Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 1 (2021) 123 - 131 ISSN Media Elektronik: 2580-0760

Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma *Naive Bayes* dan *Latent Dirichlet Allocation*

Ni Luh Putu Merawati[1], Ahmad Zuli Amrullah[2], Ismarmiaty[3]
[1,2]Teknologi Informasi, Teknik dan Desain, Universitas Bumigora
[3]Sistem Informasi, Teknik dan Desain, Universitas Bumigora
[1]putu.mera@universitasbumigora.ac.id, [2]zuli@universitasbumigora.ac.id, [3]ismarmiaty@universitasbumigora.ac.id

Abstract

Lombok Island is one of the favorite tourist destinations. Various topics and comments about Lombok tourism experience through social media accounts are difficult to manually identify public sentiments and topics. The opinion expressed by tourists through social media is interesting for further research. This study aims to classify tourists' opinions into two classes, positive and negative, and topics modelling by using the Naive Bayes method and modeling the topic by using Latent Dirichlet Allocation (LDA). The stages of this research include data collection, data cleaning, data transformation, data classification. The results performance testing of the classification model using Naive Bayes method is shown with an accuracy value of 92%, precision of 100%, recall of 84% and specificity of 100%. The results of modeling topics using LDA in each positive and negative class from the coherence value shows the highest value for the positive class was obtained on the 8th topic with a value of 0.613 and for the negative class on the 12th topic with a value of 0.528. The use of the Naive Bayes and LDA algorithms is considered effective for analyzing the sentiment and topic modelling for Lombok tourism.

Keywords: sentiment analysis, probabilistic computing, machine learning, tourism

Abstrak

Pulau Lombok menjadi salah satu tujuan wisata favorit di Indonesia. Beragam topik maupun komentar tentang Lombok disampaikan wisatawan melalui akun pribadi media sosialnya sehingga sangat sulit untuk melakukan identifikasi sentimen publik maupun topik pembicaraan secara manual. Opini yang disampaikan wisatawan melalui media sosial khususnya *twitter* tentang pariwisata Lombok, menarik untuk diteliti lebih lanjut. Penelitian ini bertujuan untuk melakukan klasifikasi opini-opini wisatawan menjadi dua kelas yaitu positif dan negatif serta melakukan pemodelan topik pada kedua kelas tersebut. Pemodelan topik bertujuan untuk mengetahui topik yang sering dibicarakan pada masing-masing kelas. Tahapan dari penelitian ini meliputi pengumpulan data, pembersihan data, transformasi data, klasifikasi data dengan metode *Naive Bayes* dan penggunaan metode *Latent Dirichlet Allocation* (LDA) untuk pemodelan topik. Hasil pengujian kinerja model menggunakan algoritma *Naive Bayes* ditunjukkan dengan nilai akurasi, presisi, recall dan spesifisitas masing-masing sebesar 92%, 100%, 83,84% dan 100%. Hasil pemodelan topik dengan metode LDA pada masing-masing kelas positif dan negatif dapat dilihat dari nilai koherensi yaitu semakin tinggi nilai koherensi suatu topik maka semakin mudah topik tersebut diinterpretasikan oleh manusia. Nilai koherensi tertinggi untuk kelas positif diperoleh pada topik ke 8 dengan nilai sebesar 0,613 dan untuk kelas negatif pada topik ke 12 dengan nilai sebesar 0,528. Penggunaan algoritma *Naive Bayes* dan LDA dinilai efektif untuk analisis sentimen serta pemodelan topik untuk pariwisata Lombok.

Kata kunci: analisis sentimen , komputasi probabilistik, pembelajaran mesin, pariwisata

1. Pendahuluan

Pulau Lombok merupakan salah satu bagian wilayah administratif daerah tingkat 1 Provinsi Nusa Tenggara Barat (NTB) sehingga pulau Lombok menjadi pilihan utama untuk pengembangan pariwisata NTB dibandingkan dengan pulau lainnya yang masuk wilayah NTB. Dinamika ini sejalan dengan upaya pemerintah daerah untuk menjadikan sektor pariwisata sebagai salah satu program unggulan provinsi Nusa Tenggara Barat (NTB). Sektor pariwisata diyakini mampu menjadi sumber pertumbuhan ekonomi baru di Lombok. Selain itu Lombok ditargetkan menjadi salah satu pintu gerbang pariwisata nasional untuk wilayah Nusa Tenggara [1] . Para wisatawan yang datang berkunjung ke suatu daerah

Diterima Redaksi: 24-10-2020 | Selesai Revisi: 18-02-2021 | Diterbitkan Online: 20-02-2021

wisata mampu memberikan andil yang besar dalam pertumbuhan ekonomi daerah sehingga pengembangan kepariwisataan Lombok sangat perlu dilakukan sebagai salah satu upaya untuk mendukung perkembangan kepariwisataan nasional [2].

Secara umum perkembangan pariwisata Lombok masih cukup lambat jika dibandingkan dengan daerah lain di Indonesia seperti Bali, Yogyakarta, Bandung, Malang, dan lain-lain. Faktor-faktor yang mempengaruhi lambatnya perkembangan wisata Lombok antara lain ketersediaan fasilitas pendukung kepariwisataan masih sangat terbatas mulai dari akses hingga infrastruktur [2] serta masih minimnya pemanfaatan teknologi informasi dalam pengelolaan kepariwisataan daerah.

