Jurnal SimanteC
Vol. 12, No. 2 Juni 2024
P-ISSN : 2088-2130
E-ISSN : 2502-4884

PENERAPAN *NAÏVE BAYES* DAN *LATENT DIRICHLET ALLOCATION* (LDA) UNTUK ANALISIS SENTIMEN DAN PEMODELAN TOPIK PADA PROYEK KERETA CEPAT JAKARTA-BANDUNG

APPLICATION OF NAÏVE BAYES AND LATENT DIRICHLET ALLOCATION (LDA) FOR SENTIMENT ANALYSIS AND TOPIC MODELING ON THE JAKARTA-BANDUNG HIGH-SPEED RAIL PROJECT

[1*]Doni Abdul Fatah, [2]Fajrul Ihsan Kamil, [3]Budi Soesilo, [4]Mulaab

[1,2,3]Prodi Sistem Informasi, Jurusan Teknik Informatika, Universitas Trunojoyo Madura
[4]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Trunojoyo Madura
Jl. Raya Telang, PO BOX 2 Kamal, Bangkalan
e-mail: [1]doni.fatah@.trunojoyo.ac.id[*],[2]200411100172@student.trunojoyo.ac.id,
[3]budi.soesilo@trunojoyo.ac.id ,[4]mulaab@trunojoyo.ac.id

Abstrak

Proyek kereta cepat Jakarta-Bandung merupakan salah satu proyek besar yang saat ini sedang dibuat di Indonesia. Proyek kereta cepat Jakarta – Bandung menjadi ramai dibicarakan di media sosial Twitter. Karena dalam pembangunannya terdapat beberapa masalah, seperti banjir yang terjadi di Bekasi dan menyebabkan kemacetan dan mengganggu kelancaran logistik. Beberapa opini masyarakat dapat berupa sentimen positif dan negatif terhadap Pembangunan Kereta Cepat ini. Untuk mengetahui opini masyarakat tersebut maka perlu dilakukan analisis sentimen dan pemodelan topik menggunakan metode *Naïve Bayes* dan *Latent Dirichlet Allocation*. Hasil pengujian menunjukkan bahwa pada analisis sentimen setelah dilakukan perhitungan menggunakan metode *Naïve Bayes* diperoleh nilai akurasi sebesar 66%. Sedangkan pada pemodelan topik menggunakan metode *Latent Dirichlet Allocation* diuji menggunakan nilai koherensi terbaik diperoleh nilai sebesar 0,472 pada 9 topik.

Kata kunci: Analisis Sentimen, Kereta Cepat, *Latent Dirichlet Allocation*, *Naïve Bayes* , Pemodelan Topik.

Abstract

The Jakarta – Bandung high-speed train project is one of the major projects currently being built in Indonesia. The Jakarta – Bandung high-speed train project has become widely discussed on Twitter social media. Because during its construction there were several problems, such as flooding that occurred in Bekasi which caused traffic jams and disrupted the smooth running of logistics. Some public opinions can include positive and negative sentiments regarding the construction of this high-speed train. To find out public opinion, it is necessary to carry out sentiment analysis and topic modeling using the Naïve Bayes and Latent Dirichlet Allocation methods. The test results show that in sentiment analysis after carrying out calculations using the Naïve Bayes method, an accuracy value of 66% was obtained. Meanwhile, topic modeling using the Latent Dirichlet Allocation method was tested using the best coherence value, obtaining a value of 0.472 on 9 topics.

Keywords: *Sentiment Analysis, Fast Train, Latent Dirichlet Allocation, Naïve Bayes , Topic Modeling.*

1 PENDAHULUAN

Proyek kereta cepat Jakarta – Bandung telah dimulai sejak tanggal 21 januari 2016 dengan dilakukannya *groundbreaking* oleh Jokowi di Perkebunan Mandalawangi Maswati, Cikalong Wetan, Bandung Barat, Jawa Barat. Menurut Rini mantan menteri BUMN, keuntungan dibangunnya kereta cepat Jakarta – Bandung diantaranya akan meningkatkan perekonomian, mengangkat sektor pariwisata, dan membuka lapangan perkerjaan yang baru.Permasalahan yang lain dalam proyek kereta cepat ini kurang memperhatikan kelancaran akses keluar – masuk jalan tol, pembiaran penumpukan material yang mengganggu fungsi *drainase*, pembangunan pilar LRT tanpa izin, sampai persoalan keselamatan dan Kesehatan kerja [1]. Oleh karena itu, diperlukan analisis sentimen untuk mengetahui bagaimana sentimen yang ada pada media sosial mengenai pembangunan proyek kereta cepat Jakarta – Bandung.

