Jurnal Sains dan Teknologi
Volume 12 Number 3, Tahun 2023, pp. 815-822
P-ISSN: 2303-3142 E-ISSN: 2548-8570
Open Access: https://doi.org/10.23887/jstundiksha.v12i3.68492

Topic Modelling Latent Dirichlet Allocation untuk Klasifikasi Komentar pada Layanan *Streaming Platform*

Noorhanida Royani[1*], Catur Edi Widodo[2], Budi Warsito[3]

[1] Sistem Informasi, Universitas Diponegoro, Semarang, Indonesia

ARTICLE INFO

Article history:
Received April 28, 2023
Accepted October 11, 2023
Available online October 25, 2023

Kata Kunci:
Latent Dirichlet Allocation, Klasifikasi Komentar, *Streaming Platform*

Keywords:
Latent Dirichlet Allocation, Comment Classification, Streaming Platform

ABSTRAK

Seiring dengan berkembangnya teknologi, memunculkan banyak platform online untuk streaming film. Streaming platform banyak digunakan masyarakat seperti netflix, disney+, hbo go, we tv, vidio. Banyaknya perbandingan antar streaming platform menjadi perbincangan dimedia sosial yaitu twitter. Opini yang disampaikan pengguna streaming platform berisi komentar positif dan komentar negatif yang mempengaruhi pengguna lainnya yang ingin menonton film. Penelitian ini dilakukan untuk mengkaji perbandingan antara komentar positif dan komentar negatif pengguna streaming platform pada media sosial Twitter. Metode *Latent dirichlet allocation* dapat digunakan sebagai topic modelling dan *Support Vector Machine* untuk klasifikasi. Pada tahapan pengambilan data dengan menggunakan tools framework scrapy dengan python, data diambil sebanyak 5.000 dan dilakukan preprocessing text. Metode LDA dapat mempresentasikan topik dan dokumen serta klasifikasi menggunakan *Support Vector Machine* (SVM) mendapatkan hasil komentar positif lebih banyak dari pada komentar negatif. Hasil evaluasi preforma didapatkan nilai akurasi 0,88, recall 0,88, F1score 0,87, precision 0,88. Topic *Modelling Latent Dirichlet Allocation* (LDA) untuk Klasifikasi Komentar pada Layanan Streaming Platform dengan menggunakan 5,000 data diambil dari sosial media yaitu twitter yang terbagi menjadi komentar positif dan komentar negatif. Hasil ini dipengaruhi dari jumlah komentar positif yang lebih dominan dari pada komentar negatif. Implikasi dari penelitian ini adalah pentingnya memperhatikan keseimbangan data dalam melakukan klasifikasi komentar pada platform streaming agar hasil prediksi klasifikasi dapat lebih akurat.

ABSTRACT

Along with the development of technology, gave rise to many online platforms for streaming movies. Streaming platforms are widely used by people such as Netflix, Disney+, HBO Go, We TV, Vidio. The number of comparisons between streaming platforms has become a conversation on social media, namely Twitter. Opinions submitted by streaming platform users contain positive comments and negative comments that affect other users who want to watch movies. This study was conducted to examine the comparison between positive comments and negative comments of streaming platform users on social media Twitter. The Latent dirichlet allocation method can be used as topic modelling and Support Vector Machine for classification. At the stage of data retrieval using scrapy framework tools in python, 5,000 data were taken and text preprocessing was carried out. The LDA method can present topics and documents and classifications using the Support Vector Machine (SVM) get more positive comments than negative comments. The results of the preforma evaluation obtained an accuracy value of 0.88, recall 0.88, F1score 0.87, precision 0.88. Topic Modelling Latent Dirichlet Allocation (LDA) for Comment Classification on Streaming Platform Services using 5,000 data taken from social media, namely Twitter, which is divided into positive comments and negative comments. This result is influenced by the number of positive comments which are more dominant than negative comments. The implication of this study is the importance of paying attention to data balance in classifying comments on streaming platforms so that the results of classification predictions can be more accurate.

1. PENDAHULUAN

Berkembangnya teknologi di Indonesia, banyak bermunculan platform untuk streaming film. Fenomena media streaming mengindikasikan bahwa situasi terkini secara teknologi menyebabkan proses komukasi bermedia dapat berlangsung secara meluas (Anggraeni, 2019; Imron, 2018). Saat ini streaming banyak diminati oleh masyarakat dari berbagai kalangan. Streaming merupakan teknologi yang menampilkan video atau audio dalam bentuk terkompresi melalui internet yang ditampilkan secara terus-menerus. Streaming platform yang banyak digunakan oleh masyarakat yaitu netflix, disney+, hbo go, we tv, video. Banyaknya platform tersebut mempengaruhi masyarakat untuk memilih platform streming yang akan digunakan. Perkembangan teknologi sosial media yaitu twitter sebagai media untuk mewadahi opini

*Corresponding author.
E-mail addresses: hanidaroy@gmail.com (Noorhanida Royani)

masyarakat terhadap suatu film (Gifari et al., 2022; Ramadhan & Ramadhan, 2022).

