

MODEL CLASSIFIER JUDUL BERITA PARIWISATA INDONESIA BERDASARKAN SENTIMEN

Ni Luh Putu Risma Dewi¹, I Nyoman Saputra Wahyu Wijaya^{*2}, I Ketut Purnamawan³, Ni Wayan Marti⁴

^{1,2,3,4}Universitas Pendidikan Ganesha, Singaraja

Email: ¹risma.dewi.2@undiksha.ac.id, ²wahyu.wijaya@undiksha.ac.id, ³purna@undiksha.ac.id,

⁴wayan.marti@undiksha.ac.id

^{*}Penulis Korespondensi

(Naskah masuk: 8 Agustus 2023, diterima untuk diterbitkan: 23 Januari 2024)

Abstrak

Kemajuan teknologi dan *platform* digital telah menyebabkan perubahan penting dalam industri pariwisata, termasuk penyebaran berita pariwisata. Artikel-artikel berita yang kerap dipublikasikan melalui portal media online dapat berdampak besar pada persepsi positif maupun negatif bagi pembaca. Oleh sebab itu, penelitian ini bertujuan untuk mengembangkan sebuah *classifier* yang mampu melakukan klasifikasi terhadap judul berita pariwisata berdasarkan sentimen. Sebelumnya, penelitian serupa biasanya lebih berfokus pada klasifikasi berita pariwisata Indonesia dengan menggunakan data dari *platform* Twitter. Namun, dalam penelitian ini, penulis melakukan pendekatan yang berbeda dengan menggunakan data dari portal berita online, dengan hanya mengambil judul berita sebagai sumber datanya. Penelitian berfokus pada analisis dan klasifikasi sentimen atau sikap emosional yang terkandung dalam judul-judul berita pariwisata. Dalam penelitian ini, digunakan metode *machine learning Support Vector Machine (SVM)*. Data dikumpulkan dari situs berita Detik.com dan diberi label secara manual sesuai dengan sentimen yang terkandung dalamnya. Proses *preprocessing teks* digunakan untuk mempersiapkan data judul berita pariwisata Indonesia sehingga fiturnya dapat diekstraksi dengan pendekatan *Binary Term Presence*. Data penelitian dibagi menjadi dua bagian, yaitu 90% untuk proses pelatihan (*training proses*) dan 10% untuk pengujian, menerapkan teknik *K-Fold Cross Validation* untuk membagi data dalam proses pelatihan. Hasil penelitian menunjukkan bahwa pendekatan *Binary Term Presence* berhasil mencapai akurasi sebesar 87,80% dengan kernel RBF. Penelitian ini memberikan kontribusi penting dalam memahami respon publik terhadap topik atau isu pariwisata yang dipublikasikan oleh media. Metode dan pendekatan yang dikembangkan diharapkan dapat menjadi panduan dalam menganalisis sentimen terkini terkait industri pariwisata dan berita yang terkait dengannya.

Kata kunci: Pariwisata, Sentimen, Klasifikasi, SVM, Detik.com

INDONESIAN TOURISM NEWS CLASSIFIER MODEL BASED ON SENTIMENT

Abstract

The advancement of technology and digital platforms has led to significant changes in the tourism industry, including the dissemination of tourism news. News articles are frequently published through online media portals and can have a considerable impact on readers' positive or negative perceptions. Therefore, this research aims to develop a classifier capable of categorizing tourism news headlines based on sentiment. Previous studies have mostly focused on classifying Indonesian tourism news using data from Twitter. However, this research takes a different approach by using data from online news portals, specifically focusing on the headlines as the data source. This study concentrates on analyzing and classifying the sentiments or emotional attitudes contained in tourism news headlines. The machine learning method used for this purpose is Support Vector Machine (SVM). Data is collected from the news site Detik.com and manually labeled according to the sentiments expressed in the headlines. Text preprocessing is applied to prepare the data of Indonesian tourism news headlines for feature extraction using the Binary Term Presence approach. The research data is divided into two parts, with 90% used for the training process and 10% for testing, implementing the K-Fold Cross Validation technique to partition the data during the training process. The research results demonstrate that the Binary Term Presence approach achieved an accuracy of 87.80% with the RBF kernel. This research contributes significantly to understanding the public's response to tourism topics or issues published by the media. The methods and approaches developed are expected to serve as guidelines for analyzing current sentiments related to the tourism industry and its associated news.