Pertumbuhan media sosial yang cukup pesat saat ini tidak lepas dari pengaruh perkembangan teknologi informasi. Media sosial adalah media informasi paling diminati masyarakat saat ini serta merupakan media untuk bersosialisasi satu sama lain secara online. Keberadaan media sosial memberikan banyak manfaat, salah satunya adalah masyarakat menggunakan media sosial sebagai tempat untuk menyampaikan opini atau pendapat, kritik, saran secara bebas [3]. Twitter merupakan salah satu media sosial yang populer dikalangan masyarakat Indonesia karena mampu memberikan informasi secara cepat dan *real time*. Twitter menjadi wadah penyampaian opini masyarakat dalam bentuk saran, kritikan, maupun pendapat. Twitter dapat menjadi sumber informasi yang tepat untuk menggali opini masyarakat melalui cuitan-cuitan yang dilontarkan terhadap suatu berita atau kejadian dengan topik tertentu. Twitter dapat digunakan untuk mengetahui isi pikiran atau sentimen pemilik akun [4].

Sentimen yang disampaikan masyarakat melalui twitter mengandung informasi yang sangat berharga untuk dianalisis [5]. Penggalian informasi dari sekumpulan *tweet* sulit dilakukan secara manual karena jumlah *tweet* yang di post perhari sangat banyak dengan topik ulasan beragam sehingga membutuhkan model analisis data yang cepat dan tepat. Model analisis data seperti analisis sentimen dan pemodelan topik dapat digunakan untuk membantu menemukan informasi-informasi tersembunyi dari sekumpulan *tweet* [6]. Analisis sentimen merupakan bagian dari text mining yang bertujuan untuk mengklasifikasi dokumen teks berupa opini sehingga menghasilkan suatu informasi sentimen yang dapat bermakna positif maupun negatif [7]. Metode analisis sentimen dapat digunakan untuk menganalisa pendapat maupun emosi seseorang dalam meyikapi sesuatu yang menyangkut topik tertentu[8] misalnya sentimen masyarakat terhadap pariwisata Lombok sehingga informasi mengenai kepuasan wisatawan dapat diketahui. Hasil analisis sentimen tersebut dapat dimanfaatkan oleh pemerintah daerah, pelaku wisata maupun stakeholder sebagai pendukung pengambilan keputusan dalam pengembangan dan pengelolaan pariwisata Lombok kedepannya.

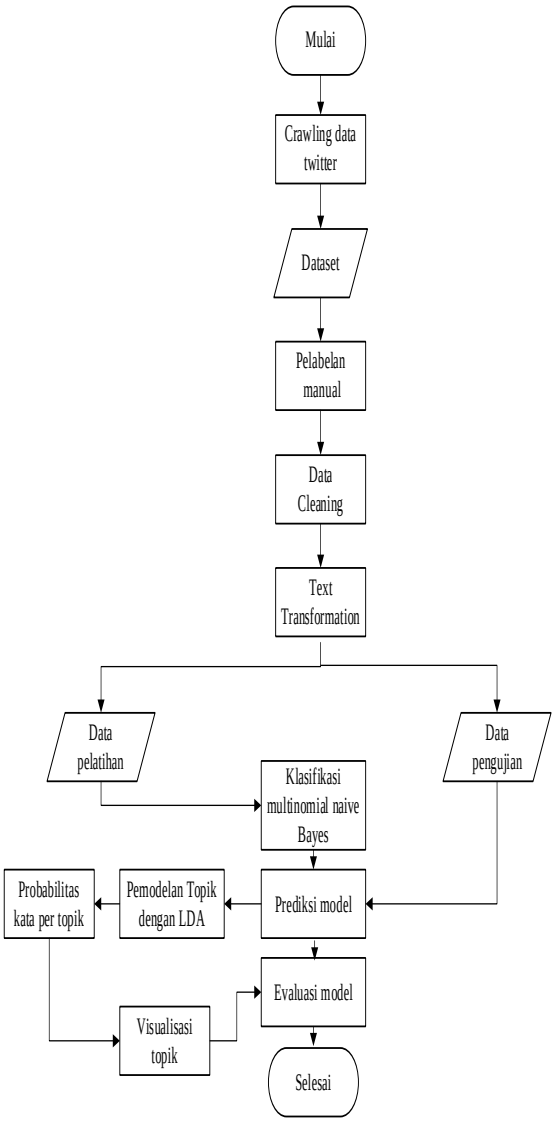
Berdasarkan penelitian yang telah dilakukan, banyak peneliti melakukan eksplorasi data teks melalui klasifikasi untuk menghasilkan informasi yang penting seperti penelitian yang dilakukan [9] tentang analisis sentimen untuk penilaian tempat tujuan wisata kota Tegal menggunakan metode *naive bayes* dan *decision tree*. Penelitian ini bertujuan untuk mendapatkan model terbaik untuk dimplementasikan pada sistem. Hasil penelitian menunjukkan bahwa nilai akurasi *naive bayes* sebesar 77,50% lebih baik dibandingkan dengan *decision tree* dengan nilai akurasi sebesar 60,83%. Penelitian lainnya dilakukan oleh [10] bertujuan untuk menganalisis ulasan masyarakat tentang pariwisata kota Malang melalui analisis sentimen. Pada penelitian ini menggunakan metode *naive bayes* untuk klasifikasi teks dan metode *query expansion ranking* untuk mengurangi jumlah fitur pada proses klasifikasi. Hasil penelitian menghasilkan akurasi sebesar 86,6%. Kemudian [11] melakukan analisis sentimen pemeringkatan popularitas tujuan wisata menggunakan algoritma *naive bayes* dengan nilai akurasi sebesar 82,67%. Selanjutnya [12] melakukan analisis sentimen twitter tentang pariwisata Lombok. Penelitian ini menggunakan metode *naive bayes* serta *mutual information* untuk seleksi fitur dengan nilai akurasi sebesar 97,9%.