Selama ini permasalahan streaming banyak menimbulkan opini di masyarakat. Banyaknya perbandingan antar streaming platform menjadi perbincangan hangat di media sosial. Streaming film membuat persoalan yang sensitive di era digital sekarang menjadi semakin kompleks (Anshari, 2019; Wibowo, 2018). Opini publik merupakan kumpulan pendapat individu terhadap masalah tertentu yang mempengaruhi suatu kelompok atau masyarakat (Kurnia & Mella, 2018; Minerva et al., 2020). Opini yang disampaikan pengguna streaming platform sangat mempengaruhi pengguna lainnya yang ingin menonton film. Penyampaian opini menjadi informasi positif dan negatif yang diterima masyarakat sehingga menyebar luas dimedia sosial yaitu twitter. Twitter salah satu sosial media paling populer yang dapat dijadikan sumber data untuk teks analisis (Ferdiana et al., 2019; Putranti & Winarko, 2014). Penyampaian opini dari segi komentar positif dan negatif menimbulkan permasalahan dan dapat mempengaruhi pengguna yang ini menonton film melalui streming platform.

Analisis sentimen sebagai teknik mengetahui, mengekstrak, dan menjalankan informasi tekstual secara otomatis untuk memperoleh keterangan sentimental dalam mengungkapkan tanggapan (Gifari et al., 2022; Ramadhan & Ramadhan, 2022). Text mining bertujuan menemukan pola unik dalam dokumen teks berjumlah besar. Probabilistic topik modeling merupakan serangkaian algoritma yang menemukan topik utama dari himpunan dokumen besar yang tidak terstruktur (Hua et al., 2020; Vulić et al., 2015). Latent dirichlet allocation (LDA) yang mempresentasikan topik dengan probabilitas kata, Latent Dirichlet Allocation dikenal dengan kemampuan dan stabilitas yang baik dalam menangani data skala besar karena memberikan parameter sebagai variabel acak (Annisa & Surjandari, 2019; I. R. Putri & Kusumaningrum, 2017). Untuk klasifikasi menggunakan support vector machine menemukan fungsi optimum untuk di pergunakan dalam memisah antar kelas atau dua kelas data yang beda serta membangun model yang baik untuk data. Sehingga svm memiliki kemampuan generalisasi tinggi dan klasifikasi akurasi lebih stabil (Hermanto et al., 2020; Indrayuni, 2019).

Penggunaan Latent Dirichlet Allocation dalam klasifikasi komentar dapat membantu meningkatkan efektivitas dalam mengelompokkan komentar ke dalam kategori atau topik yang sesuai. Ini dapat membuatnya lebih mudah untuk memahami apa yang dibicarakan oleh pengguna dalam komentar mereka (Bustami & Noviaristanti, 2022; Santoso et al., 2022). *Latent Dirichlet Allocation* dapat membantu dalam identifikasi komentar spam atau yang tidak relevan dengan topik diskusi. Dengan demikian, ini dapat membantu dalam membersihkan komentar yang tidak diinginkan dari platform streaming. Dengan klasifikasi komentar yang lebih baik, pengguna dapat dengan mudah menavigasi dan menemukan konten atau komentar yang paling relevan dengan minat mereka. Ini dapat meningkatkan pengalaman pengguna secara keseluruhan (Roiqoh et al., 2023; Widodo, A. O. et al., 2023).

Selain klasifikasi topik, *Latent Dirichlet Allocation* juga dapat digunakan untuk menganalisis komentar yang sentimen, seperti apakah komentar tersebut positif, negatif, atau netral. Ini dapat membantu platform streaming dalam memahami reaksi pengguna terhadap konten atau acara tertentu. Berdasarkan klasifikasi komentar, platform streaming dapat memberikan rekomendasi yang lebih personal kepada pengguna berdasarkan minat dan preferensi mereka, yang dapat meningkatkan retensi pengguna (Bustami & Noviaristanti, 2022; Shidqi & Febrianta, 2023). Analisis topik komentar dapat memberikan wawasan berharga kepada penyedia layanan streaming tentang apa yang paling diminati oleh pengguna mereka dan area mana yang mungkin perlu ditingkatkan.

Pada tahun 2020, tentang Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine dengan menggunakan komentar positif dan negatif pada trailer film dan hasil penelitian tersebut memberikan gambaran informasi umum kepada calon penonton tentang film yang akan ditonton. Dari metode tersebut memberikan kinerja yang baik dalam melakukan analisis komentar positif dan negatif untuk semua genre film dari video trailer film Indonesia (Khomsah, 2021; Royyan & Setiawan, 2022). Pada tahun 2021, tentang Topic modeling using latent dirichlet allocation on twitter data with Indonesia keyword dengan menggunakan analisis berita yang diambil dari twitter secara otomatis dan hasil penelitian tersebut dapat mengklasifikasi teks berdasarkan topik yang digunakan untuk meringkas, mengelompokkan, dan menghubungkan atau mengolah data yang besar serta menghasilkan daftar topik yang berbobot pada setiap dokumen. Metode tersebut dapat menganalisis dokumen yang sangat besar dan LDA ini dapat memastikan bahwa model topik yang dihasilkan pada dokumen sudah benar, baik berupa topik maupun kata-kata dalam topik (Musliadi et al., 2022; Negara & Triadi, 2021).

