tersebut, penulis memutuskan untuk menggunakan SVM dalam penelitian ini. Alasan memilih algoritma SVM adalah kemampuannya dalam mengelola jumlah besar data ulasan dan mengklasifikasikannya berdasarkan sentimen, yang dikenal memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma lainnya.

Namun analisis sentimen tidak dapat menunjukkan topik atau tema spesifik yang sering muncul dalam respon pengguna, sehingga dibutuhkan pemodelan topik untuk dapat membantu mengidentifikasi dan mengelompokkan pola dan tren yang berkaitan dengan aspek tertentu. Terdapat beberapa algoritma dalam teknik pemodelan topik, antara lain *Probabilistic Latent Semantic Analysis* (PLSA), *Latent Semantic Analysis* (LSA), *Non-Negative Matrix Factorization* (NNMF), dan *Latent Dirichlet Allocation* (LDA) [12]. Dalam penelitian sebelumnya oleh Kustyaningsih dan Permana, mereka menggunakan metode LDA untuk melakukan pemodelan topik. Hasil penelitian menunjukkan bahwa skor koherensi tertinggi yang mereka capai adalah 0,487 pada jumlah topik sebanyak 3 topik [13]. Sehingga dalam penelitian ini penulis memutuskan untuk memanfaatkan metode LDA. Dengan demikian, diharapkan penelitian ini dapat menghasilkan kelompok pendapat para pengguna Ruangguru mengenai layanan yang disediakan untuk kemudian menjadi bahan penilaian dalam peningkatan pelayanan yang dimiliki oleh Ruangguru.

## II. METODE PENELITIAN

### A. Sistematika Penyelesaian Masalah

Penelitian ini mengikuti serangkaian langkah sistematis untuk memastikan keberhasilan pelaksanaan penelitian. penelitian dimulai dari pengumpulan data, pelabelan data, *preprocessing*, pembagian data, kemudian sentimen analisis menggunakan algoritma SVM dan pemodelan topik menggunakan LDA, evaluasi model, hingga analisis. Sistematika penyelesaian ditampilkan pada Gambar 1.



Gambar. 1. Sistematika Penyelesaian Masalah

### B. Pengumpulan Data

Pengumpulan data dilakukan menggunakan *library google\_play\_scraper* yang digunakan pada bahasa pemrograman *Python*. Data yang diambil merupakan data ulasan pengguna pada aplikasi Ruangguru yang menggunakan Bahasa Indonesia dan diurutkan dari yang paling relevan. Didapatkan 64.072 ulasan yang diambil pada rentang Januari–Maret 2024. Pengambilan data dilakukan tanpa menggunakan filter rating review sehingga data yang didapatkan memiliki rentang rating 1-5 bintang. Data yang sudah berhasil dikumpulkan selanjutnya diberi label sentimen pada setiap ulasan yang ada. Namun hanya 31.070 ulasan yang digunakan karena mempertimbangkan sumber daya komputer yang ada.

### C. Pelabelan Data

Pelabelan data dilakukan berdasarkan sentimen atau kepuasan pengguna yang terkandung dalam setiap ulasan, label positif jika mengandung sentimen positif atau pengguna merasa puas, dan label negatif jika mengandung sentimen negatif atau pengguna merasa tidak puas. Label data ini digunakan untuk memberikan informasi tambahan dalam melatih model saat melakukan analisis sentimen. Proses pelabelan data dilakukan sebelum tahap *preprocessing* agar ulasan yang diberikan label dapat dipahami secara optimal oleh manusia. Hal ini dikarenakan kata-kata dalam kalimat masih utuh dan kalimat yang digunakan masih bersifat umum sehingga mudah dipahami.

Dalam pelabelan data untuk ulasan pengguna aplikasi Ruangguru, kriteria yang digunakan untuk menentukan sentimen positif atau negatif mencakup beberapa faktor. Ulasan dianggap memiliki sentimen positif jika pengguna mengekspresikan kepuasan mereka terhadap fitur-fitur aplikasi, kualitas konten pembelajaran, kemudahan penggunaan, atau responsifnya layanan pelanggan Ruangguru. Sebaliknya, ulasan diklasifikasikan sebagai sentimen negatif jika terdapat kritik terhadap performa aplikasi, kesulitan dalam navigasi, keluhan tentang layanan pelanggan yang lambat, atau masalah dengan konten pembelajaran yang disediakan. Pada dasarnya, penilaian

sentimen ini mencerminkan pengalaman pengguna dengan aplikasi Ruangguru.

#### D. Preprocessing

*Preprocessing* merujuk pada serangkaian langkah atau teknik yang diterapkan pada data mentah sebelum masuk ke proses analisis atau pemodelan. Tahap *preprocessing* diperlukan agar data yang diteliti dapat diprediksi dengan tepat, metode koreksi dalam tahap *preprocessing* yang berbeda dapat menghilangkan jenis artefak tertentu namun tetap meninggalkan beberapa efek untuk meningkatkan kualitas data [14]. Berikut langkah-langkah pada tahap *preprocessing* :

##### 1) Data Cleaning

*Data cleaning* adalah proses pengolahan data yang bertujuan untuk memeriksa dan memperbaiki data agar sesuai untuk dilakukan langkah lebih lanjut. Lalu dilakukan tahap *case folding* yang bertujuan untuk menyamakan semua huruf dalam dokumen teks menjadi huruf kecil [15]. Lalu dilakukan juga proses normalisasi kata, yaitu mengubah bahasa atau singkatan yang tidak beraturan menjadi kata dengan ejaan yang benar sehingga mengurangi gangguan yang dihasilkan oleh satu kata yang sama dengan berbagai bentuk [16].

##### 2) Tokenization

*Tokenization* adalah proses untuk membagi suatu kalimat, paragraf atau dokumen menjadi token berdasarkan suatu karakter tertentu [17] Sebagai contoh, tokenisasi dari kalimat "Saya naik motor roda tiga" menghasilkan lima token, yakni: "Saya", "naik", "motor", "roda", dan "tiga".