Fokus utama dari beberapa penelitian sebelumnya yang telah dijabarkan adalah melakukan analisis sentimen pada topik pariwisata. Oleh karena itu pada penelitian ini akan melakukan analisis sentimen pada topik pariwisata Lombok menggunakan metode *naive bayes* karena metode ini mempunyai beberapa kelebihan yaitu menurut [4] metode *naive bayes* adalah metode klasifikasi text yang mempunyai kecepatan pemrosesan dan akurasi yang cukup tinggi apabila diterapkan kasus yang jumlah datanya banyak, besar dan beragam serta dalam [13] mengemukakan bahwa *naive bayes* adalah algoritma yang sederhana dan bagus untuk klasifikasi teks. Selain itu penelitian ini juga berfokus untuk mengidentifikasi topik yang paling banyak dibicarakan pada masing-masing kelas yaitu positif dan negatif menggunakan metode *Latent Dirichlet Allocation* (LDA). LDA digunakan untuk menentukan beberapa topik yang muncul dari masing-masing opini pada setiap kelas. Kelebihan metode LDA adalah dapat mengekstrak topik secara akurat pada kumpulan data yang cukup besar [6]. Penelitian ini terbagi menjadi dua tujuan yaitu pertama melakukan analisis sentimen pada *twitter* dengan topik pariwisata Lombok menjadi dua kelas yaitu positif dan negatif dengan metode *naive bayes* kemudian mengukur hasil kinerja model berdasarkan empat kriteria yaitu akurasi, presisi, *recall* dan spesifitas. Kedua melakukan pemodelan topik pada masing-masing kelas positif dan negatif untuk mengidentifikasi topik utama yang sering dibahas pada kedua kelas tersebut menggunakan *Latent Dirichlet Allocation* (LDA) serta mengukur hasil kinerja model LDA berdasarkan nilai koherensi. Hasil penelitian ini digunakan sebagai data pendukung bagi para pelaku wisata di Pulau Lombok

DOI: https://doi.org/10.29207/resti.v5i1.2587

seperti pemerintah daerah khususnya dinas pariwisata maupun sektor swasta dalam pembuatan kebijakan maupun pengambilan keputusan yang berkaitan dengan pengembangan sektor pariwisata yang relevan dengan kebutuhan wisatawan seperti kebijakan untuk promosi, perbaikan fasilitas tempat wisata, penambahan fasilitas tempat wisata, penambahan fasilitas umum dan sebagainya.

2. Metode Penelitian



.

Gambar 1. Diagram alir penelitian

Tahapan penelitian ini secara umum ditunjukkan pada Gambar 1. Berdasarkan Gambar 1, penelitian ini mempunyai 7 tahapan penting yaitu pengumpulan data *tweet*, pelabelan secara manual, data *cleaning*, *text transformation*, klasifikasi *tweet* menggunakan *multinomial naive bayes*, pemodelan topik menggunakan LDA dan yang terakhir adalah tahap pengujian menggunakan metode *confusion matrix*

2.1. Pengambilan Data

Tahap pertama pada penelitian ini adalah pengumpulan data *tweet* dengan topik pariwisata melalui API twitter dengan menggunakan *hastag* seperti #gilitrawangan, #wonderfulllombok, #pantaikutamandalika, #lombok #beautifullombok, #lombokindah, #senggigibeach #kekmandalika, #lombokaman, dan lain-lain. Data *tweet* yang diambil merupakan rentang data 5 tahun terakhir yaitu dari tahun 2014 sampai 2019. Contoh data tweet yang berhasil dikumpulkan diperlihatkan pada tabel 1. Selanjutnya seluruh data yang terkumpul akan melalui proses *preprocessing* bertujuan untuk pembersihan data.

Tabel 1. Contoh data *tweet*

Kelas	Tweet
Negatif	Tak sebersih dulu lagi...#sampah #pantai senggigi
Positif	Dinikmati aja,,,keindahannya... #Lovepantai

2.2. *Data Cleaning*

Tahap kedua pada penelitian ini adalah proses pembersihan data bertujuan agar dataset tidak mengandung *noise* yang dapat mempengaruhi hasil klasifikasi. Contoh hasil data c*leaning* ditunjukkan tabel 2, dimana tahapan data *cleaning* meliputi hapus *hastag* pada *tweet*, hapus URL, hapus *mention*, hapus karakter *tweet*, hapus kata-kata yang diulang, hapus angka dan tanda baca.

Tabel 2. Hasil data *cleaning*

Langkah	Sebelum	Sesudah
Hapus hastag	Memandang keindahan laut dan pantai dari atas bukit... #lombok #lombokisland #bukitmerese #meresehill #visitlombok #visitindonesia #wonderfulindonesia @Bukit Merese Lombok https://www.instagram.com/p/BoBPFALlsmG/?utm_source=ig_twitter_share&amp	Memandang keindahan laut dan pantai dari atas bukit... @Bukit Merese Lombok https://www.instagram.com/p/BoBPFALlsmG/?utm_source=ig_twitter_share&amp
Hapus URL	Memandang keindahan laut dan pantai dari atas bukit.... @Bukit Merese Lombok https://www.instagram.com/p/BoBPFALlsmG/?utm_source=ig_twitter_share&amp	Memandang keindahan laut dan pantai dari atas bukit.... @Bukit Merese Lombok
Hapus mention	Memandang keindahan laut dan pantai dari atas bukit.... @Bukit Merese Lombok	Memandang keindahan laut dan pantai dari atas bukit...
Hapus karakter	Memandang keindahan laut dan pantai dari atas bukit...	Memandang keindahan laut dan pantai dari atas bukit...