Pada penelitian ini dapat dilakukan analisis sentimen untuk memberikan wawasan tentang pandangan opini dan ungkapan perasaan dari pengguna layanan platform streaming. Dengan menggunakan opini negatif dan positif di ambil dari setiap platform streaming yaitu netflix, disney+, hbo go, we tv, video dan menggunakan metode LDA untuk mempresentasikan topik yang berisikan perwakilan kata. Kata yang didapat dari topik dapat memberikan hasil dan gambaran bahwa ada topik tertentu yang paling banyak dibahas oleh pengguna. Klasifikasi pada penelitian ini dilakukan untuk mengetahui apakah komentar positif

atau negatif yang paling banyak dibahas dari data twitter yang didapatkan. Tujuan penelitian ini bahwa opini dari segi komentar positif dan negatif dapat dijadikan analisis sentimen untuk mengungkapkan dan memberi informasi dari tanggapan pengguna terhadap platform streaming.

2. METODE

Analisis data menggunakan beberapa tahapan seperti pengumpilan data, preprocessing text, pembobotan kata, topic modelling latent dirichlet allocation dan klasifikasi Suport Vector Machine (SVM). Proses ini menggunakan bahasa pemprograman phyton. Pada tahap preprocessing text dengan menggunakan punctuation removal untuk menghilangkan komponen tidak relevan yang tidak berisikan informasi. Contohnya pada tanda (“”, “?”, “}”, “()”) dan white space lebih, proses kedua menggunakan lowercasing untuk mengubah huruf kapital menjadi menjadi huruf kecil, proses ketiga tokenizing yaitu pemotongan sebuah kalimat menjadi bagian-bagian, proses keempat stemming mengubah kata yang mempunyai imbuhan jadi sebuah kata dasar, proses kelima stopword removal menghapus suatu kata kurang penting dan tidak memiliki informasi yang dibutuhkan (Hafidz & Liliana, 2021; Kusairi & Agustian, 2022). Hasil preprocessing text ditampilkan pada Tabel 1.

Tabel 1. *Preprocessing Text*

Tahapan *Preprocessing Text*	Teks
Kalimat Awal	Nonton film bareng Teman2 dan anak dirumah, Banyak pilihan Film yang bisa ditonton … #online #rekomendasi judul Terbaru
Punctuational Removal	Nonton film bareng Teman dan anak dirumah Banyak pilihan Film yang bisa ditonton online rekomendasi judul Terbaru
Lowercasing	nonton film bareng teman dan anak dirumah banyak pilihan film yang bisa ditonton online rekomendasi judul terbaru
Tokenizing	[‘nonton’, ‘film’, ‘bareng’, ‘teman’, ‘dan’, ‘anak’, ‘dirumah’, ‘banyak’, ‘pilihan’, ‘film’, ‘yang’, ‘bisa’, ‘ditonton’, ‘online’, ‘rekomendasi’, ‘judul’, ‘terbaru’]
Stemming	[‘nonton’, ‘film’, ‘bareng’, ‘teman’, ‘dan’, ‘anak’, ‘rumah’, ‘banyak’, ‘pilih’, ‘film’, ‘yang’, ‘bisa’, ‘tonton’, ‘online’, ‘rekomendasi’, ‘judul’, ‘baru’]
Stopword Removal	[‘nonton’, ‘film’, ‘teman’, ‘anak’, ‘rumah’, ‘banyak’, ‘pilih’, ‘film’, ‘online’, ‘rekomendasi’, ‘judul’, ‘baru’]

Dalam proses pembobotan kata menggunakan library Scikit-learn atau SK-Learn, modul yang digunakan adalah countvectorize Scikit-learn untuk melakukan konversi pada kumpulan dokumen teks menjadi vector jumlah kata atau token, sederhananya berfungsi mengubah teks menjadi matriks. Adapun modul lain yang digunakan yaitu tfidf transformer yang berfungsi untuk mengubah kumpulan dokumen, cara kerja dari modul ini adalah dengan menghitung jumlah kata secara sistematis menggunakan countvectorizer dan menghitung nilai inverse document frequency serta menghitung skor totalnya. Hasil pembobotan kata ditampilkan pada Tabel 2.

Tabel 2. Pembobotan Kata

No.	Kata	Pembobotan Kata
1	Seru	0,1231
2	Momen	0,1906
3	Cerita	0,1341
4	Semua	0,173
5	Film	0,1801
6	Drama	0,115
7	Bagus	0,1843
8	Wajib	0,1335

Untuk melakukan topic modeling, data dirubah dalam bentuk dictionary dan corpus. Dictionary merupakan format data yang mengandung himpunan perwakilan kata yang diberi indeks sehingga dapat memudahkan dalam menampilkan kata yang termasuk dalam model. Corpus merupakan format data yang berbentuk dokumen term matrix yang digunakan dalam melakukan pembentukan model nantinya. Kamus dan corpus adalah dua input untuk topik modelling LDA yang dibuat menggunakan modul gensim corpora.