Keywords: *Tourism, Sentiment, Classification, SVM, Detik.com*

1. PENDAHULUAN

Dampak kemajuan teknologi dan *platform* digital pada industri pariwisata, termasuk penyebaran berita terkait pariwisata di Indonesia, semakin terasa. Dimana masyarakat kini lebih sering mencari informasi pariwisata melalui internet dan media sosial (Mursalina, Wibowo Yunanto dan Oktaviani, 2019). Portal berita online menjadi salah satu media massa digital yang populer dan banyak digunakan masyarakat. Kemunculan portal berita online ini sejalan dengan perkembangan masyarakat yang semakin dinamis dalam mencari informasi di media massa (Setiowati, Cheril dan Sary, 2021), yang juga menjadi sumber informasi paling populer bagi masyarakat Indonesia. Berita-berita pariwisata di Indonesia yang banyak dipublikasikan di portal media online dapat berdampak besar pada pandangan masyarakat yang membacanya, yang awalnya bersifat positif bisa berubah menjadi negatif, atau sebaliknya, dalam skala yang luas (Rameshbhai dan Paulose, 2019).

Berita yang dipublikasikan melalui portal berita online adalah salah satu bentuk media untuk mengekspresikan opini tentang beragam kejadian, informasi, fenomena, dan fakta terkait pariwisata di Indonesia. Berita-berita tersebut bersifat penilaian subjektif atau opini yang tetap memegang teguh etika jurnalistik (Juditha, 2017). Publikasi berita dapat menyebabkan masyarakat membentuk persepsi baru yang dapat mempengaruhi citra destinasi pariwisata yang diangkat dalam berita tersebut (Ugy Soebiyantoro, 2009).

Sebelumnya telah dilakukan penelitian mengenai klasifikasi sentimen: (Asiyah dan Fithriasari, 2016), (Hidayat, Garno dan Ridha, 2021), dan (Zalyhaty, 2021) yang dijadikan sebagai rujukan dalam penelitian ini. Meskipun banyak peneliti yang telah melakukan klasifikasi untuk mengetahui sentimen, berdasarkan penelusuran peneliti, belum terdapat penelitian yang secara khusus mengembangkan model *classifier* untuk mengklasifikasikan sentimen pada berita dengan menggunakan data judul berita sebagai sumber data penelitian dan berfokus pada kata kunci pariwisata, wisata, dan liburan.

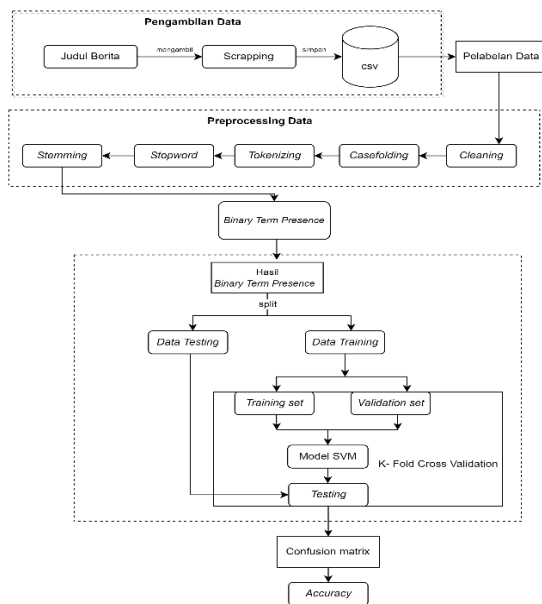
Penelitian ini bertujuan mengklasifikasikan judul berita pariwisata Indonesia berdasarkan sentimen. Klasifikasi sentimen dilakukan pada judul berita didasari karena kebanyakan orang cenderung menilai isi berita berdasarkan judulnya saja dari pada membaca seluruh artikel (Dor, 2003). Disamping itu, penelitian yang dilakukan oleh (Mursalina, Wibowo Yunanto dan Oktaviani, 2019) menunjukkan bahwa