DOI: https://doi.org/10.29207/resti.v5i1.2587

Hapus kata yang berulang	Memandang keindahan laut dan pantai dari atas bukit...	Memandang keindahan laut dan pantai dari atas bukit...
Hapus angka	Memandang keindahan laut dan pantai dari atas bukit...	Memandang keindahan laut dan pantai dari atas bukit...
Hapus tanda baca	Memandang keindahan laut dan pantai dari atas bukit	Memandang keindahan laut dan pantai dari atas bukit

2.3. *Text Transformation*

Tahap ketiga pada penelitian ini adalah proses *text transformation.* Hasil *text transformation* ditunjukkan pada tabel 3. Tahapan *text transformation* yang dilakukan pada penelitian ini terdiri dari 6 proses yaitu *case folding*, tokenisasi, *stopword removal*, *lematisasi*, pembobotan dan ekstraksi fitur.

a. *Case folding* yaitu proses mengubah semua karakter huruf menjadi huruf kecil.
b. Tokenisasi yaitu proses pemotongan dokumen inputan berdasarkan tiap kata yang menyusunnya.
c. *Stopword removal* yaitu proses pemilihan kata-kata penting yang mempunyai arti dan tidak, sehingga kata yang tidak mempunyai arti akan dibuang, contohnya "ini", "itu", "yang", "di" dan lain-lain.
d. Lematisasi adalah proses normalisasi kata untuk menemukan bentuk dasar dari kata tersebut berdasarkan bentuk lemmanya.
e. Ekstraksi fitur adalah proses mengubah kata menjadi fitur menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini akan menghitung kemunculan kata pada setiap dokumen kemudian memberi bobot pada kata tersebut. Proses pemberian bobot kata dipengaruhi oleh tiga faktor utama, yaitu *Term Frequency* (TF), *Inverse Document Frequency* (IDF) dan *Document Length* [14]. TF dihitung berdasarkan jumlah kemunculan setiap kata pada sebuah dokumen sedangkan IDF dihitung berdasarkan jumlah kemunculan kata pada seluruh dokumen. Kemudian dilakukan proses normalisasi lalu nilai TF dibandingkan terhadap nilai IDF [15]. Setelah semua kata teridentifikasi langkah selanjutnya adalah membuat tabel untuk menampung kumpulan kata tersebut dimana setiap kata akan memiliki fitur kolom yang disebut dengan *text vectorization*. Hasil vektor ini digunakan sebagai fitur untuk pelatihan dan pengujian klasifikasi.

Tabel 3. Hasil *text transformation*

Langkah	Sebelum	Sesudah
Case folding	Memandang keindahan laut dan pantai dari atas bukit	memandang keindahan laut dan pantai dari atas bukit
Tokenisasi	memandang keindahan laut dan pantai dari atas bukit	'memandang', 'keindahan', 'laut', 'dan', 'pantai', 'dari', 'atas', 'bukit'
Stopword removal	memandang keindahan laut dan pantai dari atas bukit	'memandang', 'keindahan', 'pantai', 'bukit'
Lematisasi	memandang keindahan pantai bukit	'pandang', 'indah', 'pantai', 'bukit'

2.4. Klasifikasi *Naive Bayes*

Tahap keempat dari penelitian ini adalah klasifikasi teks menggunakan metode *multinomial naive bayes*. Tujuan dari klasifikasi yaitu untuk mengetahui kelas sebuah *tweet* apakah masuk kelas positif atau kelas negatif. Langkah penting dari proses klasifikasi adalah representasi sebuah kalimat atau dokumen ke dalam bentuk angka agar dapat dipahami oleh komputer, dimana pada penelitian ini menggunakan metode *bag of word*.

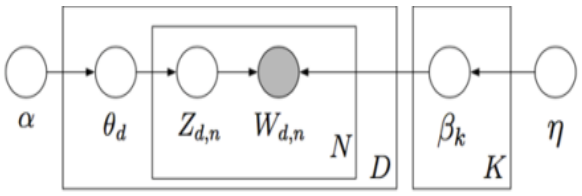
Perhitungan *multinomial naive bayes* dilakukan dengan cara menghitung jumah kemunculan kata dalam dokumen dengan rumus diperlihatkan pada persamaan 1.

$$P(X|c) = \log\frac{N_c}{N} + \sum_{i=1}^{n} log\,\frac{t_i+\alpha}{\sum_{i=1}^{n} t_i+\alpha} \quad (1)$$

dengan P(X|c) adalah probabilitas dokumen X pada kelas c, Nc adalah total dokumen pada kelas c, N adalah total dokumen, t_i adalah bobot *term* t, $\sum_{i=1}^{n} t$ adalah total bobot *term* pada kelas c, α adalah nilai parameter *smoothing* [16].

2.5. Pemodelan Topik dengan LDA

Tahap kelima pada penelitian ini adalah pemodelan topik menggunakan metode LDA. Masing-masing sentimen positif dan negatif akan diproses menggunakan metode LDA untuk mengetahui interpretasi topik utama yang sering dibahas oleh wisatawan pada kedua kelas sentimen tersebut. Prinsip dasar dari LDA adalah setiap dokumen direpresentasikan sebagai campuran topik-topik yang tersembunyi dan belum diketahui, dimana setiap topik terdiri dari distribusi banyak kata [8]. Blei merepresentasikan metode LDA sebagai sebuah *probabilistic model* seperti ditunjukkan pada Gambar 2 [17].



Gambar 2. Representasi Model LDA

Berdasarkan ilustrasi pada gambar 2 maka LDA dapat digambarkan secara khusus penggunaan notasi-notasi matematika dapat dilihat pada persamaan 2 [17].

$$p(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n-1}^{N} p(Z_{d,n}|\theta_d) p(W_{d,n}|\beta_{1:k}, Z_{d,n}) \right) \quad (2)$$

DOI: https://doi.org/10.29207/resti.v5i1.2587

D adalah kumpulan dokumen, K adalah kumpulan topik, N adalah jumlah kata dalam dokumen (N_d), $W_{d,n}$ adalah kata ke-n pada dokumen d, $Z_{d,n}$ adalah topik ke-n pada dokumen d, θ_d adalah jumlah topik per dokumen yang terindentifikasi, β_k adalah distribusi topik pada *vocabulary* serta α, ɳ adalah parameter *dirichlet* [17].