##### 3) Stopwords Removal

Teknik ini melibatkan perbandingan kata-kata dalam teks dengan daftar stopword, sehingga jika sebuah kata teridentifikasi dalam daftar tersebut, maka kata itu akan dihilangkan dari teks [18]. Proses penghapusan *stopwords*, atau *stopword removal*, bertujuan untuk mengeliminasi kata-kata yang memberikan informasi minimal dari teks, sehingga dapat lebih fokus pada kata-kata yang memiliki nilai informasi lebih tinggi [19].

##### 4) Stemming

*Stemming* adalah teknik yang digunakan untuk mengurangi kata-kata yang mengalami infleksi ke bentuk dasar mereka, meskipun bentuk dasar ini tidak selalu merupakan akar kata yang sebenarnya [20]

#### E. Pembagian Data

Pembagian data merupakan proses membagi data yang akan digunakan dalam penelitian menjadi dua atau lebih bagian. Biasanya, proses ini digunakan untuk menguji model atau algoritma dengan membagi dataset menjadi set pelatihan dan set pengujian, atau juga set validasi [21]. Dataset biasanya dibagi menjadi dua subset, yaitu *data training* dan *data testing*. *Data training* digunakan untuk melatih model, sementara *data testing*, yang belum pernah digunakan selama pelatihan, digunakan untuk mengevaluasi keberhasilan atau akurasi dari model yang telah dilatih [22].

#### F. Algoritma Support Vector Machine

*Support Vector Machine* adalah metode pembelajaran mesin yang populer karena berbagai keunggulan yang dimilikinya. SVM dipilih karena efektif dalam menemukan batas keputusan yang kompleks di antara kelas-kelas sentimen yang berbeda dalam data teks, yang sering kali memiliki pola yang kompleks. Kernel yang akan digunakan pada penelitian ini yaitu *kernel linear*. Sedangkan untuk parameternya, terdapat satu parameter yang dapat disesuaikan yaitu parameter 'C' dengan nilai 1. SVM merupakan metode *data mining* dalam kelas *supervised learning*. Tujuan utama dari teknik ini adalah untuk mencari *hyperplane* terbaik dari sekumpulan *hyperplane* yang ada. Metode *Support Vector Machine* dirancang untuk menemukan fungsi pemisah yang optimal, yang memungkinkan pemisahan efektif antara dua set data yang berasal dari dua kelas yang berbeda [23]. Persamaan *hyperplane* pada SVM dapat diselesaikan dengan persamaan berikut [24] :

$$f(x) = w^T \cdot x + b \quad (1)$$

Keterangan:

$w$  = vektor bobot (*weight*)

$x$  = vektor fitur input

$b$  = bias

#### G. Evaluasi dengan Confusion Matrix

*Confusion Matrix* digunakan dalam mengevaluasi kinerja model dengan cara mempresentasikan hasil dalam bentuk tabel matriks. Dalam dataset yang terdiri dari dua kelas, kelas pertama biasanya dianggap sebagai kelas

positif, sementara kelas kedua dianggap sebagai kelas negatif dengan tujuan membantu dalam mengukur akurasi serta kesalahan klasifikasi yang dilakukan oleh model [25]. Dengan menganalisis nilai *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN) dapat diperoleh pemahaman mendalam tentang model. Misalnya, evaluasi *False Positives* (FP) membantu mengidentifikasi ulasan yang salah diprediksi sebagai positif sehingga perlunya peningkatan dalam mengenali ulasan yang seharusnya memiliki sentimen negatif. Evaluasi *confusion matrix* juga menghasilkan metrik seperti *accuracy*, *precision*, *recall*, dan *f-measure*.

#### H. K-fold Cross Validation

*Cross validation* merupakan teknik validasi model yang digunakan untuk mengevaluasi kemampuan model statistik dalam menggeneralisasi terhadap data independen, terutama jika digunakan untuk memprediksi performa model dan memperkirakan akurasi model prediktif ketika diterapkan pada data asli [26]. Salah satu teknik dari *cross validation* yaitu *K-Fold Cross Validation*, di mana metode ini digunakan untuk mengevaluasi kinerja suatu algoritma. Dalam *K-Fold Cross Validation*, data dibagi menjadi k subset yang sama besar, yang disebut *fold* (lipatan), dan algoritma dievaluasi sebanyak k kali. Pada setiap iterasi, salah satu lipatan digunakan sebagai data pengujian, sementara lipatan lainnya digunakan sebagai data pelatihan, kemudian hasilnya berupa nilai akurasi dari algoritma terhadap data [27]. Dengan melihat rata-rata dan deviasi standar dari metrik akurasi pada setiap iterasi K-fold, peneliti dapat menilai seberapa konsisten model dalam performa di berbagai subset data. Jika deviasi standar rendah, ini menunjukkan bahwa model konsisten dalam melakukan prediksi yang stabil terhadap variasi dalam data uji. Perbedaan yang signifikan antara skor performa pada data pelatihan dan data uji dari *K-fold cross-validation* juga dapat mengindikasikan keberadaan *overfitting* atau *underfitting*.

#### I. Metode Latent Dirichlet Allocation

*Latent Dirichlet Allocation* adalah metode pemodelan topik yang memungkinkan identifikasi pola-pola tertentu dalam kumpulan dokumen, sehingga dokumen-dokumen tersebut dapat dikelompokkan ke dalam beberapa topik yang berbeda secara otomatis [28]. Sedangkan menurut K. B. Putra dan Kusumawardani, *Latent Dirichlet Allocation* adalah metode *text mining* yang berguna untuk mengidentifikasi pola dalam teks dokumen dengan menghasilkan berbagai topik berbeda [29]. LDA merupakan salah satu metode paling populer dalam *topic modelling* dan analisis topik saat ini. Metode ini sering digunakan dalam analisis dokumen yang memiliki ukuran sangat besar dan tidak teratur. Di sisi lain, LDA dipilih untuk pemodelan topik karena cocok digunakan untuk mengidentifikasi topik atau tema yang dominan dalam data teks ulasan yang besar dan kompleks. Pendekatan untuk menentukan jumlah topik yang digunakan dalam penelitian ini adalah dengan membangun banyak model LDA dengan nilai jumlah topik yang berbeda dan memilih salah satu yang memberikan nilai koherensi tinggi.