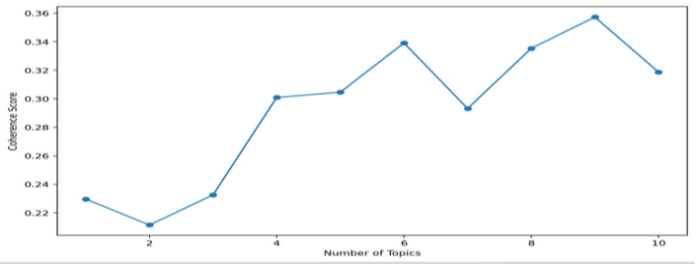
Untuk menentukan jumlah topik atau num topics dengan menggunakan modul gensim coherencemodel dapat divisualisasikan dalam bentuk grafik coherence score. Coherence score adalah ukuran untuk menentukan jumlah topik dengan nilai tertinggi yang akan diambil. Parameter yang digunakan dalam topik modelling ini yaitu corpus, dictionary, jumlah topik atau num topics, iterasi.

Untuk *support vector machine* (SVM), pembobotan kata dan label menjadi input pada klasifikasi. Data di klasifikasikan menjadi positif dan negatif menggunakan support vector machine (SVM). Dalam pembuatan support vector machine terdapat beberapa library yang digunakan yaitu sklearn svm untuk melatih data. Pada klasifikasi, data di lakukan pelabelan positif dengan nilai 1 dan pelabelan negatif dengan nilai -1. Pengujian performa dapat dilakukan dengan menggunakan confusion matrix untuk mengevaluasi performa dengan hasil akurasi, precision, recall dan f-1 score berdasarkan dataset yang digunakan.

3. HASIL DAN PEMBAHASAN

Hasil

Topic modelling latent dirichlet alloaction menghasillkan nilai coherence tertinggi. Adapun hasil dari nilai coherence pada Gambar 1.



Gambar 1. Hasil *Coherence Score*

Pada grafik nilai coherence, dapat dilihat bahwa nilai grafik tertinggi pada topik ke 9 dengan nilai 0,3573 , sehingga pada penulisan ini digunakan 9 topik pada nilai coherence. Nilai coherence tertinggi dapat dijadikan acuan dalam topic modelling ini. Topic modelling dapat menemukan nilai kata terhadap topik dan topik terhadap dokumen. Adapun hasil dari topik terhadap kata pada Tabel 3.

Tabel 3. Hasil Topik terhadap Kata

Jumlah Topik	Hasil Kata terhadap Topik
Topik 1	netflix 0,0178 + disney 0,0149 + indo 0,0115 + video 0,0083 + viral 0,0083 + link 0,0074 + little 0,0066 + film 0,0057 +series 0,0051
Topik 2	disney 0,0183 + hbo 0,0151 + go 0,0147 + film 0,0144 + netflix 0,0130 + vidio 0,0128 + series 0,0060 + seru 0,0055 + skandal 0,0051
Topik 3	disney 0,0276 + tv 0,0155 + hotstar 0,0108 + netflix 0,0106 + video 0,0103 + vidio 0,0100 + banget 0,0079 + film 0,0064 + indihome 0,0062
Topik 4	video 0,0217 + netflix 0,0172 + banget 0,0141 + film 0,0090 + main 0,0083 + random 0,0078 + disney 0,0075 + series 0,0070+ nomor 0,0058
Topik 5	bagus 0,0089 + disney 0,0055 + member 0,0053 + hbo 0,0051 + he 0,0049 + model 0,0045 + seru 0,0037 + favorit 0,0036 + banget 0,0032
Topik 6	netflix 0,0182 + kak 0,0074 + youtube 0,0072 + full 0,0061, sayap 0,0054 + film 0,0048 + disney 0,0047 + tv 0,0047 + bio 0,0046
Topik 7	nonton 0,0637 + game 0,0291 + vidio 0,0269 + netflix 0,0177 + film 0,0096 + piala 0,0075, hbo 0,0071 + dunia 0,0064 + asik 0,0128
Topik 8	netflix 0,0119 + hbo 0,0108 + keren 0,0095 + disney 0,0092, banget 0,0089, sinyal 0,0070, buruk 0,0063, susah 0,0061 + bayar 0,0087
Topik 9	disney 0,0306 + netflix 0,0143 + black 0,0132 + panther 0,0103 + film 0,0084 + video 0,0064 + bioskop 0,0057 + illegal 0,0055 + gratis 0,0095

Pada hasil Tabel 3, dapat diketahui label aspek setiap topik seperti pada topik 1 label aspek tertinggi yaitu netflix 0,0178, pada topik 2 yaitu disney 0,0183, pada topik 3 yaitu disney 0,0276, pada topik 4 yaitu video 0,0217, pada topik 5 bagus 0,0089, pada topik 6 netflix 0,0637, pada topik 7 nonton 0,0637, pada topik 8 netflix 0,0119, pada topik 9 disney 0,0306. Nilai tertinggi pada setiap kata terhadap topik dapat mewakili label aspek bahwa kata tersebut menjadi pembahasan yang dominan. Topic modelling latent

dirichlet alloaction (LDA) menghasilkan nilai topik terhadap dokumen. Nilai topik terhadap dokumen dapat menentukan aspek yang mewakili sebuah komentar pengguna platform streaming. Adapun hasil dari topik terhadap dokumen pada Tabel 4.