Analisis sentimen adalah proses mengidentifikasi dan mengelompokan opini yang masih berbentuk teks ke dalam sentimen positif atau negatif [2]. Sebuah sistem analisis sentimen dibangun menggunakan algoritma klasifikasi *Naïve Bayes*. Fitur utama dari algoritma *Naïve Bayes* adalah asumsi yang sangat kuat dari setiap kondisi atau kejadian . Kelebihan dari *Naïve Bayes* adalah proses klasifikasi data dapat disesuaikan dengan sifat dan kebutuhan setiap orang. Dengan adanya sistem analisis sentimen, diharapkan dapat membantu atau meningkatkan algoritma dengan melihat tingkat keakuratan [3].

Penelitian sebelumnya sudah dilakukan penelitain tentang analisis sentiment Masyarakat terhadap Pembangunan Kereta Cepat dengan cara pengklasifikasian review tersebut ke dalam class positif dan negatif. Teknik klasifikasi yang digunakan untuk klasifikasi data adalah *K-Nearest Neighbors* (KNN). Kemudian dalam menentukan evaluasinya peneliti menggunakan *Accuracy* dan AUC (*Area Under Curve*). Hasil penelitian menunjukkan akurasi *K-Nearest Neighbors* yaitu 82.70 %[1].

Selanjutnya juga pada penelitian sebelumnya tentang analisis sentiment terhadap Pembangunan Kereta Cepat menggunakan metode *Naïve Bayes*. Hasil dari penelitian tersebut menunjukkan setelah melalui pemrosesan dengan hasil sentimen negatif sebanyak 673, hasil sentimen positif sebanyak 668, dan hasil sentimen netral sebanyak 665 hasil *accuracy* 71%, *precision* 73%, *recall* 89%[4].

Berdasarkan uraian tersebut, berbeda dengan penelitian-penelitian yang telah dilakukan sebelumnya, dalam penelitian ini tidak hanya menampilkan hasil sentimen saja akan tetapi dapat memvisualisaikan hasil topik sehingga memudahkan pengguna mengetahui sebaran kata dan frasa di setiap topik. Berdasarkan penelitian dengan metode LDA yang telah diuraikan. Metode LDA dapat menyelesaikan permasalahan yang berkaitan tentang pemodelan topik. Oleh karena itu pada penelitian ini peneliti menggunakan Algoritma *Naïve Bayes* dan Pemodelan Topik LDA untuk menganalia sentiment Masyarakat terhadap Kereta Cepat Jakarta-Bandung pada Twitter.

2 TINJAUAN PUSTAKA

Preprocessing data merupakan tahapan penting dalam persiapan data untuk analisis. Proses ini dilakukan untuk mengubah data mentah menjadi data yang siap digunakan dalam analisis, sehingga dapat memperoleh hasil analisis yang akurat dan relevan diantaranya [5].

Penelitian tentang *Sentiment Analysis of Madura Tourism Opinion Using Support Vector Machine* (SVM), dimana pada artikel tersebut membahas tentang analisis sentimen terhadap opini wisata Madura menggunakan metode *Support Vector Machine* (SVM). Dalam uji coba dengan validasi silang K = 5 *fold*, diperoleh hasil sentimen positif sebanyak 192 *tweet* dengan akurasi 92,592% menggunakan *Confusion Matrix*. Hasil ini menunjukkan bahwa SVM efektif dalam mengklasifikasikan opini publik terkait pariwisata Madura menjadi sentimen positif, yang dapat digunakan sebagai dasar untuk memberikan rekomendasi perbaikan dalam meningkatkan kualitas pariwisata Madura berdasarkan pandangan Masyarakat [6].

Penelitian lainnya tentang *Topic Modelling Latent Dirichlet Allocation* untuk Klasifikasi Komentar pada Layanan *Streaming Platform*, dimana pada artikel tersebut membahas penggunaan metode Topic Modelling *Latent Dirichlet Allocation* (LDA) dan *Support Vector Machine* (SVM) dalam mengklasifikasikan komentar pada layanan streaming platform di media sosial Twitter. Penelitian ini dilakukan dengan mengambil 5.000 data komentar dari Twitter yang

terbagi menjadi komentar positif dan komentar negatif. Tujuan penelitian ini adalah untuk membandingkan antara komentar positif dan komentar negatif yang diberikan oleh pengguna *platform streaming* di Twitter. Metode LDA digunakan untuk mempresentasikan topik dan dokumen, sementara SVM digunakan untuk klasifikasi. Hasil penelitian menunjukkan bahwa metode LDA dan SVM menghasilkan lebih banyak komentar positif daripada komentar negatif. Evaluasi kinerja menunjukkan bahwa penelitian ini mencapai nilai akurasi sebesar 0,88, *recall* sebesar 0,88, F1-*score* sebesar 0,87, dan presisi sebesar 0,88. Hal ini menunjukkan bahwa metode ini efektif dalam mengklasifikasikan komentar pada layanan *streaming platform*. Penelitian ini juga menyoroti pentingnya menjaga keseimbangan data dalam melakukan klasifikasi komentar untuk mendapatkan hasil prediksi yang lebih akurat [7].