#### J. Evaluasi dengan Topic Coherence

*Topic Coherence* adalah pengujian yang digunakan untuk mengevaluasi seberapa konsisten atau koheren topik yang dihasilkan oleh model, dengan tujuan untuk menilai seberapa mudah topik tersebut dapat dipahami atau diinterpretasikan oleh manusia [30]. *Topic coherence* bertujuan untuk mengevaluasi model dengan mengukur tingkat kesamaan semantik dari kata-kata yang membentuk topik tersebut. *Topic coherence* diterapkan pada model LDA untuk membantu menentukan jumlah topik terbaik berdasarkan nilai koherensi topik yang dihasilkan.

#### K. Visualisasi Wordcloud

*Wordcloud* adalah teknik visualisasi yang menampilkan kumpulan kata-kata yang paling sering muncul dalam sebuah dataset setelah dianalisis [31]. Teknik ini digunakan untuk menyoroti kata-kata kunci atau topik yang paling dominan dalam suatu korpus teks. Pada dasarnya, *wordcloud* memberikan gambaran visual tentang frekuensi kata-kata dalam sebuah dokumen atau kumpulan dokumen.

### III. HASIL DAN PEMBAHASAN

#### A. Pengumpulan Data

Pengumpulan data dilakukan dengan metode *scraping* menggunakan *library google\_play\_scraper* dengan menjalankan *code* yang telah dibuat pada *platform Google Colab*. *Scraping* dilakukan pada data ulasan aplikasi Ruangguru yang diurutkan dari yang paling relevan kemudian hasilnya disimpan dalam format CSV. Jumlah data yang diperoleh mencapai 64.072 data dengan 11 atribut. Namun hanya sebanyak 30.070 data dan atribut *content* saja yang akan digunakan. Tabel 1 menampilkan sampel dari data mentah yang berhasil diambil.

TABEL I  
 SAMPEL DATA ULASAN APLIKASI RUANGGURU

re-viewId	user Name	userImage	content	score	thumbsUp Count	review-CreatedVersion	at	replyContent	re-pliedAt	ap-pVersion
22922a96039f4159ad049d09cb4b0058	Bang Jago	<a href="https://play-lh.googleusercontent.com/a/ACg8ocLqf6_jiaBuP7R74AEIv7EeQRrdFr0x6HeO4UV7dsSb=mo">https://play-lh.googleusercontent.com/a/ACg8ocLqf6_jiaBuP7R74AEIv7EeQRrdFr0x6HeO4UV7dsSb=mo</a>	Aplikasi ini sangat membantu kami siswa siswi yg melakukan kegiatan belajar mengajar dengan cara online (DARING) membantu kami agar bisa lebih memahami materi yg diberikan oleh guru kami terimakasih @ruangguru Semoga aplikasi ini bisa bermanfaat bagi kami semua	5	5	6.9.0	16-02-21 07:10	Hai. Terima kasih atas review dan bintang 5-nya. Senang rasanya jika Ruangguru bisa membantumu. Ruangguru akan terus berusaha untuk meningkatkan kualitas pendidikan sesuai dengan standar pendidikan yang ada di Indonesia. Tetap semangat belajar. Follow juga Instagram @ruangguru ya	16-02-21 07:43	6.9.0
f260fe da8cc5432781f188837ce4123d	M Daffa Yusman-syah	<a href="https://play-lh.googleusercontent.com/a/AL-VUjX7AL7bcVADxpL0hzPkWk9WRFpObgUcMxtsFTOATDwo kyQ">https://play-lh.googleusercontent.com/a/AL-VUjX7AL7bcVADxpL0hzPkWk9WRFpObgUcMxtsFTOATDwo kyQ</a>	saya pengguna ruang guru,kemarin baru saja beli paket sbmptn untuk 6 bulan,setelah pembelian waktu menonton vidio yang keluar cuma suaranya saja,tolong diperbaiki,saya jadi menyesal,dan waktu untuk tes sbmptn sudah mepet,kembalikan uang saya yang sia sia ini,olongggggggg	1	97	6.9.0	23-02-21 07:52	Halo Edelwais. Jangan cemas ya, Pastikan koneksi internetmu lancar dan stabil. Jika masih terkendala, bisa sampaikan detail kendalanya ke email layanan@ruangguru.co m dengan no.tiket APPRG20737 ya. Tetap semangat dan sukses selalu! Cheers, Sonia	23-02-21 08:14	6.9.0

### B. Pelabelan Data

Pelabelan dilakukan dengan bantuan model gpt-3.5-turbo-0125 dari Openai yang diakses menggunakan API, lalu dilakukan pengecekan secara manual untuk menghindari kesalahan dalam melakukan klasifikasi teks berdasarkan sentimennya masing-masing. Tabel 2 menampilkan sampel data dari hasil pelabelan yang sudah dilakukan.

TABEL 2  
 SAMPEL DATA HASIL PELABELAN

content	label
Aplikasi ini sangat membantu kami siswa siswi yg melakukan kegiatan belajar mengajar dengan cara online (DARING) membantu kami agar bisa lebih memahami materi yg diberikan oleh guru kami terimakasih @ruangguru Semoga aplikasi ini bisa bermanfaat bagi kami semua	positif
saya pengguna ruang guru,kemarin baru saja beli paket sbmptn untuk 6 bulan,setelah pembelian waktu menonton vidio yang keluar cuma suaranya saja,tolong diperbaiki,saya jadi menyesal,dan waktu untuk tes sbmptn sudah mepet,kembalikan uang saya yang sia sia ini,olongggggggg	negatif

### C. Preprocessing

Tujuan dari *preprocessing* adalah untuk membersihkan, menormalkan, dan mempersiapkan data sehingga lebih sesuai untuk analisis atau pemodelan yang dilakukan. Berikut tahap-tahap yang dilakukan pada proses ini :

#### 1) Data Cleaning

Proses *data cleaning* melibatkan langkah-langkah seperti menghapus data yang bernilai *null* atau duplikat, *case folding*, menghapus karakter tunggal, angka, spasi kosong, tanda baca, baris baru, dan emotikon, serta

normalisasi kata, seperti mengubah kata "bgs" menjadi "bagus". Tabel 3 menampilkan hasil sampel sebelum dan sesudah dilakukan proses *data cleaning*.