judul berita juga memiliki potensi untuk mengarahkan pembaca ke arah ideologi tertentu yang tercermin dalam isi berita. Oleh sebab itu, judul sering kali dianggap merepresentasikan pandangan media terkait terhadap suatu hal yang diharapkan. Menggunakan model *classifier* untuk mengklasifikasikan sentimen atau sikap emosional yang terdapat dalam judul berita, dan metode ekstraksi fitur *Binary Term Presence* digunakan untuk memperoleh informasi yang relevan dari judul tersebut. Informasi dari *classifier* ini diharapkan dapat membantu pihak-pihak terkait dalam mendapatkan informasi yang mendukung pemahaman dan pengelolaan mengenai bagaimana masyarakat bereaksi terhadap topik atau isu pariwisata yang diangkat oleh media. Dengan demikian, model ini diharapkan menjadi alat bantu dalam pengambilan keputusan dan pemantauan terkait aspek pariwisata Indonesia secara lebih responsif, mengingat pentingnya dampak tersebut terhadap citra destinasi pariwisata Indonesia di mata masyarakat (Rohmatul dkk., 2022). Sebagai sumber data, penelitian ini menggunakan portal berita online Detik.com, yang dikenal sebagai portal berita yang menjaga netralitas berita (Hutami, Wijayanto dan Sulvianti, 2022), (Indrayani dan Dwi Wardani, 2018), dan sering dikunjungi oleh masyarakat Indonesia sesuai data yang ditemukan pada website alexa.com, sebuah *platform* yang menawarkan layanan analisis data komersial terkait lalu lintas situs web. Detik.com termasuk dalam 10 besar portal berita online yang sering dikunjungi. Metode *Support Vector Machine* (SVM) digunakan untuk mengklasifikasikan sentimen judul berita pariwisata di Indonesia. Dalam penelitian ini judul berita diklasifikasikan ke dalam kelas sentimen positif, netral dan negatif, pada pelabelan data dilakukan secara manual oleh anotator yaitu mahasiswa program studi Bahasa Indonesia pada keseluruhan data.

Penelitian ini menggunakan metode *Support Vector Machine* (SVM) karena metode ini dianggap tepat dalam mencari pemisah antara kelas atau *hyperplane*. SVM bekerja dengan memproyeksikan data ke dalam ruang dimensi yang lebih tinggi dan secara optimal membagi data ke dalam kelas-kelas pada *hyperplane* (Awalina, Bachtiar dan Indriati, 2022) Selain itu, berdasarkan beberapa penelitian sebelumnya, SVM telah terbukti unggul dalam mengklasifikasikan dokumen teks dibandingkan dengan metode klasifikasi tradisional machine learning lainnya seperti *K-Nearest Neighbor* (KNN) dan *Naïve Bayes* (Setiowati, Cheril dan Sary, 2021).

2. METODE PENELITIAN

Gambar 1 menggambarkan alur penelitian yang dilakukan oleh penulis yang mana pada penelitian ini dimulai dari proses pengambilan data dari website, pelabelan data, *preprocessing* data, ekstraksi fitur dengan *binary term presence*, pembangunan model, evaluasi model dan analisis hasil *wordcloud* data berdasarkan sentimen untuk melihat penyebaran kata pada masing-masing sentimen.



Gambar 1. Kerangka Alur Penelitian

2.1. Pengumpulan Data

Data pada penelitian ini diperoleh dari judul berita portal berita online melalui teknik *scraping* menggunakan bahasa pemrograman *Python* pada platform *Google Collaboratory* dengan library *BeautifulSoup*. Pengambilan data dilakukan dengan menggunakan *keyword* “pariwisata, wisata, dan liburan”. Data yang terkumpul disimpan dalam bentuk format *.csv* (*comma separated value*) dan diberi label. Pengambilan data dilakukan secara bertahap dalam rentang waktu dari bulan Januari – Desember tahun 2022, sesuai dengan *keyword* yang telah di tentukan. Hasil pengumpulan data dibagi menjadi tiga jenis: data *training*, *validation*, dan *testing*. Proporsi data *training* dan *testing* adalah 90% : 10%, dengan pembagian data dilakukan secara manual. Data 90% digunakan untuk proses *training* dan *validation*, yang dilakukan secara bersamaan dengan *K-Fold Cross Validation*, sementara data 10% digunakan sebagai data *testing* untuk menguji model yang di hasilkan oleh *K-Fold Cross Validation*.