Penelitian lainnya membahas penggunaan metode *Latent Dirichlet Allocation* (LDA) dalam analisis topik dan analisis sentimen terkait dengan sirkuit Mandalika di media sosial, dengan fokus pada *platform* Twitter. Penelitian ini dilakukan untuk memanfaatkan informasi yang tersebar luas di media sosial terkait dengan sirkuit Mandalika yang belum dimanfaatkan dengan baik oleh pihak-pihak terkait. Metode SVM digunakan untuk analisis sentimen, sementara metode LDA digunakan untuk pemodelan topik. Hasil penelitian menunjukkan bahwa mayoritas pengguna Twitter memberikan respons positif terhadap sirkuit Mandalika berdasarkan analisis sentimen terhadap 500 *tweet* yang dikumpulkan. Algoritma SVM mampu mengklasifikasikan sentimen dengan baik, dengan akurasi 87%, *presisi* 77%, *recall* 84,81%, dan spesifisitas 98,52%. Hasil penelitian ini diharapkan dapat digunakan sebagai data pendukung bagi pemerintah dan sektor swasta dalam pengambilan keputusan terkait pengembangan sektor pariwisata yang relevan dengan kebutuhan wisatawan [8].

Penelitian lainhya tentang membahas implementasi metode Klasifikasi *Naïve Bayes* dan Pemodelan Topik dengan *Latent Dirichlet Allocation* (LDA) untuk data ulasan video game lokal pada platform Steam. Penelitian ini bertujuan untuk meningkatkan prediksi sentimen yang akurat dalam ulasan online serta mengidentifikasi topik-topik dominan dalam ulasan video game. Hasil penelitian menunjukkan bahwa setelah penyeimbangan data, tingkat akurasi klasifikasi *Naïve Bayes* mencapai 81%. Selain itu, penelitian ini berhasil mengidentifikasi 5 topik yang dominan dalam ulasan yang direkomendasikan dan 3 topik dalam ulasan yang tidak direkomendasikan, termasuk fitur permainan, *puzzle, visual art, soundtrack, update patch*, karakter, gameplay, dan masalah bug serta performa video game [9].

Penelitian lainnya tentang membahas pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA) pada ulasan masyarakat tentang aplikasi PeduliLindungi melalui Play Store. Penelitian ini dilakukan untuk membagi data komentar berdasarkan topik yang dibahas dalam ulasan tersebut. Hasil penelitian menunjukkan bahwa menggunakan LDA dapat mengidentifikasi lima topik utama yang meliputi kendala pendaftaran, sertifikat vaksin, kesalahan tanggal lahir, kendala membuka aplikasi, dan keluhan pengguna aplikasi [10].

Penelitian lainnya membahas tentang analisis sentimen terkait wabah COVID-19 di Twitter dengan menggunakan metode *Latent Dirichlet Allocation* (LDA) dan *Naïve Bayes* Classifier (NBC). Tujuan dari penelitian ini adalah untuk membentuk pemodelan topik terkait wabah COVID-19 di Twitter dan menganalisis sentimen positif dan negatif dalam setiap topik. Dengan menggabungkan kedua metode tersebut, penelitian ini berhasil mencapai akurasi yang baik sebesar 89%. Hasil penelitian menunjukkan bahwa topik yang paling banyak dibahas adalah vaksinasi booster, dan topik tersebut lebih cenderung memiliki sentimen negatif daripada sentimen positif [11].

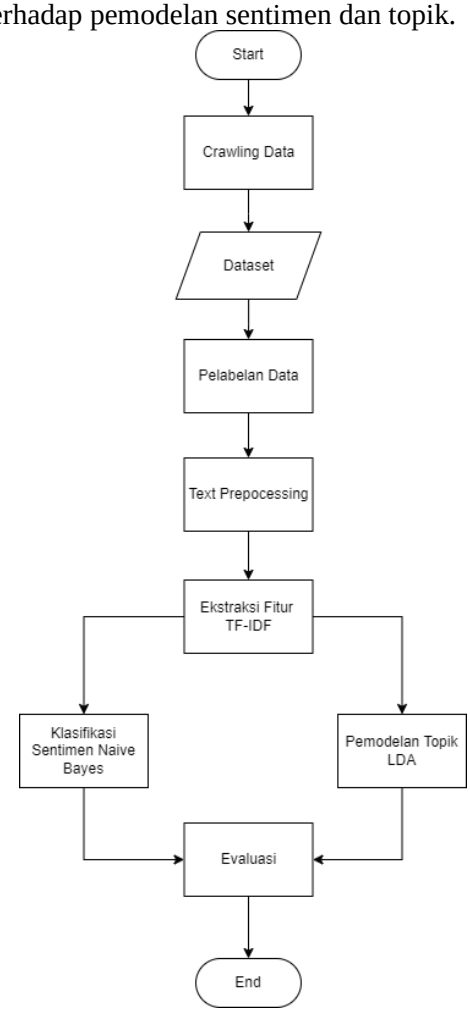
Penelitian lainnya membahas tentang analisis sentimen dan pemodelan topik pariwisata di Pulau Lombok menggunakan algoritma *Naïve Bayes* dan *Latent Dirichlet Allocation* (LDA). Tujuan penelitian ini adalah untuk mengklasifikasikan opini wisatawan menjadi kelas positif dan negatif serta memodelkan topik yang dibicarakan. Hasil penelitian menunjukkan bahwa metode *Naïve Bayes* memiliki akurasi 92% dengan presisi 100% dan LDA menghasilkan nilai koherensi tertinggi pada topik ke-8 untuk kelas positif dan topik ke-12 untuk kelas negatif. Penggunaan algoritma *Naïve Bayes* dan LDA dinilai efektif untuk analisis sentimen dan pemodelan topik pariwisata di Lombok [12].