Secara umum cara kerja metode LDA pada penelitian ini adalah [18]:

a. Membuat kamus dan korpus dari kumpulan sentimen positif dan negatif
b. Melakukan inisialisasi parameter yaitu jumlah dokumen, jumlah topik, jumlah iterasi, random state, nilai alpha, nilai beta dan lain-lain.
c. Menentukan kata-kata untuk topik tertentu berdasarkan distribusi *dirichlet*.
d. Menampilkan probabilitas kata per topik
e. Mengulang alur b sampai d untuk semua kata dalam korpus

2.6. Pengujian

Tahap keenam dari penelitian ini adalah pengujian kinerja dari metode. Pengujian algoritma *naive bayes* dilakukan menggunakan metode *confusion matrix* berdasarkan pengukuran akurasi, *precision*, *recall* dan *specificity* seperti yang diperlihatkan persamaan 3, 4, 5 dan 6.

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

dengan TP adalah *True Positive*, TN adalah *True Negative*, FP adalah *False Positive* dan FN adalah *False Negative* [19].

Pengujian LDA dilakukan dengan melihat nilai *koherence* topik yaitu seberapa mudah topik tersebut diinterpretasikan menggunakan rumus pada persamaan 7.

$$Coherence\ (V) = \sum_{(v_i, v_j) \in V} Score\ (v_i, v_j, \in) \quad (7)$$

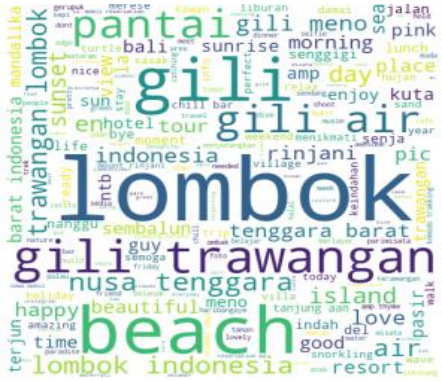
dengan V adalah kumpulan kata untuk menjelaskan sebuah topik dan ϵ adalah faktor *smoothing* untuk mengembalikan nilai ke bilangan ril [20].

3. Hasil dan Pembahasan

3.1. Dataset

Data yang berhasil dikumpulkan sebanyak 12.971 *tweet* kemudian dilakukan pelabelan manual oleh 3 orang anotator untuk dikelompokan menjadi 2 kelas yaitu kelas positif dan kelas negatif. Proses pemilihan label *tweet* dilakukan dengan memilih label mayoritas yang diberikan oleh 3 orang anotator. Tahapan selanjutnya yaitu proses pembersihan data dari unsur *hastag*, url, tanda baca, kata yang berulang, *mention*, karakter dan kelebihan spasi, kemudian dilanjutkan dengan proses *text transformation* yang meliputi proses *case folding* tokenisasi, *stopword removal*, lematisasi, pembobotan dan ekstraksi fitur. Hasil *preprocessing* data menunjukkan bahwa tidak semua *tweet* dapat digunakan sebagai dataset, dimana jumlah data yang memenuhi kriteria sebanyak 9.496 data dengan pembagian 8.996 sentimen positif dan 500 sentimen negatif. Jumlah data yang tidak sama (*imbalance* data) untuk setiap kelasnya dimana jumlah data kelas positif lebih besar daripada kelas negatif. *Imbalance* data yang terjadi akan berpengaruh terhadap hasil prediksi dari model yang dibuat karena data *training* lebih dominan pada satu kelas saja. Untuk mengatasi permasalahan tersebut diterapkan teknik *undersampling* atau pengurangan jumlah data pada kelas mayoritas agar jumlah data menjadi sama dengan kelas minoritas. Pembagian dataset setelah diterapkan teknik *undersampling* adalah 500 *tweet* kelas positif dan 500 *tweet* kelas negatif.

Profile hasil klasifikasi untuk kelas positif dan negatif dapat dilihat pada visualisasi *word cloud* seperti yang ditunjukkan Gambar 3 dan 4. Pada kelas positif sering muncul kata Lombok, gili, pantai, hotel, sunrise, good, love, beautiful sedangkan pada kelas negatif kata yang sering muncul adalah sampah, gempa bumi, toilet, rinjani, gunung dan pantai. Dari kata-kata yang sering muncul pada word cloud positif dapat disimpulkan bahwa pantai dan gili trawangan menjadi tujuan wisata favorit wisatawan sedangkan untuk kelas negatif opini wisatawan lebih banyak mengarah ke masalah sampah.



Gambar 3. Tampilan *word cloud* kelas positif

Gambar 4. Tampilan *word cloud* kelas negatif

DOI: https://doi.org/10.29207/resti.v5i1.2587

3.2. Klasifikasi *Naive Bayes*

Proses klasifikasi diawali dengan membagi dataset menjadi 2 bagian yaitu 80% data *training* dengan jumlah 800 *twee*t dan 20% data *testing* dengan jumlah 200 *tweet*. Data *training* digunakan untuk membangun model menggunakan algoritma *multinomial naive bayes.* Data *testing* digunakan untuk mengevaluasi perfoma model yang diperoleh melalui proses *training*. Proses *training* dimulai dengan menghitung nilai *prior* dari setiap kelas (positif dan negatif) setelah itu dilanjutkan dengan menghitung peluang *term* ke n pada sebuah dokumen. Kemudian dilanjutkan menghitung peluang sebuah dokumen masuk ke dalam suatu kelas dan tahapan terakhir adalah menentukan kelas dokumen dengan memilih nilai probabilitas tertinggi.