TABEL 3  
 SAMPEL DATA HASIL *DATA CLEANING*

sebelum	sesudah
Aplikasi ini sangat membantu kami siswa siswi yg melakukan kegiatan belajar mengajar dengan cara online (DARING) membantu kami agar bisa lebih memahami materi yg diberikan oleh guru kami terimakasih @ruangguru Semoga aplikasi ini bisa bermanfaat bagi kami semuaδY~S	aplikasi ini sangat membantu kami siswa siswi yang melakukan kegiatan belajar mengajar dengan cara online daring membantu kami agar bisa lebih memahami materi yang diberikan oleh guru kami terimakasih semoga aplikasi ini bisa bermanfaat bagi kami semua

## 2) *Tokenization*

*Tokenization* merupakan proses yang bertujuan untuk memecah teks menjadi unit-unit yang lebih kecil, seperti kata-kata atau frasa. Proses ini memungkinkan teks yang kompleks diganti menjadi format yang lebih terstruktur sehingga memudahkan komputer untuk melakukan proses selanjutnya. Tabel 4 menampilkan hasil sampel sebelum dan sesudah dilakukan proses *tokenization*.

TABEL 4  
 SAMPEL DATA HASIL *TOKENIZATION*

sebelum	sesudah
aplikasi ini sangat membantu kami siswa siswi yang melakukan kegiatan belajar mengajar dengan cara online daring membantu kami agar bisa lebih memahami materi yang diberikan oleh guru kami terimakasih semoga aplikasi ini bisa bermanfaat bagi kami semua	['aplikasi', 'ini', 'sangat', 'membantu', 'kami', 'siswa', 'siswi', 'yang', 'melakukan', 'kegiatan', 'belajar', 'mengajar', 'dengan', 'cara', 'online', 'daring', 'membantu', 'kami', 'agar', 'bisa', 'lebih', 'memahami', 'materi', 'yang', 'diberikan', 'oleh', 'guru', 'kami', 'terimakasih', 'semoga', 'aplikasi', 'ini', 'bisa', 'bermanfaat', 'bagi', 'kami', 'semua']

## 3) *Stopwords Removal*

Penghapusan *stopwords* merupakan proses mengeliminasi kata-kata umum yang sering muncul tetapi tidak memberikan kontribusi penting dalam analisis teks. *Stopwords removal* membantu menyaring teks sehingga analisis dapat lebih fokus terhadap kata-kata kunci yang lebih relevan. Tabel 5 menampilkan hasil sampel sebelum dan sesudah dilakukan proses *stopwords removal*.

TABEL 5  
 SAMPEL DATA HASIL *STOPWORDS REMOVAL*

sebelum	sesudah
['aplikasi', 'ini', 'sangat', 'membantu', 'kami', 'siswa', 'siswi', 'yang', 'melakukan', 'kegiatan', 'belajar', 'mengajar', 'dengan', 'cara', 'online', 'daring', 'membantu', 'kami', 'agar', 'bisa', 'lebih', 'memahami', 'materi', 'yang', 'diberikan', 'oleh', 'guru', 'kami', 'terimakasih', 'semoga', 'aplikasi', 'ini', 'bisa', 'bermanfaat', 'bagi', 'kami', 'semua']	['aplikasi', 'membantu', 'siswa', 'siswi', 'kegiatan', 'belajar', 'mengajar', 'online', 'daring', 'membantu', 'memahami', 'materi', 'terimakasih', 'semoga', 'aplikasi', 'bermanfaat']

## 4) *Steaming*

Proses *stemming* dilakukan dengan menghapus awalan dan/atau akhiran sebuah kata sehingga kata-kata yang memiliki kata dasar yang sama diubah menjadi bentuk yang seragam, hal ini mengurangi variasi kata dalam teks dan mampu meningkatkan konsistensi dalam analisis atau pemodelan teks. Tabel 6 menampilkan hasil sampel sebelum dan sesudah dilakukan proses *stemming*.

TABEL 6  
 SAMPEL DATA HASIL *DATA CLEANING*

sebelum	sesudah
['aplikasi', 'membantu', 'siswa', 'siswi', 'kegiatan', 'belajar', 'mengajar', 'online', 'daring', 'membantu']	aplikasi bantu siswa siswi giat ajar ajar online daring bantu paham materi terimakasih moga aplikasi manfaat

tu', 'memahami', 'materi', 'terimakasih', 'semoga',  
 'aplikasi', 'bermanfaat']

#### D. Pembagian Data

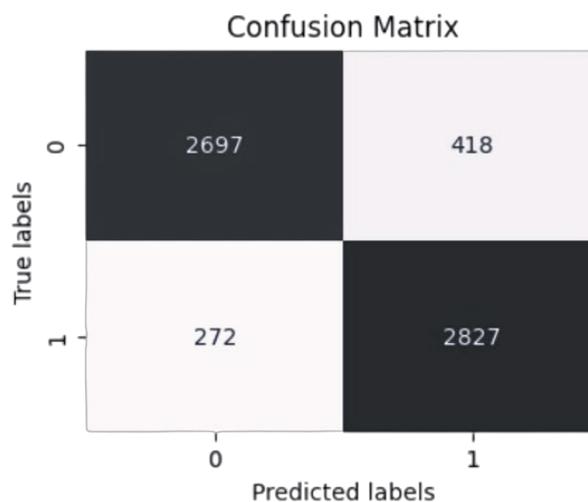
Pembagian data dilakukan menggunakan fungsi *train\_test\_split* yang didapat dari modul *model\_selection* yang terdapat dalam *library Scikit-learn*. Fungsi ini diaplikasikan untuk memecah dataset menjadi dua subset, yaitu *data training* dan *data testing*. Pembagian data berguna untuk melatih model pada *data training* dan menguji performanya pada *data testing*. Pada penelitian ini dataset dibagi menjadi *data testing* dan *data training* dengan perbandingan 80:20.