2.2. Pelabelan Data

Pada penelitian ini, setiap data yang terkumpul melalui proses *scraping* akan diberikan label yang sesuai. Pelabelan dilakukan dengan

mengklasifikasikan judul berita menjadi sentimen negatif, sentimen netral, atau sentimen positif. Pelabelan mencakup seluruh data dan mempertimbangkan karakteristik informasi yang terdapat dalam masing-masing judul berita. Jika sebuah judul berita cenderung memberikan rekomendasi, pujian, atau ulasan positif terhadap suatu destinasi wisata yang layak dikunjungi, judul tersebut akan dikategorikan sebagai judul berita dengan sentimen positif. Judul berita dengan sentimen netral meujuk pada judul yang memberikan informasi deskriptif tanpa mencoba memengaruhi opini, sementara judul berita dengan sentimen negatif cenderung mengungkapkan kekurangan, kesulitan, atau ulasan yang kurang puas terhadap suatu destinasi wisata. Agar proses klasifikasi sentimen pada model menjadi lebih mudah, sentimen positif diberi label 1, sentimen netral diberi label 0, dan sentimen negatif diberi label -1. Dalam penelitian ini, proses pelabelan data dilakukan secara manual oleh tiga anotorator, yaitu mahasiswa tingkat akhir dari program studi pendidikan Bahasa Indonesia, dengan kualifikasi sebagai mahasiswa yang berada di semester 8 dan memiliki Indeks Prestasi Kumulatif (IPK) di atas 3,5. Contoh hasil pelabelan data dapat ditemukan dalam Tabel 1.

Tabel 1. Contoh Pelabelan Data

Label	Judul Berita
1	Bangga, Salah Satu Hotel Mewah Bali Raih Wordl Travel Award 2022
0	Masjid Agung Jamik Singaraja Bukti Toleransi Beragama Sejak Lama
-1	Bule Curhat Dilecehkan di Pantai Kuta, Jadi Kapok Liburan di Bali

2.3. Preprocessing Data

Tahapan selanjutnya setelah data diberikan label, dilakukan tahap-tahap *preprocessing*, data judul yang diperoleh dari portal berita online merupakan data teks yang tidak terstruktur karena masih mengandung banyak noise atau gangguan pada data tersebut. Oleh karena itu, data harus diseleksi terlebih dahulu agar data menjadi terstruktur dan memiliki keseragaman yang akan memudahkan dalam proses analisis. Berikut adalah urutan tahapan *preprocessing* yang dilakukan pada penelitian ini (Pramudita, Putro dan Makhmud, 2018):

- 1) *Cleaning* : *cleaning* data berfungsi menghapus karakter-karakter yang tidak diinginkan atau noise dari teks (Budiman dan Widjaja, 2020). Tanda baca, bilangan angka, kata-kata yang duplikat pada judul berita, spasi karakter dari teks kiri dan kanan dan menghapus kata yang kurang dari 3 huruf. Hal ini dilakukan karena keberadaan elemen-elemen tersebut dapat membuat data menjadi tidak efektif. Proses *cleaning* pada penelitian ini dilakukan dengan menggunakan *Python* dengan library ‘*re*’ dan ‘*string*’.