Pengujian model menggunakan metode *confusion matrix*, sebanyak 200 *tweet* akan diujicobakan untuk mengetahui perfoma model berdasarkan akurasi, presisi, recall dan spesifisitas. Metode *confusion matrix* dipilih karena mampu memberikan perbandingan hasil klasifikasi model dengan klasifikasi sebenarnya menggunakan 4 kombinasi nilai prediksi yaitu *True Positive* (TP) menunjukkan , *False Positive* (FP), *False Negative* (FN) dan *True Negative* (TN). Selain itu metode *confusion matrix* cocok untuk mengukur kinerja dari model klasifikasi yang menghasilkan dua *output* kelas seperti pada penelitian ini yaitu kelas positif dan negatif.

Berdasarkan hasil pengujian yang diperlihatkan pada tabel 4 menggunakan 200 *tweet* diperoleh 99 *tweet* masuk kelas positif dan 101 *tweet* masuk kelas negatif. Sedangkan hasil prediksi model menunjukkan bahwa prediksi untuk kelas positif dan benar adalah 83 *tweet*, prediksi untuk kelas negatif dan benar adalah 101 *tweet*, prediksi kelas positif dan salah adalah 0 serta prediksi kelas negatif dan salah adalah 16 *tweet*.

Tabel 4. Hasil matriks konfusi metode *Naive Bayes*

Prediksi	Aktual	
	Positif	Negatif
Positif	83	0
Negatif	16	101

Hasil matrik konfusi diatas dijadikan acuan untuk perhitungan kinerja model dengan parameter akurasi, presisi, recall dan spesifisitas. Hasil perhitungan menunjukkan nilai akurasi model sebesar 92%, presisi 100%, recall 83,84% dan Spesifisitas 100%. Sehingga secara keseluruhan dapat disimpulkan bahwa metode *naive bayes* memiliki kinerja yang baik untuk mengklasifikasi sentimen dengan topik pariwisata Lombok.

3.3. Pemodelan topik dengan LDA

Data hasil sentimen akan dieksplorasi lebih lanjut menggunakan metode *Laten Dirichlet Alocation* (LDA). Pemodelan topik dilakukan untuk masing-masing kelas yaitu kelas positif dan kelas negatif. Tujuan pemodelan topik pada penelitian ini untuk menggali informasi dari kumpulan opini wisatawan yang pernah berkunjung ke Lombok. Informasi-informasi tersebut akan diinterpretasikan dalam bentuk kumpulan topik utama pada kelas positif dan negatif. Tahap awal dari pemodelan topik adalah membuat *dictionary* dan *corpus* untuk kelas positif serta negatif. Data yang digunakan merupakan dataset yang sama untuk klasifikasi *naive bayes* sebanyak 500 *tweet* kelas positif dan 500 *tweet* kelas negatif.

Tahapan penting pada proses pemodelan topik adalah pembentukan kamus dan korpus untuk data kelas positif dan kelas negatif. Selanjutnya dilakukan proses pembentukan model LDA menggunakan bantuan *library* gensim. Langkah awal adalah menetapkan nilai parameter yang akan digunakan seperti nomor topik, *random state* = 100, *update every* = 1, *chunksize* = 100, *passes* = 10, alpha = auto. Penelitian ini akan mengambil rentang 1 sampai 20 topik untuk diuji guna mencari kelompok topik terbaik. Topik yang dihasilkan dari pemodelan topik belum tentu mudah untuk diinterpretasikan. Oleh karena itu dilakukan perhitungan koherensi topik untuk membedakan mana topik yang baik dan buruk. Indikator suatu topik dikatakan baik berdasarkan tingkat kemudahan kata-kata di dalam topik untuk ditafsirkan secara semantik, sedangkan suatu topik dikatakan buruk jika kata-kata di dalam topik sulit untuk dimaknai.

Hasil perhitungan nilai koherensi untuk topik pada kelas positif dan negatif ditunjukkan pada tabel 5. Berdasarkan hasil perhitungan pada tabel 5, nilai koherensi terbaik untuk kelas positif diperoleh pada kelompok ke 8 dengan nilai 0,613 sedangkan untuk kelas negatif diperoleh pada kelompok ke 12 dengan nilai 0,528. Semakin tinggi nilai koherensi suatu topik maka semakin mudah topik tersebut diinterpretasikan 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. Kelompok kata-kata penyusun topik untuk kelas positif dan negatif dapat dilihat pada tabel 6 dan 7.

Berdasarkan tabel 6 terdapat 8 topik yang dapat diinterpretasikan untuk kelas positif. Hasil interpretasi topik digunakan untuk melihat tren komentar wisatawan terhadap pariwisata Lombok baik itu dilihat dari sisi keindahan alamnya, kenyamanan, harapan wisatawan, makanan, dan lain-lain. Hasil interpretasi topik untuk kelas positif adalah Topik "0" mengandung informasi wisatawan banyak yang datang ke pantai pink pada saat *weekend*. Topik "1" mengandung informasi air terjun di Sembalun sangat indah. Topik "2" mengandung informasi selain Pulau Bali, Nusa Tenggara Barat mempunyai pantai yang indah. Topik "3" mengandung informasi keindahan sunset di gili meno, air dan trawangan. Topik "4" mengandung informasi keindahan

DOI: https://doi.org/10.29207/resti.v5i1.2587

pantai gili nanggu dapat dinikmati dari hotel sambil makan siang. Topik "5" mengandung informasi *sunrise* di gili trawangan. Topik "6" mengandung informasi pantai senggigi dan tanjung aan merupakan pantai dengan pemandangan mengagumkan di NTB. Topik "7" mengandung informasi jasa sewa mobil untuk tour wisata.