#### E. Implementasi Algoritma Support Vector Machine

Implementasi algoritma SVM dilakukan dengan memanfaatkan kelas *SVC* dari modul *svm* yang terdapat dalam *library Scikit-learn*. *SVC* bekerja dengan menemukan *hyperplane* yang membagi data ke dalam kelas-kelas yang berbeda dalam ruang fitur. *Hyperplane* terbaik akan ditentukan sehingga jarak antara *hyperplane* dengan titik-titik data (*support vectors*) dapat maksimal. Pada penelitian ini, *SVC* digunakan untuk klasifikasi biner (dua kelas) antara kelas ulasan dengan sentimen positif dan kelas ulasan dengan sentimen negatif. Kemudian peneliti memilih kernel yang akan digunakan pada penelitian ini yaitu kernel 'linear' dengan parameter 'C' bernilai 1.

#### F. Evaluasi dengan Confusion Matrix

Penelitian ini memakai teknik *confusion matrix* untuk mengevaluasi hasil prediksi dari model SVM dengan membandingkan data test asli dan prediksi. Hal ini bertujuan untuk melihat hasil *true negative*, *false positive*, *false negative*, dan *true positive*. Dari klasifikasi tersebut, nilai akurasi, presisi, *recall*, dan *F1-score* dapat dihitung. Dalam penerapan metode *confusion matrix*, peneliti menerapkan *library sklearn.metrics* dengan modul *confusion\_matrix*, *classification\_report*, dan *accuracy\_score*. Gambar 2 memperlihatkan hasil *confusion matrix* pada model SVM yang telah dilatih.



Gambar. 2. Hasil *Confusion Matrix*

Berikut merupakan hasil perhitungan *Confusion Matrix*.

##### 1) Akurasi

$$\text{Akurasi} = \frac{TP + TN}{\text{Total}} = \frac{2827 + 2697}{2697 + 418 + 272 + 2827} = \frac{5524}{6214} = 0.8889 \quad (2)$$

##### 2) Presisi

$$\text{Presisi} = \frac{TP}{TP + FP} = \frac{2827}{2827 + 418} = \frac{2827}{3245} = 0.8711 \quad (3)$$

### 3) Recall

$$Recall = \frac{TP}{TP + FN} = \frac{2827}{2827 + 272} = \frac{2827}{3099} = 0.9122 \quad (4)$$

### 4) F1-score

$$F1-score = \frac{2(precision \times recall)}{(precision + recall)} = \frac{2(0.8711 \times 0.9122)}{(0.8711 + 0.9122)} = \frac{1.5892}{1.7833} = 0.8911 \quad (5)$$

Berdasarkan hasil evaluasi model klasifikasi sentimen yang dilakukan, diperoleh nilai akurasi sebesar 88,89%, presisi sebesar 87,11%, *recall* sebesar 91,22%, dan *F1-score* sebesar 89,11%. Nilai-nilai ini mencerminkan efektivitas model dalam mengklasifikasikan sentimen ulasan dengan baik. Akurasi yang tinggi menandakan kemampuan model dalam memprediksi secara tepat. Sementara itu, presisi yang mencapai 87,11% menunjukkan bahwa model jarang melakukan kesalahan dalam mengidentifikasi ulasan yang sebenarnya positif. *Recall* yang tinggi sebesar 91,22% mengindikasikan bahwa model mampu mendeteksi sebagian besar ulasan positif yang ada. Secara keseluruhan, *F1-score* sebesar 89,11% menunjukkan keseimbangan yang baik antara presisi dan *recall*, menggambarkan performa yang solid dalam konteks klasifikasi sentimen ulasan.

### G. K-fold Cross Validation

Tahapan validasi model SVM dilakukan dengan pengujian menggunakan *k-fold cross validation* dengan *library sklearn* pada modul *model\_selection* kelas *cross\_val\_score* dan *KFold*. Data latih akan dibagi secara menjadi sejumlah bagian selaras dengan nilai *k*, dengan setiap bagian memiliki perbandingan yang sama besar. Selanjutnya, akan didapatkan akurasi untuk setiap bagian, dan rata-rata akurasi semua bagian dihasilkan dari penjumlahan akurasi per bagian dibagi nilai *k*. Eksperimen ini akan dilakukan sebanyak 10 kali, mulai dari *k*=1 hingga *k*=10. Hasil perhitungan dari setiap bagian dapat dilihat pada Tabel 7.

TABEL 7  
 HASIL PERHITUNGAN K-FOLD CROSS VALIDATION

<i>fold</i>	akurasi
<i>Fold 1</i>	0,8865
<i>Fold 2</i>	0,8934
<i>Fold 3</i>	0,9022
<i>Fold 4</i>	0,8873
<i>Fold 5</i>	0,8921
<i>Fold 6</i>	0,8897
<i>Fold 7</i>	0,8853
<i>Fold 8</i>	0,8905
<i>Fold 9</i>	0,8993
<i>Fold 10</i>	0,8801
<b>Rata-rata</b>	<b>0,8906</b>

Berdasarkan Tabel 7, pengujian dengan metode *K-fold Cross Validation* memiliki skor akurasi tertinggi pada *fold* ke-3, yaitu 90,22% dan secara keseluruhan rata-rata akurasi adalah 89,06%. Hal ini menunjukkan bahwa algoritma *Support Vector Machine* layak dan dapat digunakan dalam klasifikasi sentimen ulasan aplikasi Ruangguru.