Penelitian serupa lainnya ini membahas tentang analisis sentimen dan pemodelan topik terkait pandemi Covid-19 di media sosial Twitter. Penelitian ini menggunakan metode *Naïve*

Bayes Classifier (NBC) untuk analisis sentimen dan metode *Latent Dirichlet Allocation* (LDA) untuk pemodelan topik. Hasil penelitian menunjukkan bahwa dengan menggunakan NBC, dapat diklasifikasikan sentimen positif, negatif, dan netral dari data tweet terkait Covid-19. Selain itu, dengan menggunakan LDA, dapat diidentifikasi topik-topik yang sering dibicarakan oleh pengguna Twitter terkait Covid-19. Penelitian ini menggunakan 5000 data *tweet* sebagai dataset dan melakukan *preprocessing* teks sebelum pemodelan topik. Hasil pemodelan topik divisualisasikan menggunakan word cloud untuk memperlihatkan kata-kata yang banyak muncul dalam tweet terkait Covid-19 [13].

3 METODE PENELITIAN

Tahapan-tahapan yang dilakukan pada penelitian ini ditunjukan oleh Diagram Penelitian pada Gambar 1. Pada penelitian ini memiliki 6 tahapan yaitu pertama melakukan pengumpulan data yaitu feedback pengguna aplikasi, *Preprocessing, labeling* sentimen, pemodelan analisis sentimen menggunakan *Naïve Bayes*, pemodelan topik menggunakan LDA dan tahap terakhir yaitu melakukan evaluasi terhadap pemodelan sentimen dan topik.



Gambar 1. Diagram Penelitian

a. Pengumpulan Data

Langkah ini dilakukan dengan mengumpulkan data untuk keperluan analisis. Data dikumpulkan dari media sosial Twitter. Data yang diambil merupakan tweet (*tweet, reply, retweet*) berbahasa indonesia yang mengandung kata kunci yang telah ditentukan yaitu "#keretacepat", "Pembangunan kereta cepat" dan "#kcjb". Pengambilan data dilakukan dengan menggunakan *crawling* data dari twitter dengan python yang dilakukan dalam rentang waktu s/d 15 September 2023. Data yang diambil sebanyak 1000 *Tweet*. Data yang dihasilkan dari proses *crawling* memiliki format CSV. Data tersebut memiliki beberapa atribut seperti *username*, *tweet* dan atribut lainnya.

b. Preprocessing

Proses ini dilakukan untuk mengubah data mentah menjadi data yang siap digunakan dalam analisis, sehingga dapat memperoleh hasil analisis yang akurat dan relevan diantaranya[5]:

1. *Case folding*
   Proses mengubah text *tweet* menjadi text yang sama untuk melakukan pencarian atau analisis seperti mengkonversi text menjadi huruf kecil.
2. *Cleansing*
   Proses pembersihan teks dan menghilangkan komponen yang tidak diperlukan seperti url, *mention*, *emoticon*, dan hastag.
3. *Tokenizing*
   Proses pemecahan text menjadi unit yang lebih kecil dengan memisahkan tanda koma atau spasi.
4. *Slangword*
   Proses perubahan pada kata-kata atau frasa yang bersifat informal, tidak resmi, dan seringkali singkat
5. *Stopword Removal*
   Proses menghapus kata-kata umum yang dianggap memiliki makna yang sedikit dari sebuah teks.
6. *Stemming*
   Proses menyederhanakan kata-kata dalam suatu teks menjadi kata dasar.
   Berikut merupakan contoh dari data yang sudah dilakukan *preprocessing pada tabel 1 berikut:*