Tabel 5. Nilai koherensi topik

Topik	Koherensi	
	Positif	Negatif
1	0,582	0,409
2	0,584	0,439
3	0,596	0,459
4	0,604	0,478
5	0,606	0,527
6	0,616	0,472
7	0,612	0,510
8	0,613	0,512
9	0,581	0,489
10	0,556	0,510
11	0,552	0,521
12	0,553	0,528
13	0,500	0,469
14	0,510	0,465
15	0,486	0,457
16	0,475	0,443
17	0,462	0,421
18	0,451	0,423
19	0,444	0,430
20	0,469	0,424

Tabel 6. Kumpulan topik untuk kelas positif

Topik 0	Topik 1	Topik 2	Topik 3
beach	lombok	nusa	gili
place	indonesia	tenggara	trawangan
pink	air	barat	meno
time	pic	morning	en
selamat	island	indonesia	lombok
nice	sembalun	pasir	air
weekend	terjun	good	sea
pagi	senja	pantai	beautiful
moment	kuta	bali	sunset
dive	mandalika	selfi	gilis

Topik 4	Topik 5	Topik 6	Topik 7
lombok	day	lombok	rinjani
pantai	sunrise	beach	tour
love	trawangan	happy	sewa
view	beach	ntb	mobil
hotel	enjoy	aan	kawan
keindahan	happy	resort	info
nanggu	semoga	senggigi	bar
halal	semaian	indonesia	chill
indah	minggu	amazing	wisata
lunch	angkatan	village	life

Berdasarkan tabel 7 terdapat 12 topik yang dapat diinterpretasikan untuk kelas negatif yaitu topik "0" mengandung informasi banyaknya sampah plastik dan botol di pantai. Topik "1" mengandung informasi disepanjang jalan pantai pink jarang terpasang lampu jalan. Topik "2" mengandung informasi banyak ditemukan sampah di gunung rinjani. Topik "3" mengandung informasi fasilitas toilet umum masih terbatas. Topik "4" mengandung informasi menolak pembangunan kereta gantung gunung rinjani. Topik "5" mengandung informasi *coral bleaching* terumbu karang. Topik "6" mengandung informasi banyak kotoran kuda di jalan. Topik "7" mengandung informasi musholla kurang bersih. Topik "8" mengandung informasi trotoar untuk pejalan kaki masih jarang. Topik "9" mengandung informasi ditemukan banyak coretan pada batu di gunung rinjani dan air terjun benang kelambu. Topik "10" mengandung informasi penurunan jumlah wisatawan akibat gempa. Topik "11" mengandung informasi penutupan gunung rinjani akibat kebakaran hutan.

Tabel 7. Kumpulan topik untuk kelas negatif

Topik 0	Topik 1	Topik 2	Topik 3
pantai	lampu	bawa	fasilitas
nyari	jalan	gunung	kotor
sampah	gelap	turun	jumlah
buang	pantai	puntung	bau
botol	penerangan	rinjani	campur
plastik	malam	sampah	cewek
pecinta	pink	rokok	pantai
sembarangan	pasang	pulang	toilet
gili	buang	kotor	umum
bikin	jarang	alam	sedikit

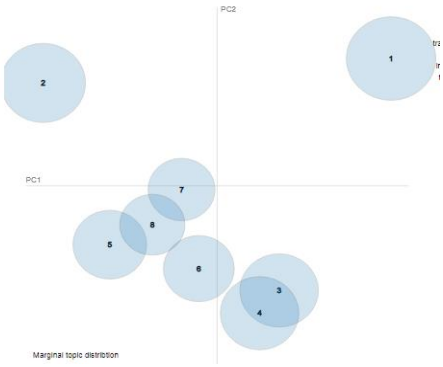
Topik 4	Topik 5	Topik 6	Topik 7
kereta	karang	jalan	air
gantung	terumbu	bikin	mati
rinjani	coral	liat	sulit
lombok	bleaching	wisata	musholla
tolak	mata	kuda	pantai
manusia	memandang	kotoran	cidomo
sampai	mengalami	daerah	salah
wisata	perairan	kondisi	uang
gunung	sebelah	lombok	bersih
indah	diambil	cidomo	kurang

Topik 8	Topik 9	Topik 10	Topik 11
trotoar	coretan	gempa	tutup
jalan	sembarangan	bumi	rinjani
kaki	rinjani	lombok	bakar
jarang	batu	turun	pendaki
lengkap	rembige	sepi	orang
sarana	cidomo	tumpukan	gunung
buat	kelambu	pembenahan	wisata
belum	lingkungan	jalan	arah
ATM	sadar	lokal	cuman
hotel	banget	wisatawan	main

Visualisasi pemodelan topik pada penelitian ini dibuat dalam bentuk *pyLDAvis* ditunjukkan oleh Gambar 5 dan 6. Pada Gambar 5 menunjukkan visualisasi untuk kelas positif dan Gambar 6 menunjukkan visualisasi untuk kelas negatif. Berdasarkan visualisasi sebaran topik yang diperlihatkan pada kelas positif dan negatif terdapat beberapa topik saling beririsan yang menandakan topik-

DOI: https://doi.org/10.29207/resti.v5i1.2587

topik tersebut mempunyai beberapa kata penyusun yang sama. Topik–topik yang saling beririsan pada kelas positif yaitu topik 7 dengan topik 8, topik 5 dengan topik 8, topik 8 dengan topik 6 dan topik 3 dengan topik 4 sedangkan topik-topik yang saling beririsan pada kelas negatif yaitu topik 2 dengan 5 dan topik 5 dengan 4, serta topik 7 dengan topik 12, topik 9, topik 6, topik 10, topik 11.