### H. Implementasi Metode Latent Dirichlet Allocation

Implementasi LDA dilakukan dengan menggunakan kelas *LdaModel* dari modul *models* yang terdapat dalam *library gensim*. Dua model LDA terpisah dilatih untuk setiap kategori sentimen, yaitu positif dan negatif, dengan mencari topik-topik dominan dalam teks. Melalui pemilihan jumlah topik (*num\_topics*) dan iterasi pelatihan, model LDA memetakan teks ke dalam topik-topik yang mewakili konten teks secara keseluruhan. Pendekatan yang digunakan untuk menentukan jumlah topik adalah dengan membangun banyak model LDA dengan nilai jumlah topik yang berlainan dan menggunakan jumlah topik dengan nilai koherensi tertinggi. Hasil dari perhitungan nilai koherensi untuk setiap jumlah topik pada kelas sentimen positif dan negatif ditunjukkan pada Tabel 8.

TABEL 8  
 NILAI KOHERENSI TOPIK

jumlah topik	nilai koherensi	
	positif	negatif
2	0,3979	0,4269
3	0,4561	0,3766
4	0,4632	0,4899
5	0,4779	0,4495
6	0,4424	0,4659
7	0,4476	0,4663
8	0,4435	0,4636
9	0,3651	0,4559
10	0,4516	0,4387

Berdasarkan hasil perhitungan pada Tabel 8, pada kelas sentimen positif digunakan 5 topik utama, sedangkan pada kelas sentimen negatif digunakan 4 topik utama.

### I. Evaluasi dengan Topic Coherence

Evaluasi pemodelan topik dilakukan dengan menggunakan kelas *CoherenceModel* dari modul *models* yang terdapat dalam *library gensim*. Perhitungan koherensi topik ini berdasarkan pendekatan C\_V dengan empat tahap yaitu *Segmentation*, *Probability Estimation*, *Confirmation Measure*, dan *Aggregation*. Semakin tinggi koherensi topiknya, menandakan semakin mudah penafsirannya dilakukan oleh manusia. Berdasarkan hasil perhitungan, nilai koherensi untuk kelas positif dengan jumlah topik sebanyak 5 mendapatkan nilai 0,4779 menunjukkan bahwa model mampu mengidentifikasi dan mengelompokkan kata-kata yang terkait dengan topik-topik tertentu, serta menandakan bahwa masih terdapat beberapa tantangan dalam menjaga konsistensi antar topik. Sedangkan nilai koherensi untuk kelas negatif dengan jumlah topik sebanyak 4 mendapatkan nilai 0,4899 menunjukkan kemampuan yang baik dalam menghasilkan topik-topik yang konsisten serta mengindikasikan bahwa model dapat mengelompokkan kata-kata yang sering muncul bersama-sama dan memiliki keterkaitan yang kuat. Hal ini menunjukkan bahwa metode *Latent Dirichlet Allocation* layak digunakan untuk menemukan struktur tematik dalam kumpulan ulasan aplikasi Ruangguru pada setiap kelas sentimen.

### J. Visualisasi Wordcloud

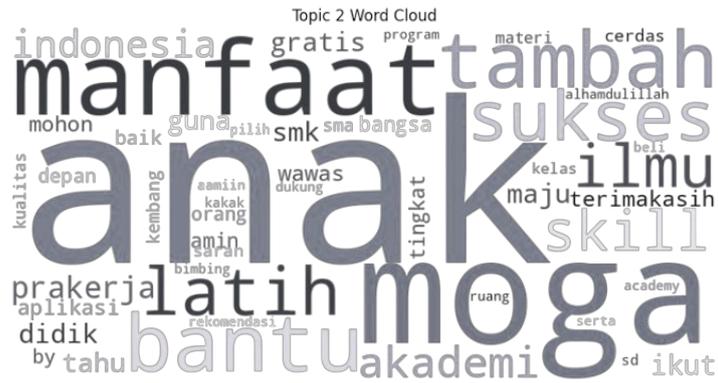
*Wordcloud* berguna untuk menampilkan kata yang mendominasi dalam suatu dataset secara visual. Kata yang divisualisasikan pada penelitian ini adalah kata yang terdapat dari hasil pemodelan topik pada setiap topik. Pada penelitian ini, *wordcloud* dibuat menggunakan kelas *WordCloud* dari *library wordcloud*. Setiap kata akan dihitung frekuensinya dan diambil 50 kata teratas dengan bobot tertinggi dalam topik tersebut untuk ditampilkan dalam sebuah figur menggunakan kelas *pyplot* dari *library matplotlib*. *Wordcloud* membantu dalam visualisasi distribusi kata dalam setiap topik, memungkinkan interpretasi yang lebih mudah dan cepat terhadap kata-kata kunci yang membentuk topik tersebut dan menentukan topik utama pada setiap topik yang ada. Visualisasi ditampilkan dalam bentuk *wordcloud* ditunjukkan oleh Gambar 3 sampai Gambar 11. Pada Gambar 3 sampai Gambar 7 menunjukkan visualisasi untuk setiap topik dalam kelas sentimen positif.



Gambar. 3. Visualisasi Topik "0" Kelas Positif



Gambar. 4. Visualisasi Topik “1” Kelas Positif



Gambar. 5. Visualisasi Topik “2” Kelas Positif



Gambar. 6. Visualisasi Topik “3” Kelas Positif



Gambar. 7. Visualisasi Topik “4” Kelas Positif





Gambar. 11. Visualisasi Topik “3” Kelas Negatif

Pada visualisasi sebaran kata-kata dalam setiap topik yang diperlihatkan pada kelas negatif, terdapat 4 topik yang dapat diinterpretasikan. Hasil interpretasi pada setiap topik untuk kelas sentimen negatif adalah Topik “0” mengandung informasi berupa permintaan pengguna untuk dapat disediakan materi untuk lebih banyak jenjang sekolah dan mata pelajaran. Topik “1” mengandung informasi mengenai keluhan pengguna tentang biaya berlangganan yang dianggap mahal. Topik “2” mengandung kritik tentang video pembelajaran, kualitas aplikasi secara keseluruhan, serta fitur yang ada. Topik “3” mengandung keluhan pengguna mengenai iklan aplikasi yang muncul di televisi dan promosi lainnya melalui email dan Whatsapp.