Tabel 4. Hasil Topik terhadap Dokumen

Kalimat	Topik 1	Topik 2	Topik 3	Topik 4	Topik 5
Nonton sampai final series terakhir	0,0124	0,0124	0,901	0,0124	0,0124
Umroh nonton film seru kayaknya	0,016	0,016	0,016	0,8722	0,016
Nonton film detektif sampai nangis	0,0159	0,0159	0,0159	0,0159	0,8727
Takut telat nonton *live*	0,0222	0,0222	0,0222	0,8221	0,0222
Seru banget nonton walaupun ilegal	0,014	0,0139	0,014	0,014	0,014

Pada proses klasifikasi svm di bagi dua klasifikasi sentimen yaitu positif dan negatif. Ketentuan dari bobot score, apabila kata >0 maka klasifikasi kelas positif dan apabila <0 maka klasifikasi kelas negatif. Data pada klasifikasi ini dibagi menjadi data uji dan data latih bahwa ketentuan jumlah data latih 70 dan data uji 30. Adapun hasil klasifikasi pada Tabel 5.

Tabel 5. Hasil Klasifikasi Support Vector Machine (SVM)

Kalimat	Klasifikasi Sentimen	Prediksi Klasifikasi Sentimen
Nonton sampai final series terakhir	1	Positif
Umroh nonton film seru kayaknya	1	Positif
Nonton film detektif sampai nangis	1	Negatif
Takut telat nonton *live*	-1	Negatif
Seru banget nonton walaupun ilegal	1	Positif

Pada Tabel 5 terdapat beberapa perbedaan pada klasifikasi sentimen dan prediksi klasifikasi sentimen. Pada kalimat ”nonton film detektif sampai nangis” bahwa kata ”nangis” mendapatkan klasifikasi sentimen positif dan pada hasil prediksi klasifikasi sentimen menjadi negatif, karna kata ”nangis” terprediksi sebagai kata negatif. Pada kalimat ”umrok nonton film seru kayaknya” bahwa kata ”seru” mendapatkan sentimen positif sehingga hasil prediksi klasifikasi sentimen menjadi positif.

Pembahasan

Topic Modelling dengan menggunakan metode *Latent Dirichlet Allocation* adalah teknik yang kuat untuk mengklasifikasikan komentar pada layanan streaming platform. *Latent Dirichlet Allocation* adalah model probabilistik yang digunakan untuk mengidentifikasi topik-topik yang tersembunyi dalam koleksi teks. Dalam konteks ini, topik-topik ini dapat mewakili jenis komentar yang berbeda atau topik-topik yang dibahas oleh pengguna dalam komentar mereka. Langkah-langkah umum untuk mengimplementasikan *Latent Dirichlet Allocation* dalam klasifikasi komentar pada layanan streaming platform yakni dalam beberapa langkah. Pertama, yang bersangkutan perlu mengumpulkan sebanyak mungkin data komentar dari layanan streaming platform. Data ini harus mencakup teks komentar beserta metadata tambahan seperti waktu posting, nama pengguna, dll. Proses teks komentar untuk membersihkannya dari karakter khusus, tanda baca, dan kata-kata yang tidak relevan. Pengguna juga perlu melakuakan tokenisasi (mengubah teks menjadi kata-kata individual) dan mengonversi teks menjadi bentuk yang seragam (Firdaus et al., 2020; Nurmawati & Amanda, 2023). Kemudian, langkah selanjutnya adalah membangun model Latent Dirichlet Allocation menggunakan toolkit atau pustaka seperti Gensim atau Scikit-Learn. Penyusun perlu menentukan jumlah topik yang diinginkan sebelumnya atau menggunakan teknik penentuan jumlah topik otomatis seperti Coherence Score atau Perplexity. Langkah selanjutnya yakni melatih Model Latent Dirichlet Allocation. Komentar yang telah dibersihkan untuk melatih model Latent Dirichlet Allocation. Model ini akan menemukan pola topik dalam komentar (Jia, 2019; Xu et al., 2022).

Setelah melatih model, penyusun dapat menggunakannya untuk mengidentifikasi topik dalam komentar baru. Ini akan membantu mengklasifikasikan komentar ke dalam berbagai topik. Kemudian, berdasarkan topik yang diidentifikasi oleh model *Latent Dirichlet Allocation*, penyusun dapat mengklasifikasikan komentar ke dalam kategori-kategori yang sesuai (Sutherland et al., 2020; Zou et al., 2022). Misalnya, jika model Latent Dirichlet Allocation mengidentifikasi bahwa suatu komentar berbicara tentang "musik" atau "kualitas streaming," dapat mengkategorikannya sesuai. Dalam langkah ini penting untuk mengevaluasi kinerja model Anda menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Penyusun juga dapat menggunakan teknik cross-validation untuk mengukur keandalan model. Jika model Latent Dirichlet Allocation tidak memberikan hasil yang memuaskan, dan dapat mencoba mengoptimasi