Tabel 1. Hasil *Prepocessing*

Proses	Sebelum	Sesudah
Case Folding	Sejumlah fasilitas mewah terdapat di dalam Kereta Cepat Jakarta-Bandung salah satunya adalah toilet yang ramah untuk pengguna berkebutuhan khusus.	sejumlah fasilitas mewah terdapat di dalam kereta cepat jakarta-bandung salah satunya adalah toilet yang ramah untuk pengguna berkebutuhan khusus.
Cleansing	sejumlah fasilitas mewah terdapat di dalam kereta cepat jakarta-bandung salah satunya adalah toilet yang ramah untuk pengguna berkebutuhan khusus.	sejumlah fasilitas mewah terdapat di dalam kereta cepat jakarta bandung salah satunya adalah toilet yang ramah untuk pengguna berkebutuhan khusus
Tokenizing	sejumlah fasilitas mewah terdapat di dalam kereta cepat jakarta bandung salah satunya adalah toilet yang ramah untuk pengguna berkebutuhan khusus	['sejumlah', 'fasilitas', 'mewah', 'terdapat', 'di', 'dalam', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satunya', 'adalah', 'toilet', 'yang', 'ramah', 'untuk', 'pengguna', 'berkebutuhan', 'khusus']
Slangword	['sejumlah', 'fasilitas', 'mewah', 'terdapat', 'di', 'dalam', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satunya', 'adalah', 'toilet', 'yang', 'ramah', 'untuk', 'pengguna', 'berkebutuhan', 'khusus']	['fasilitas', 'mewah', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satunya', 'toilet', 'ramah', 'pengguna', 'berkebutuhan', 'khusus']
Stemming	['fasilitas', 'mewah', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satunya', 'toilet', 'ramah', 'pengguna', 'berkebutuhan', 'khusus']	['fasilitas', 'mewah', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satu', 'toilet', 'ramah', guna, butuh, 'khusus']
Stopword Removal	['fasilitas', 'mewah', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satu', 'toilet', 'ramah', guna, butuh, 'khusus']	['fasilitas', 'mewah', 'kereta', 'cepat', 'jakarta', 'bandung', 'salah', 'satu', 'toilet', 'ramah', 'butuh', 'khusus']

c. Klasifikasi *Naïve Bayes*

Naïve Bayes Classifier adalah metode klasifikasi statistik yang dapat memprediksi probabilitas keanggotaan kelas, seperti probabilitas bahwa sampel yang diberikan termasuk dalam kelas tertentu. Metode ini yang akan digunakan pada penelitian ini untuk klasifikasi data yang diambil dari Twitter dan data tersebut akan diklasifikasikan menjadi kelas positif dan kelas negatif. *Naïve Bayes Classifier* memiliki bentuk umum seperti berikut[14].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (1)$$

Keterangan dengan X merupakan data dengan kelas tidak dikenal, H adalah hipotesis data X adalah kelas khusus, P(H|X) adalah probabilitas hipotesis H didasarkan pada kondisi X, P(H)

adalah probabilitas hipotesis H, P(X|H) adalah probabilitas hipotesis X didasarkan pada kondisi H, P(X) adalah probabilitas X

d. Pemodelan Topik dengan LDA

Latent Direchlet Analysis (LDA) adalah model yang digunakan dalam menemukan topik *laten* (tersembunyi) dibalik dokumen-dokumen. Hal tersebut terjadi karena model LDA menginterpretasikan dokumen sebagai campuran yang random[15]. Model LDA melakukan interpretasi topik-topik dalam dokumen dalam model statistik. Proses LDA bersifat imaginary random proses pada model bahwa tiap dokumen berasan dari tema atau topik tertentu dan setiap topiknya memiliki struktur distribusi kata-kata atau *term*.

e. Evaluasi

Pengujian untuk algoritma *Naïve Bayes* menggunakan *Confusion Matrix* berdasarkan pengukuran akurasi, *precision*, dan *recall* seperti pada persamaan 2,3,4 berikut :

a. Akurasi (%) = $\frac{TP + TN}{TP + FP + FN + TN}$ (2)

b. Recall (%) = $\frac{TP}{TP + FN}$ (3)

c. Presisi (%) = $\frac{TP}{TP + FP}$ (4)

dengan TP adalah *True Positive*, TN adalah *True Negative*, FP adalah *False Positive* dan FN adalah *False Negative*[16].

Pengujian untuk algoritma *Latent Dirichlet Allocation* menggunakan pengukuran nilai koherensi terbaik berdasarkan jumlah topik yang akan dibentuk.

4 HASIL DAN PEMBAHASAN

a. Hasil Pengujian *Naïve Bayes*

Setelah melakukan pelabelan data dan ekstraksi fitur menggunakan *TF-IDF*, tahap selanjutnya penelitian ini yaitu melakukan pengujian analisis sentimen menggunakan metode *Naïve Bayes*. Dalam pengujian metode ini dilakukan beberapa skenario uji coba dengan menggunakan parameter alpha sebesar (1,0.1,0.01,0.001), *random state* dan data *test* sebesar (20%, 30%, 40%) untuk melihat berapa akurasi yang terbaik dari beberapa parameter yang ada.