- 2) *CaseFolding*: *CaseFolding* bertujuan untuk mengubah keseluruhan huruf pada data menjadi huruf kecil (Nurrochman dan Prasasti, 2021). Hal ini dilakukan agar data yang diperoleh memiliki bentuk yang sama. Proses ini dilakukan dengan menggunakan fungsi *lower()* yang telah di sediakan oleh *python*.
- 3) *Tokenizing*: *Tokenizing* berfungsi untuk memisahkan atau membagi teks yang awalnya berbentuk kalimat menjadi sebuah kata atau *token* (Amrizal, 2018). Proses tokenizing pada penelitian ini dilakukan dengan menggunakan *library* NLTK dengan fungsi *word_tokenize()* pada *python*.
- 4) *Stopword* : *Stopword* berfungsi untuk menghapus kata-kata yang tidak memiliki arti, seperti dalam Bahasa Indonesia yaitu kata “yang”, “di”, “ke” (Rofiqi dkk., 2019). Proses *stopword* pada penelitian ini menggunakan modul ‘*stopword*’ dari *library* NLTK pada bahasa pemrograman *python*.
- 5) *Stemming* : *Stemming* berfungsi untuk menemukan bentuk kata dasar dan menghapus kata imbuhan awal (prefix) dan imbuhan akhir (sufiks) dalam sebuah kata (Irham, Adiwijaya dan Wisesty, 2019). Pada penelitian ini implementasi proses *stemming* dilakukan dengan menggunakan *library* *Sastrawi* dengan modul *Stemmer*.

2.4. Binary Term Presence

Setelah melalui proses *preprocessing*, penelitian ini melanjutkan pemrosesan data dengan mengubah data teks menjadi data numerik menggunakan metode ekstraksi fitur *Binary Term Presence*. *Binary Term Presence* adalah metode pembobotan teks yang mengukur keberadaan term atau fitur dalam sebuah dokumen. Jika *term* tersebut ada dalam dokumen, maka nilai fitur tersebut dalam vektor fitur akan diatur menjadi 1; sedangkan jika *term* tersebut tidak ada dalam dokumen, maka nilai fitur tersebut diatur menjadi 0 dalam ruang fitur (O’Keefe dan Koprinska, 2009). Pemilihan *binary term presence* sebagai metode ekstraksi fitur dikarenakan penelitian ini berfokus pada keberadaan atau ketiadaan kata dalam kalimat, tanpa mempertimbangkan relevansi kata secara menyeluruh. Pendekatan ini bertujuan untuk menyederhanakan representasi data tanpa memperhatikan signifikansi atau hubungan kata. Dalam penelitian ini menggunakan modul *CountVectorizer* dari *scikit-learn* dalam bahasa pemrograman *Python* dengan menggunakan parameter *Binary = True*.

2.5. Membangun Model

Pada tahap pembangunan model *machine learning*, data yang telah melalui proses *preprocessing* dan ekstraksi fitur digunakan. Metode yang digunakan dalam penelitian ini adalah *Support*

Vector Machine (SVM) untuk melakukan klasifikasi. Proses pembentukan model SVM dimulai dengan menggunakan hasil ekstraksi fitur dari *CountVectorizer*. Selanjutnya, dilakukan klasifikasi sentimen dengan metode SVM dan menerapkan teknik *K-Fold Cross Validation* untuk menghasilkan model. Hasil klasifikasi sentimen yang diperoleh dari metode SVM terdiri dari sentimen kelas positif, netral, atau negatif. Model yang terbentuk kemudian disimpan dan diuji menggunakan data testing, dan performa model dievaluasi menggunakan *confusion matrix*. Dalam penelitian ini, untuk melakukan klasifikasi sentimen, peneliti menggunakan fungsi *Support Vector Classifier* (SVC) yang disediakan oleh *library sklearn* dalam bahasa pemrograman *Python*. Parameter-parameter yang relevan, seperti kernel yang digunakan, akan disesuaikan dengan kebutuhan penelitian. Parameter lain yang juga relevan, seperti parameter C yang mengontrol pinalti *hyperplane*, dan *gamma* yang mengontrol jangkauan pengaruh sampel, akan diatur sesuai dengan kebutuhan penelitian guna mendapatkan performa model yang optimal.

2.6. Evaluasi Model

Evaluasi model pada penelitian ini menggunakan *confusion matrix* untuk melihat nilai *accuracy* yang dihasilkan untuk mengetahui seberapa tepat model mampu melakukan klasifikasi dengan benar.

3. HASIL DAN PEMBAHASAN

3.1. Pengambilan dan Pelabelan Data

Hasil pengambilan data yang dilakukan dengan teknik *scraping* pada halaman website detik.com, dengan *keyword* yang digunakan adalah “pariwisata, wisata, dan liburan” dengan komponen yang di ambil adalah *title*. Didapatkan total data judul berita adalah 3680 judul berita dengan rentang waktu pengambilan dari bulan Januari hingga Desember tahun 2022.