jumlah topik atau melakukan preprocessing yang lebih canggih (Min et al., 2020; Xue et al., 2020). Terakhir, penyusun dapat mengintegrasikan model klasifikasi ini ke dalam layanan streaming platform Anda untuk mengelompokkan dan menampilkan komentar sesuai dengan topik-topik yang diidentifikasi (Manullang et al., 2023; A. J. Putri et al., 2022). Penerapan *Latent Dirichlet Allocation* dalam klasifikasi komentar pada layanan streaming platform dapat membantu meningkatkan pengalaman pengguna dengan mengorganisir dan mengkategorikan komentar secara otomatis, sehingga memudahkan pengguna untuk menemukan informasi yang mereka cari atau berinteraksi dengan komentar lain yang relevan (Farsiah et al., 2022; Singgalen, 2021).

Tahapan dalam analisis ini dimulai dari pengambilan data dengan framework scrapy menggunakan bahasa pemprograman phyton sebanyak 5.000 data dan dilanjutkan proses preprocessing text yang terdiri dari tahapan punctuatuional removal, lowercasing, tokenizing, stemming, stopword removal, pembobotan kata dan topic modelling latent dirichlet allocation dan klasifikasi support vector machine. Pada topic modelling latent dirichlet allocation ditentukan nilai coherence, kata terhadap topik dan topik terhadap dokumen menggunakan library gensim dengan menetapkan nilai dari parameter seperti jumlah topik dan iterasi. Untuk klasifikasi menggunakan support vectore machine bahwa data sangat mempengaruhi hasil klasifikasi karna apabila dalam data dominan positif dari pada negatif maka hasil prediksi positif akan lebih banyak dari pada negatif dan klasifikasi mempengaruhi jika ada judul film atau nama seorang aktor maka akan terprediksi positif pada label prediksi klasifikasi sentimen. Pada penelitian ini didapatkan bahwa opini dari segi komentar positif lebih dominan dari pada komentar negatif. Pada tahap evaluasi menggunakan mengevaluasi performa dengan hasil akurasi, precision, recall dan f-1 score. Adapun dari hasil data, dapat dipaparkan bahwa akurasi sebesar 88%, precision 88%, recall 88%, dan F1 Score 87%.

Penggunaan Latent Dirichlet Allocation dalam klasifikasi komentar dapat membantu meningkatkan efektivitas dalam mengelompokkan komentar ke dalam kategori atau topik yang sesuai. Ini dapat membuatnya lebih mudah untuk memahami apa yang dibicarakan oleh pengguna dalam komentar mereka. Latent Dirichlet Allocation dapat membantu dalam identifikasi komentar spam atau yang tidak relevan dengan topik diskusi (Çallı & Çallı, 2023; Santoso et al., 2022). Dengan demikian, ini dapat membantu dalam membersihkan komentar yang tidak diinginkan dari platform streaming. Dengan klasifikasi komentar yang lebih baik, pengguna dapat dengan mudah menavigasi dan menemukan konten atau komentar yang paling relevan dengan minat mereka. Ini dapat meningkatkan pengalaman pengguna secara keseluruhan.

Selain klasifikasi topik, *Latent Dirichlet Allocation* juga dapat digunakan untuk menganalisis komentar yang sentimen, seperti apakah komentar tersebut positif, negatif, atau netral (Roiqoh et al., 2023; Widodo, A. O. et al., 2023). Ini dapat membantu platform streaming dalam memahami reaksi pengguna terhadap konten atau acara tertentu. Berdasarkan klasifikasi komentar, platform streaming dapat memberikan rekomendasi yang lebih personal kepada pengguna berdasarkan minat dan preferensi mereka, yang dapat meningkatkan retensi pengguna. Analisis topik komentar dapat memberikan wawasan berharga kepada penyedia layanan streaming tentang apa yang paling diminati oleh pengguna mereka dan area mana yang mungkin perlu ditingkatkan.

4. SIMPULAN

Topic Modelling Latent Dirichlet Allocation (LDA) untuk Klasifikasi Komentar pada Layanan Streaming Platform dengan menggunakan 5.000 data diambil dari sosial media yaitu twitter yang terbagi menjadi komentar positif dan komentar negatif. Hasil ini dipengaruhi dari jumlah komentar positif yang lebih dominan dari pada komentar negatif. Oleh karena itu, perlunya data yang seimbang dalam melakukan klasifikasi komentar platform streaming agar pada saat klasifikasi dapat memberikan hasil prediksi klasifikasi yang baik.

5. DAFTAR PUSTAKA

Anggraeni, S. (2019). Pengaruh Pengetahuan Tentang Dampak Gadget Pada Kesehatan Terhadap Perilaku Penggunaan Gadget Pada Siswa SDN Kebun Bunga 6 Banjarmasin. *Faletehan Health Journal*, *6*(2), 64–68. https://doi.org/10.33746/fhj.v6i2.68.

Annisa, R., & Surjandari, I. (2019). Opinion Mining on Mandalika Hotel Reviews Using Latent Dirichlet Allocation. *Procedia Computer Science*, *161*, 739–746. https://doi.org/10.1016/j.procs.2019.11.178.