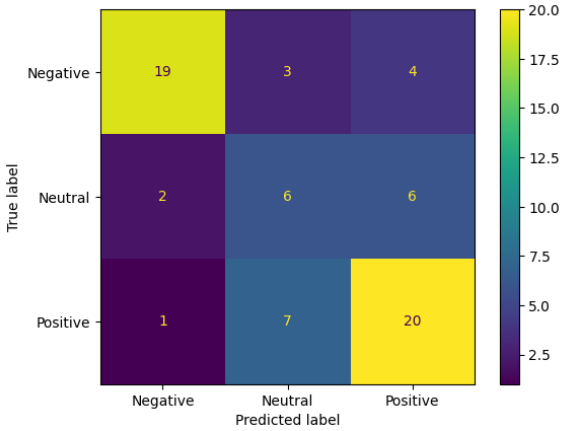
Tabel 2. Percobaan parameter

Percobaan	Random State	Alpha	Test size	Akurasi
1	1	1	0.2	0.6304
2	1	0.1	0.2	0.6304
3	1	0.01	0.2	0.5869
4	1	0.001	0.2	0.5652
5	1	1	0.3	0.5882
6	1	0.1	0.3	0.6617
7	1	0.01	0.3	0.6176
8	1	0.001	0.3	0.6029
9	1	1	0.4	0.6483
10	1	0.1	0.4	0.6483
11	1	0.01	0.4	0.6153
12	1	0.001	0.4	0.6153

Berdasarkan 12 skenario dan beberapa parameter yang diuji pada tabel ditemukan hasil akurasi terbesar yaitu pada nilai *random state* = 1, *alpha* = 0.1, *test_size* = 0.3 dengan hasil akurasi 0.661765. Grafik hasil evaluasi dapat dilihat pada gambar 2 berikut

Gambar 2. Grafik Perbandingan Akurasi

Berdasarkan hasil skenario uji coba pada tabel 2, dapat disimpulkan bahwa akurasi terbaik diperoleh oleh model yang menggunakan 30% data testing dan 70% data training dengan nilai akurasi sebesar 0,66 atau 66%. Dari permodelan tersebut, dapat diketahui jumlah data yang diprediksi salah dengan menggunakan *confusion matrix*. Hasil *confusion matrix* yang diperoleh dari permodelan tersebut ditunjukan pada Gambar 3 di bawah ini:

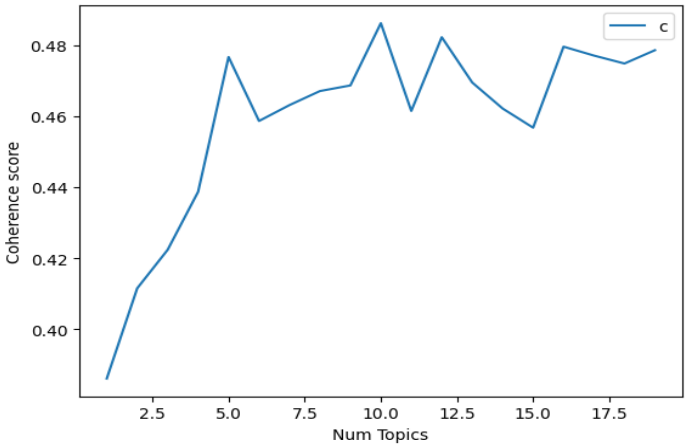


Gambar 3. *Confusion Matrix*

Hasil *confusion matrix*, dapat disimpulkan bahwa dari data testing yang telah dilakukan prediksi, terdapat 23 data yang memiliki prediksi tidak sesuai dengan data aslinya dan yang sesuai sebanyak 45 data. Hal ini terjadi karena banyak data positif yang diprediksi netral oleh sistem karena data positif dan netral hamper sama. Hal ini menimbulkan akurasi yang didapat menjadi kurang maksimal yaitu 66%.

b. Hasil Pengujian LDA

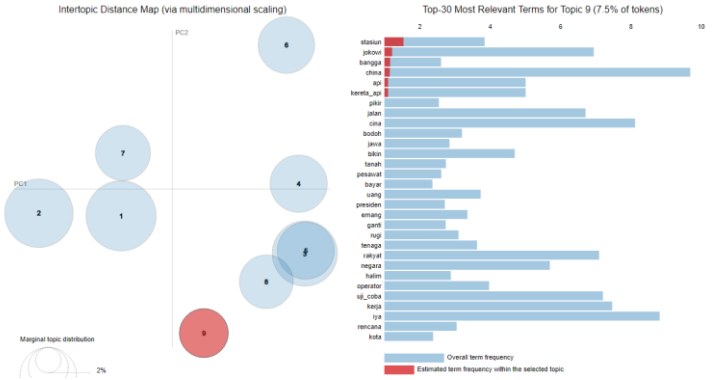
Setelah hasil sentiment ditemukan, kemudian data akan dieksplorasi lebih lanjut untuk dilakukan pemodelan topik menggunakan metode *Latent Dirichlet Alocation* (LDA). Tahap awal dari pemodelan topik adalah membuat *dictionary* dan *corpus* dari *dataset* yang sama. Selanjutnya dilakukan proses pembentukan model LDA menggunakan bantuan *library gensim*.