#### K. Analisis Hasil

Hasil model SVM yang ada digunakan untuk mengklasifikasikan sentimen pada 10.000 data ulasan baru. Hasilnya menunjukkan bahwa terdapat 3.950 ulasan atau 39,5% yang menunjukkan sentimen negatif. Namun, sebagian besar dari ulasan tersebut, yaitu 6.050 ulasan atau 60,5%, menunjukkan sentimen positif. Algoritma SVM dalam mengklasifikasikan sentimen memiliki kelebihan dan keterbatasan tersendiri. Kelebihannya terletak pada efektivitasnya dalam menangani kata-kata kunci dalam ulasan, serta efektif dalam menangani *outlier*. Namun, penggunaan SVM dapat memakan waktu komputasi yang signifikan, terutama pada dataset besar. Penelitian terkait yang berjudul Analisis Sentimen pada Aplikasi Grab di *Google Play Store* Menggunakan *Support Vector Machine* menghasilkan nilai akurasi model sebesar 85,54% dengan dataset berjumlah 1.000 ulasan [32]. Hasil tersebut memperlihatkan perbedaan nilai akurasi dengan model SVM pada penelitian ini yaitu sebesar 88,89% dengan dataset berjumlah 64.072 ulasan. Analisis sentimen ini dapat digunakan untuk memahami perasaan dan kepuasan pengguna terhadap aplikasi. Dengan mengetahui apakah ulasan pengguna cenderung positif, atau negatif, perusahaan dapat mengidentifikasi area yang memerlukan perbaikan dan aspek yang perlu dipertahankan atau ditingkatkan

Sedangkan pada penelitian yang berjudul Analisis Pemodelan Topik untuk Ulasan tentang PeduliLindungi, mendapatkan nilai koherensi sebesar 0,3936 untuk 5 topik yang berhasil diidentifikasi menggunakan metode LDA [33]. Hasil tersebut memperlihatkan perbedaan nilai koherensi pada penelitian ini yaitu sebesar 0,4779 untuk 5 topik pada kelas sentimen positif dan sebesar 0,4899 untuk 4 topik pada kelas sentimen negatif dalam melakukan pemodelan topik dengan menggunakan metode LDA. Hasil pemodelan topik ulasan aplikasi Ruangguru ini dapat memberikan manfaat bagi perusahaan, hasil ini dapat digunakan untuk meningkatkan kualitas produk dengan mengidentifikasi masalah umum dan keluhan pengguna. Hal ini memungkinkan tim pengembang untuk lebih fokus dalam melakukan perbaikan dan pengembangan fitur yang sesuai dengan kebutuhan pengguna. Selain itu, perusahaan dapat memahami aspek positif yang paling dihargai oleh pengguna, sehingga dapat mempertahankan dan meningkatkan fitur-fitur tersebut untuk memberikan pengalaman pengguna yang lebih baik.

#### IV. KESIMPULAN

Penelitian ini sukses melakukan analisis sentimen menggunakan algoritma *Support Vector Machine* (SVM) dan pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) untuk ulasan pengguna aplikasi Ruangguru di Google Play. Hasil pengujian menunjukkan bahwa SVM mampu mengklasifikasi sentimen pengguna dengan baik, dengan akurasi 88,89%, presisi 87,11%, recall 91,22%, dan F1-score 89,11%. Analisis terhadap 10.000 ulasan baru menunjukkan 6.050 ulasan positif dan 3.950 ulasan negatif, mengindikasikan mayoritas tanggapan positif. Pemodelan topik dengan LDA mengidentifikasi 5 topik utama pada sentimen positif dengan nilai koherensi 0,4779, yang berfokus pada pengalaman positif dan kegunaan aplikasi dalam membantu proses belajar. Pada sentimen negatif, ditemukan 4 topik utama dengan nilai koherensi 0,4899, yang berfokus pada keluhan tentang kurang

lengkapnnya materi pembelajaran.

Penelitian selanjutnya disarankan untuk meningkatkan jumlah data dan mengatasi kata-kata ambigu atau satir untuk meningkatkan akurasi model. Selain itu, disarankan untuk menggunakan metode evaluasi lain selain nilai koherensi untuk menguji validitas topik LDA.