Gambar 5. Visualisasi topik kelas positif

Gambar 6. Visualisasi topik kelas negatif

4. Kesimpulan

Penelitian ini berhasil melakukan analisis sentimen menggunkan metode *Naive Bayes* serta pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) untuk topik pariwisata Lombok. Berdasarkan hasil analisis sentimen diperoleh 9.496 *tweet* dengan pembagian 8.996 *tweet* sentimen positif dan 500 *tweet* sentimen negatif, sehingga dapat disimpulkan bahwa lebih banyak wisatawan memberikan respon positif daripada negatif terhadap pariwisata Lombok. Hasil pengujian menunjukkan bahwa algoritma *Naive Bayes* mampu mengklasifikasi sentimen wisatawan dengan baik yang ditunjukkan dengan pengukuran hasil kinerja algoritma melalui 4 parameter yaitu nilai akurasi sebesar 92%, presisi 100%, recall 83,84% dan Spesifisitas 100%. Sedangkan pemodelan topik menggunakan algoritma LDA menghasilkan topik terbaik untuk kelas positif pada 8 topik dengan nilai koherensi 0,613 kemudian untuk kelas negatif topik terbaik pada 12 topik dengan nilai koherensi 0,528. Hasil interpretasi topik-topik yang sering diperbincangkan oleh masyarakat pada kelas positif yaitu tentang keindahan pantai di Pulau Lombok khususnya gili trawangan dan pantai senggigi. Sedangkan untuk kelas negatif topik yang banyak dibicarakan adalah mengenai sampah.

Saran untuk pengembangan penelitian ini kedepannya adalah menambah jumlah dataset yang digunakan serta melakukan identifikasi terhadap kata-kata yang mempunyai makna ambigu sehingga mampu meningkatkan nilai akurasi pada model klasifikasi sentimen dan pemodelan topik. Selain itu perlu dilakukan penambahan metode selain koherensi untuk pengujian validasi topik yang dihasilkan pada LDA seperti uji kesesuaian distribusi topik.

Ucapan Terimakasih

Peneliti memberikan apresiasi yang sebesar-besarnya serta mengucapkan terima kasih kepada Kementerian Riset dan Teknologi/Badan Riset dan Inovasi Nasional sebagai pemberi dukungan dana melalui hibah kompetitif nasional skim dosen pemula tahun pelaksanaan 2020.

Daftar Rujukan

[1] N. Islamy, "Analisis Sektor Potensial, Dapatkah Pariwisata Menjadi Lokomotif Baru Ekonomi Nusa Tenggara Barat?," *J. Indones. Tour. Hosp. Recreat.*, vol. 2, no. 1, pp. 1–10, 2019.

[2] K. a n o m K a n o m, "Strategi Pengembangan Kuta Lombok Sebagai Destinasi Pariwisata Berkelanjutan," *J. Master Pariwisata*, vol. 1, pp. 25–42, 2015.

[3] H. S. Utama, D. Rosiyadi, B. S. Prakoso, and D. Ariadarma, "Analisis Sentimen Sistem Ganjil Genap di Tol Bekasi Menggunakan Algoritma Support Vector Machine," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 243–250, 2019.

[4] S. N. J. Fitriyyah, N. Safriadi, and E. E. Pratama, "Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, pp. 279–285, 2019.

[5] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, pp. 334–339, 2019.

[6] A. Alamsyah, W. Rizkika, D. D. A. Nugroho, F. Renaldi, and S. Saadah, "Dynamic large scale data on Twitter using sentiment analysis and topic modeling case study: Uber," in *2018 6th International Conference on Information and Communication Technology, ICoICT 2018*, 2018, vol. 0, no. c, pp. 254–258.

[7] R. Ardianto, T. Rivanie, Y. Alkhalifi, F. S. Nugraha, and W. Gata, "Sentiment Analysis on E-Sports For Education Curriculum Using Naive Bayes and Support Vector Machine," *J. Comput. Sci. Inf.*, vol. 13, no. 2, pp. 109–122, 2020.

[8] M. Cendana and S. D. H. Permana, "Pra-Pemrosesan Teks Pada Grup Whatsapp Untuk Pemodelan Topik," *Junal Mantik Penusa*, vol. 3, no. 3, pp. 107–116, 2019.

[9] O. Somantri and D. Dairoh, "Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, pp. 191–196, 2019.

[10] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.

[11] Murnawan and A. Sinaga, "Pemanfaatan Analisis Sentimen untuk Pemeringkatan Popularitas Tujuan Wisata," *J. Penelit. Pos dan Inform.*, vol. 7, no. 2, pp. 109–120, 2017.

[12] M. A. Ulfa, B. Irmawati, and A. Y. Husodo, "Twitter Sentiment Analysis using Naive Bayes Classifier with Mutual Information

Feature Selection," *J. Comput. Sci. Informatics Eng.*, vol. 2, no. 2, pp. 106–111, 2018.
[13] G. R. Gustisa Wisnu, Ahmadi, A. R. Muttaqi, A. B. Santoso, P. K. Putra, and I. Budi, "Sentiment analysis and topic modelling of 2018 central java gubernatorial election using twitter data," *2020 Int. Work. Big Data Inf. Secur. IWBIS 2020*, pp. 35–40, 2020.
[14] P. M. R. C. Dinatha and N. A. Rakhmawati, "Komparasi Term Weighting dan Word Embedding pada Klasifikasi Tweet Pemerintah Daerah," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 9, no. 2, pp. 155–161, 2020.
[15] P. M. Prihatini, "Implementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia," *J. Matrix*, vol. 6, no. 3, pp. 174–178, 2016.
[16] V. Balakrishnan and W. Kaur, "String Based Multinomial Naive Bayes for Emotion Detection among Facebook Diabetes Community," *Procedia Comput. Sci.*, vol. 159, pp. 30–37, 2019.
[17] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010.
[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[19] W. Chen, Z. Xu, X. Zheng, Q. Yu, and Y. Luo, "Research on sentiment classification of online travel review text," *Appl. Sci.*, vol. 10, no. 15, 2020.
[20] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference*, 2012, no. July, pp. 952–961.

DOI: https://doi.org/10.29207/resti.v5i1.2587