Gambar 4. Grafik Nilai Koherensi

Berdasarkan nilai koherensi pada gambar 4, telah dilakukan perhitungan menggunakan metode *Latent Dirichlet Allocation* dengan pembentukan topik 1 sampai dengan 20. Diperoleh nilai koherensi terbaik untuk data kereta cepat pada kelompok topik ke 9 dengan nilai 0,472. Semakin tinggi nilai koherensi suatu topik maka semakin mudah topik tersebut diinterpretasi-kan maknanya berdasarkan kumpulan kata yang menyusunnya, dengan kata lain semakin sering kata-kata dalam topik tersebut muncul secara bersamaan maka nilai koherensi dari topik tersebut semakin tinggi.

Setelah pada gambar 4 disimpulkan bahwa hasil perhitungan pada metode LDA diperoleh nilai koherensi terbaik 0,472. Dari hasil nilai koherensi tersebut digunakan visualisasi pemodelan topik menggunakan bentuk pyLDAvis ditunjukkan pada gambar 5.



Gambar 5. Visualisasi Kata per Topik

Berdasarkan 9 topik yang dibentuk dapat diinterprentasikan pada tabel yaitu topik “1” mengandung informasi masih digunakannya tenaga cina saat uji coba kereta cepat. Topik “2” mengandung informasi tiket uji coba gratis yang habis menandakan Masyarakat antusias dengan adanya kereta cepat ini. Topik “3” mengandung informasi kcic atau kereta cepat ini bermanfaat bagi Masyarakat. Topik “4” mengandung informasi tentang pembangunannya yang menggunakan apbn ataukah utang dari china. Topik “5” mengandung informasi keterkaitan dengan ikn dan juga operator yang masih menggunakan orang dari china. Topik “6” mengandung informasi saat Pembangunan kereta cepat sering mengakibatkan macet pada saat pulang kerja. Topik “7” mengandung informasi tentang tiket dan hutang terhadap cina. Topik “8” mengandung informasi jalan macet ke tol halim saat pulang kerja efek dari kereta cepat. Topik “9” mengandung informasi rakyat yang bangga terhadap Jokowi dengan berhasilnya kereta cepat.

5 KESIMPULAN

Penelitian ini berhasil melakukan analisis sentimen menggunakan metode *Naïve Bayes* dan pemodelan topik menggunakan *Latent Dirichlet Allocation* untuk topik Kereta Cepat Jakarta-Bandung. Hasil pengujian menunjukkan bahwa pada analisis sentimen setelah dilakukan perhitungan menggunakan metode *Naïve Bayes* diperoleh nilai akurasi sebesar 66%. Sedangkan

pada pemodelan topik menggunakan metode *Latent Dirichlet Allocation* diuji menggunakan nilai koherensi terbaik diperoleh nilai sebesar 0,472 pada 9 topik. Lalu untuk topik yang sering diperbincangkan oleh masyarakat pada Pembangunan kereta cepat ini yaitu tentang pengoperasian yang masih menggunakan tenaga dari cina dan masyarakat yang antusias dengan dibangunnya kereta cepat.

Sebagai saran untuk pengembangan penelitian berikutnya diharapkan untuk menambahkan dataset dan juga melihat sebaran kata-kata yang terlihat memiliki makna ambigu sehingga bisa menambah akurasi dari metode klasifikasi ataupun pemodelan topik nantinya. Selain itu perlu dilakukannya pemodelan topik terhadap masing-masing kelas agar distribusi topik yang menyebar menyesuaikan dengan kata yang ada pada masing-masing kelas

DAFTAR PUSTAKA

[1] C. Pricylia, A. Mulya, P. Nugraha, and I. Santoso, "Analisis Sentimen Masyarakat Terhadap Pembangunan Kereta Cepat Jakarta-Bandung Menggunakan Algoritma *K-Nearest Neighbors* (Knn)," *J. IKRAITH-INFORMATIKA*, vol. 7, no. 2, pp. 139–143, 2023, [Online]. Available: www.gataframework.com/textmining.Peneliti

[2] D. Rusdiaman and D. Rosiyadi, "Analisa Sentimen Terhadap Tokoh Publik Menggunakan Metode *Naïve Bayes* Classifier dan *Support Vector Machine*," *J. Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 2502–7131, 2019, doi: https://doi.org/10.24114/cess.v4i2.13796.