#### DAFTAR PUSTAKA

- [1] S. B. Gupta dan M. Gupta, "Technology and e-learning in higher education," *International Journal of Advanced Science and Technology*, vol. 29, no. 4, 2020.
- [2] Ruangguru, "Tentang Ruangguru." Diakses: 11 Oktober 2023. [Daring]. Tersedia pada: <https://www.ruangguru.com/about-us>
- [3] B. W. Arianto dan G. Anuraga, "Pemodelan Topik Pengguna Twitter Mengenai Aplikasi 'Ruangguru' Application," *Jurnal ILMU DASAR*, vol. 21, no. 2, 2020.
- [4] U. M. Lida dan I. Eliya, "Peran Startup Digital 'Ruangguru' Sebagai Metode Long Distance Learning dalam Pembelajaran Bahasa," *Jurnal Edul-lingua*, vol. 6, no. 2, 2019.
- [5] N. H. Tien dkk., "Factors impacting customer satisfaction at Vietcombank in Vietnam," *Himalayan Journal Economics and Business Management*, vol. 2, no. 4, 2021.
- [6] K. Khadka dan S. Maharjan, "Customer satisfaction and customer loyalty, Master's Thesis," *Centria University of Applied Sciences, Business Management, Lahti, Finland*, no. November, 2017.
- [7] A. Novantirani, M. K. Sabariah, dan V. Effendy, "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine," *e-Proceeding of Engineering*, vol. 2, no. 1, 2015.
- [8] B. Brahim, M. Touahria, dan A. Tari, "Improving sentiment analysis in Arabic: A combined approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 10, 2021, doi: 10.1016/j.jksuci.2019.07.011.
- [9] I. Rozi, S. Pramono, dan E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," *Jurnal EECIS*, vol. 6, no. 1, 2012.
- [10] D. T. Lukmana, S. Subanti, dan Y. Susanti, "Analisis Sentimen Terhadap Calon Presiden 2019 Dengan Support Vector Machine Di Twitter," *Seminar Nasional Penelitian Pendidikan Matematika (SNP2M) 2019 UMT*, no. 2002, 2019.
- [11] A. A. Annur, A. A. Murtopo, dan N. Fadilah, "Analisis Sentimen Aplikasi E-Learning Quipper Selama Pandemi Covid-19 Dengan Menggunakan Metode Support Vector Machine (SVM)," *Indonesian Journal of Informatics and Research*, vol. 3, no. 2, 2022.
- [12] P. Kherwa dan P. Bansal, "Topic Modeling: A Comprehensive Review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, 2020, doi: 10.4108/eai.13-7-2018.159623.
- [13] Y. Kustiyahningsih dan Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink," *Teknika*, vol. 13, no. 1, hlm. 127-136, Mar 2024, doi: 10.34148/teknika.v13i1.746.
- [14] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, dan D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC - Trends in Analytical Chemistry*, vol. 132, 2020. doi: 10.1016/j.trac.2020.116045.
- [15] M. A. Rosid, A. S. Fitriani, I. R. I. Astutik, N. I. Mulloh, dan H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," dalam *IOP Conference Series: Materials Science and Engineering*, 2020. doi: 10.1088/1757-899X/874/1/012017.
- [16] T. L. Nikmah, M. Z. Ammar, Y. R. Allatif, R. M. P. Husna, P. A. Kurniasari, dan A. S. Bahri, "Comparison of LSTM, SVM, and naive bayes for classifying sexual harassment tweets," *Journal of Soft Computing Exploration*, vol. 3, no. 2, 2022, doi: 10.52465/josce.v3i2.85.
- [17] T. A. L. Jaya dan M. Ayub, "Pengembangan Knowledge Management System dengan Teknik Information Retrieval," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, 2021, doi: 10.28932/jutisi.v7i1.3316.
- [18] A. P. Wibawa, F. Miftahuddin, dan S. Suyono, "K-Medoids Clustering untuk Pembentukan Database Stopword Bahasa Jawa," *Ranah: Jurnal Kajian Bahasa*, vol. 10, no. 2, 2021, doi: 10.26499/rh.v10i2.2125.
- [19] Muhammad Harris Syafa'at, E. R. Setyaningsih, dan Y. Kristian, "SVM UNTUK SENTIMENT ANALYSIS CALON KEPALA DAERAH BERDASAR DATA KOMENTAR VIDEO DEBAT PILKADA DI YOUTUBE," *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 15, no. 2, 2021, doi: 10.35457/antivirus.v15i1.1539.
- [20] E. Mulyani, F. P. B. Muhammad, dan K. A. Cahyanto, "Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, 2021, doi: 10.30865/mib.v5i1.2672.
- [21] A. Putri dkk., "Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, 2023, doi: 10.57152/malcom.v3i1.610.
- [22] E. S. R. Br. Situmorang, M. K. Anam, R. Rahmaddeni, dan A. N. Ulfah, "Perbandingan Algoritma Svm Dan Nbc Dalam Analisa Sentimen Pilkada Pada Twitter," *CSRID (Computer Science Research and Its Development Journal)*, vol. 13, no. 3, 2021, doi: 10.22230/csrid.13.3.2021.169-179.
- [23] K. R. Kavitha, K. V. Aiswarya Rajan, dan A. Pillai, "An Improved Feature Selection and Classification of Gene Expression Profile using SVM," dalam *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2019*, 2019. doi: 10.1109/ICICICT46008.2019.8993358.
- [24] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, dan A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, 2020, doi: 10.1016/j.neucom.2019.10.118.
- [25] D. Putra dan A. Wibowo, "Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes," *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, vol. 2, 2020.
- [26] F. Tempola, M. Muhammad, dan A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 5, 2018, doi: 10.25126/jtiik.201855983.
- [27] A. C. Rumahorbo dan K. A. Sekarwati, "Penerapan Data Mining Dengan Menggunakan Algoritma C4.5 Pada Klasifikasi Fasilitas Kesehatan Provinsi Di Indonesia," *Jurnal Ilmiah Komputasi*, vol. 19, no. 1, 2020, doi: 10.32409/jikstik.19.1.153.
- [28] P. Patmawati dan M. Yusuf, "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 3, 2021, doi: 10.47065/bits.v3i3.1012.
- [29] K. B. Putra dan R. P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)," *Jurnal Teknik ITS*, vol. 6, no. 2, 2017, doi: 10.12962/j23373539.v6i2.23205.
- [30] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, dan D. M. Blei, "Reading tea leaves: How humans interpret topic models," dalam *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 2009.
- [31] R. F. Amin, "IMPLEMENTASI RESTFULL API MENGGUNAKAN ARSITEKTUR MICROSERVICE UNTUK MANAJEMEN TUGAS KULIAH(STUDI KASUS: MAHASISWA STMIK AKAKOM)," *STMIK AKAKOM Yogyakarta*, 2020.
- [32] R. Wahyudi dan G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *Jurnal Informatika*, vol. 8, no. 2, 2021, doi: 10.31294/ji.v8i2.9681.
- [33] D. L. C. Pardede dan M. A. I. Waskita, "ANALISIS PEMODELAN TOPIK UNTUK ULASAN TENTANG PEDULI LINDUNGI," *Jurnal Ilmiah Informatika Komputer*, vol. 28, no. 1, 2023, doi: 10.35760/ik.2023.v28i1.7925.