Anshari, I. N. (2019). Sirkulasi Film dan Program Televisi di Era Digital: Studi Kasus Praktik Download dan Streaming melalui Situs Bajakan. *Komuniti: Jurnal Komunikasi Dan Teknologi Informasi*, *10*(2), 88–102. https://doi.org/10.23917/komuniti.v10i2.7125.

Bustami, D. K., & Noviaristanti, S. (2022). Service Quality Analysis of Tokopedia Application Using Text Mining Method. *International Journal of Management, Finance and Accounting*, *3*(1), 1–21. https://doi.org/10.33093/ijomfa.2022.3.1.1.

Çallı, L., & Çallı, F. (2023). Understanding airline passengers during covid-19 outbreak to improve service quality: topic modeling approach to complaints with latent dirichlet allocation algorithm. *Transportation Research Record*, *2677*(4), 656–673. https://doi.org/10.1177/03611981221112096.

Farsiah, L., Misbullah, A., & Husaini, H. (2022). Analisis Sentimen Menggunakan Arsitektur Long Short-Term Memory (Lstm) Terhadap Fenomena Citayam Fashion Week. *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, *6*(2), 86–94. https://doi.org/10.22373/cj.v6i2.14687.

Ferdiana, R., Jatmiko, F., Purwanti, D. D., Ayu, A. S. T., & Dicka, W. F. (2019). Dataset Indonesia untuk Analisis Sentimen. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *8*(4), 334–339. https://journal.ugm.ac.id/v3/JNTETI/article/view/2558.

Firdaus, M. R., Rizki, F. M., Gaus, F. M., & Susanto, I. K. (2020). Analisis sentimen dan topic modelling dalam aplikasi ruangguru. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *4*(1), 66–76. https://doi.org/10.30645/j-sakti.v4i1.188.

Gifari, O. I., Adha, M., Hendrawan, I. R., & Durrand, F. F. S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *Journal of Information Technology*, *2*(1), 36–40. https://doi.org/10.46229/jifotech.v2i1.330.

Hafidz, N., & Liliana, D. Y. (2021). Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(2), 213–219. https://doi.org/10.29207/resti.v5i2.2960.

Hermanto, H., Mustopa, A., & Kuntoro, A. Y. (2020). Algoritma klasifikasi naive bayes dan support vector machine dalam layanan komplain mahasiswa. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *5*(2), 211–220. https://doi.org/10.33480/jitk.v5i2.1181.

Hua, T., Lu, C. T., Choo, J., & Reddy, C. K. (2020). Probabilistic topic modeling for comparative analysis of document collections. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *14*(2), 1–27. https://doi.org/10.1145/3369873.

Imron, R. (2018). Hubungan Penggunaan Gadget dengan Perkembangan Sosial dan Emosional Anak Prasekolah di Kabupaten Lampung Selatan. *Jurnal Ilmiah Keperawatan Sai Betik*, *13*(2), 148–154. https://doi.org/10.26630/jkep.v13i2.922.

Indrayuni, E. (2019). Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Khatulistiwa Informatika*, *7*(1), 29–36. https://doi.org/10.31294/jki.v7i1.5740.g3245.

Jia, S. (2019). Toward a better fitness club: Evidence from exerciser online rating and review using latent Dirichlet allocation and support vector machine. *International Journal of Market Research*, *61*(1), 64–76. https://doi.org/10.1177/1470785318770571.

Khomsah, S. (2021). Sentiment Analysis On YouTube Comments Using Word2Vec and Random Forest. *Telematika: Jurnal Informatika Dan Teknologi Informasi*, *18*(1), 61–72. https://doi.org/10.31315/telematika.v18i1.4493.g3346.

Kurnia, P., & Mella, N. F. (2018). Opini Audit Going Concern: Kajian Berdasarkan Kualitas Audit, Kondisi Keuangan, Audit Tenure, Ukuran Perusahaan, Pertumbuhan Perusahaan dan Opini Audit Tahun Sebelumnya pada Perusahaan yang Mengalami Financial Distress pada Perusahaan Manufaktur (Studi pa. *Jurnal Riset Akuntansi Dan Keuangan*, *6*(1), 105–122. https://doi.org/10.17509/jrak.v6i1.8937.

Kusairi, M. M., & Agustian, S. (2022). SVM Method with FastText Representation Feature for Classification of Twitter Sentiments Regarding the Covid-19 Vaccination Program. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(2), 140–150. https://doi.org/10.31849/digitalzone.v13i2.11531.

Manullang, O., Prianto, C., & Harani, N. H. (2023). Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest. *Jurnal Ilmiah Dan Informatika*, *11*(2), 159–169. https://doi.org/10.33884/jif.v11i02.7987.

Min, K. B., Song, S. H., & Min, J. Y. (2020). Topic modeling of social networking service data on occupational accidents in Korea: latent dirichlet allocation analysis. *Journal of Medical Internet Research*, *22*(8), 1–12. https://doi.org/10.2196/19222.