[3] D. A. Fatah, E. M. S. Rochman, W. Setiawan, A. R. Aulia, F. I. Kamil, and A. Su'ud, "Sentiment Analysis of Public Opinion Towards Tourism in Bangkalan Regency Using *Naïve Bayes* Method," *E3S Web Conf.*, vol. 499, p. 01016, 2024, doi: 10.1051/e3sconf/202449901016.

[4] J. Florensius Sianipar, Y. R. Ramadhan, and I. Jaelani, "Analisis Sentimen Pembangunan Kereta Cepat Jakarta-Bandung di Media Sosial Twitter Menggunakan Metode *Naïve Bayes* ," *Media Online)*, vol. 4, no. 1, pp. 360–367, 2023, doi: 10.30865/klik.v4i1.1033.

[5] Y. Khoiruddin, A. Fauzi, and A. M. Siregar, "Analisis Sentimen Gojek Indonesia Pada Twitter Menggunakan Algoritme *Naïve Bayes* Dan *Support Vector Machine*," *J. Ilm. Komput.*, vol. 19, pp. 391–400, 2023.

[6] Doni Abdul Fatah, Eka Mala Sari Rochman, Fajrul Ihsan Kamil, and Ahmad Su'ud, "Sentiment Analysis of Madura Tourism Opinion Using *Support Vector Machine* (SVM)," *Tech. Rom. J. Appl. Sci. Technol.*, vol. 16, pp. 243–249, Oct. 2023, doi: 10.47577/technium.v16i.9988.

[7] N. Royani, C. E. Widodo, and B. Warsito, "Topic Modelling *Latent Dirichlet Allocation* untuk Klasifikasi Komentar pada Layanan Streaming Platform," *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 3, pp. 815–822, 2024, doi: 10.23887/jstundiksha.v12i3.68492.

[8] R. Hammad, V. C. Hardita, and A. Z. Amrullah, "Topic modeling and sentiment analysis about Mandalika on social media using the *Latent Dirichlet Allocation* method," *MATRIX J. Manaj. Teknol. dan Inform.*, vol. 12, no. 3, pp. 109–116, 2022, doi: 10.31940/matrix.v12i3.109-116.

[9] Y. S. Wardhana and A. K. Ayundyah Kesumawati, "Implementasi Klasifikasi *Naïve Bayes* dan Pemodelan Topik dengan *Latent Dirichlet Allocation* untuk Data Ulasan Video Game Lokal Pada Platform Steam," *Emerg. Stat. Data Sci. J.*, vol. 1, no. 3, pp. 345–353, 2023, doi: 10.20885/esds.vol1.iss.3.art41.

[10] J. Akbar, T. A. M., Y. Tolla, A. E. Ahmad, A. Yaqin, and E. Utami, "Pemodelan Topik Menggunakan *Latent Dirichlet Allocation* pada Ulasan Aplikasi PeduliLindungi," *InComTech J. Telekomun. dan Komput.*, vol. 13, no. 1, p. 40, 2023, doi: 10.22441/incomtech.v13i1.15572.

[11] P. P. Aziztiya, M. Habibi, and N. I. Kusumaningtyas, "Analisis Sentimen Berdasarkan Topik Terkait Wabah Covid-19 di Twitter Menggunakan *Latent Dirichlet Allocation* (LDA) dan *Naïve Bayes* Classifier (NBC)," *Teknomatika J. Inform. dan Komput.*, vol. 15, no. 2, pp. 76–85, 2022, doi: 10.30989/teknomatika.v15i2.1098.

[12] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Sentiment Analysis and Lombok Tourism Topic Modeling Using *Naïve Bayes* and *Latent Dirichlet Allocation* Algorithms,"

J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 1, pp. 123–131, 2021.

[13] Y. A. Singgalen, "Analisis Sentimen dan Pemodelan Topik dalam Optimalisasi Pemasaran Destinasi Pariwisata Prioritas di Indonesia," *J. Inf. Syst. Informatics*, vol. 3, no. 3, pp. 459–470, 2021, doi: 10.51519/journalisi.v3i3.171.

[14] K. M. Leung and others, "*Naïve Bayes* ian classifier," *Polytech. Univ. Dep. Comput. Sci. Risk Eng.*, vol. 2007, pp. 123–156, 2007.

[15] J. C. Campbell, A. Hindle, and E. Stroulia, "Chapter 6 - *Latent Dirichlet Allocation*: Extracting Topics from Software Engineering Data," in *The Art and Science of Analyzing Software Data*, C. Bird, T. Menzies, and T. Zimmermann, Eds., Boston: Morgan Kaufmann, 2015, pp. 139–159. doi: https://doi.org/10.1016/B978-0-12-411519-4.00006-9.

[16] K. M. Ting, "Confusion Matrix," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2017, p. 260. doi: 10.1007/978-1-4899-7687-1_50.