Minerva, L., Sumeisey, V. S., Stefani, S., Wijaya, S., & Lim, C. A. (2020). Pengaruh Kualitas Audit, Debt Ratio, Ukuran Perusahaan dan Audit Lag terhadap Opini Audit Going Concern. *Owner: Riset Dan Jurnal Akuntansi*, *4*(1), 254–266. https://doi.org/10.33395/owner.v4i1.180.

Musliadi, K. H., Zainuddin, H., & Wabula, Y. (2022). Twitter Social Media Conversion Topic Trending Analysis Using Latent Dirichlet Allocation Algorithm. *Journal of Applied Engineering and Technological*

Science (JAETS), *4*(1), 390–399. https://doi.org/10.37385/jaets.v4i1.1143.

Negara, E. S., & Triadi, D. (2021). Topic modeling using latent dirichlet allocation (LDA) on twitter data with Indonesia keyword. *Bulletin of Social Informatics Theory and Application*, *5*(2), 124–132. https://doi.org/10.31763/businta.v5i2.455.

Nurmawati, E., & Amanda, A. (2023). Analisis Sentimen dan Pemodelan Topik Pada Tweet Terkait Data Badan Pusat Statistik. *Jurnal Sistem Informasi Dan Informatika (Simika)*, *6*(2), 165–176. https://doi.org/10.47080/simika.v6i2.2789.

Putranti, N. D., & Winarko, E. (2014). Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *8*(1), 91–100. https://doi.org/10.22146/ijccs.3499.

Putri, A. J., Syafira, A. S., Purbaya, M. E., & Purnomo, D. (2022). Analisis Sentimen E-Commerce Lazada pada Jejaring Sosial Twitter Menggunakan Algoritma Support Vector Machine. *Jurnal TRINISTIK: Jurnal Teknik Industri, Bisnis Digital, Dan Teknik Logistik*, *1*(1), 16–21. https://doi.org/10.20895/trinistik.v1i1.447.

Putri, I. R., & Kusumaningrum, R. (2017). Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia. *Journal of Physics: Conference Series*, *801*(1). https://doi.org/10.1088/1742-6596/801/1/012073.

Ramadhan, N. G., & Ramadhan, T. I. (2022). Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, *7*(1), 39–45. https://doi.org/10.33395/sinkron.v7i1.11204.

Roiqoh, S., Zaman, B., & Kartono, K. (2023). Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Mobile JKN dengan Lexicon Based dan NaÃ¯ve Bayes. *Jurnal Media Informatika Budidarma*, *7*(3), 1582–1592. https://doi.org/10.30865/mib.v7i3.6194.

Royyan, A. R., & Setiawan, E. B. (2022). Feature Expansion Word2Vec for Sentiment Analysis of Public Policy in Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(1), 78–84. https://doi.org/10.29207/resti.v6i1.3525.

Santoso, K. R. A. P., Husna, A., Putri, N. W., & Rakhmawati, N. A. (2022). Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *7*(1), 1–9. https://doi.org/10.14421/jiska.2022.7.1.1-9.

Shidqi, F., & Febrianta, M. Y. (2023). Analisis Kualitas Layanan Internet Service Provider Menggunakan Metode Analisis Sentimen Dan Topic Modelling. *SEIKO : Journal of Management & Business*, *6*(2), 439–450. https://doi.org/10.37531/sejaman.v6i2.5305.

Singgalen, Y. A. (2021). Analisis Sentimen dan Pemodelan Topik dalam Optimalisasi Pemasaran Destinasi Pariwisata Prioritas di Indonesia. *Journal of Information Systems and Informatics*, *3*(3), 459–470. https://doi.org/10.51519/journalisi.v3i3.171.

Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020). Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. *Sustainability*, *12*(5), 1821. https://doi.org/10.3390/su12051821.

Vulić, I., De Smet, W., Tang, J., & Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, *55*(4), 77–84. https://doi.org/10.1016/j.ipm.2014.08.003.

Wibowo, T. O. (2018). Fenomena website streaming film di era media baru: Godaan, perselisihan, dan kritik. *Jurnal Kajian Komunikasi*, *6*(2), 191–203. https://doi.org/10.24198/jkk.v6i2.15623.

Widodo, A. O., Septiadi, F., & Rakhmawati, N. A. (2023). Analisis Tren Konten Pada Vtuber Indonesia Menggunakan Latent Dirichlet Allocation. *Jurnal Informatika Dan Rekayasa Elektronik*, *6*(1), 56–63. https://doi.org/10.36595/jire.v6i1.718.

Xu, H., Zhang, M., Zeng, J., Hao, H., Lin, H. C. K., & Xiao, M. (2022). Use of Latent Dirichlet Allocation and Structural Equation Modeling in Determining the Factors for Continuance Intention of Knowledge Payment Platform. *Sustainability (Switzerland)*, *14*(15), 8992. https://doi.org/10.3390/su14158992.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PLoS ONE*, *15*(9 September), 1–12. https://doi.org/10.1371/journal.pone.0239441.

Zou, Y., Luh, D. B., & Lu, S. (2022). Public perceptions of digital fashion: An analysis of sentiment and Latent Dirichlet Allocation topic modeling. *Frontiers in Psychology*, *13*(December), 1–21. https://doi.org/10.3389/fpsyg.2022.986838.