Gambar 2. Website Detik.Com

Dari total 3.680 judul berita yang di peroleh, selanjutnya data tersebut diberikan label oleh para anotator sesuai kriteria Adapun hasil pelabelan didapatkan pada kelas sentimen positif memiliki frekuensi tertinggi, yaitu sebanyak 2.587 judul, diikuti oleh kelas sentimen netral sebanyak 550 judul, dan kelas sentimen negatif sebanyak 545 judul.

Adapun tabel perbandingan hasil pelabelan data dapat dilihat pada Tabel 2.

Kelas	Jumlah
Positif	2.587
Netral	550
Negatif	545

3.2. Preprocessing Data

Tabel 3 menampilkan hasil data setelah melewati proses *preprocessing* yang terdiri dari beberapa tahapan proses, seperti pembersihan data (*cleaning*), penyesuaian huruf kecil (*casefolding*), pembagian menjadi token (*tokenizing*), penghapusan kata-kata umum (*stopword*), dan pemangkasan kata-kata menjadi bentuk dasar (*stemming*). Data yang terlihat pada Gambar 3 merupakan contoh 2 data dari keseluruhan data yang sudah bersih dan siap digunakan dalam proses klasifikasi.

Sebelum	Metode	Sesudah
"Polisi Terapkan Sistem Ring Saat Pengamanan Tahun Baru di GWK"	<i>Cleaning</i>	"Polisi Terapkan Sistem Ring Saat Pengamanan Tahun Baru"
	<i>CaseFolding</i>	"polisi terapkan sistem ring saat pengamanan tahun baru"
	<i>Tokenizing</i>	['polisi', 'terapkan', 'sistem', 'ring', 'saat', 'pengamanan', 'tahun', 'baru']
	<i>Stopword</i>	['polisi', 'terapkan', 'sistem', 'ring', 'pengamanan']
	<i>Stemming</i>	['polisi', 'terap', 'sistem', 'ring', 'aman']
"4 Ribu Turis Kunjungi Goa Batu Cermin Labuan Bajo Selama 2,5 Bulan"	<i>Cleaning</i>	"Ribu Turis Kunjungi Batu Cermin Labuan Bajo Selama Bulan"
	<i>CaseFolding</i>	['ribu', 'turis', 'kunjungi', 'batu', 'cermin', 'labuan', 'bajo', 'selama', 'bulan']
	<i>Tokenizing</i>	['ribu', 'turis', 'kunjungi', 'batu', 'cermin', 'labuan', 'bajo', 'selama', 'bulan']
	<i>Stopword</i>	['ribu', 'turis', 'kunjungi', 'batu', 'cermin', 'labuan', 'bajo']
	<i>Stemming</i>	['ribu', 'turis', 'kunjung', 'batu', 'cermin', 'labu', 'bajo']

Gambar 3. Hasil *Preprocessing* Data

3.3. Support Vector Machine Model

Dalam proses klasifikasi, algoritma menggunakan data *training* untuk membentuk model klasifikasi sentimen. Model yang terbentuk merupakan hasil dari pengetahuan yang direpresentasikan dan akan digunakan untuk memprediksi kelas dari data baru yang belum pernah ada sebelumnya. Keberhasilan mesin memahami pola data sangat bergantung pada jumlah data *training set* yang digunakan. Penelitian ini peneliti melakukan empat percobaan, dengan perbedaan utama terletak pada jumlah *Fold* atau (*k*) yang bervariasi. Pada *K-Fold* setiap iterasi salah satu fold digunakan sebagai data *validation set* sementara *K-1 Fold* lainnya digunakan sebagai data *training set*. Proses ini diulang sebanyak *K* kali sehingga setiap iterasi menjadi data *validation set* sebanyak 1 kali. Metode ini dipilih untuk memastikan bahwa model diperiksa dengan beragam data pada setiap iterasi karena pada setiap iterasi sehingga mendapatkan model yang lebih akurat. Tabel 3 menampilkan jumlah *Fold* yang digunakan pada percobaan.

Percobaan	Jumlah Fold
Ke – 1	10
Ke – 2	15
Ke – 3	20
Ke – 4	25

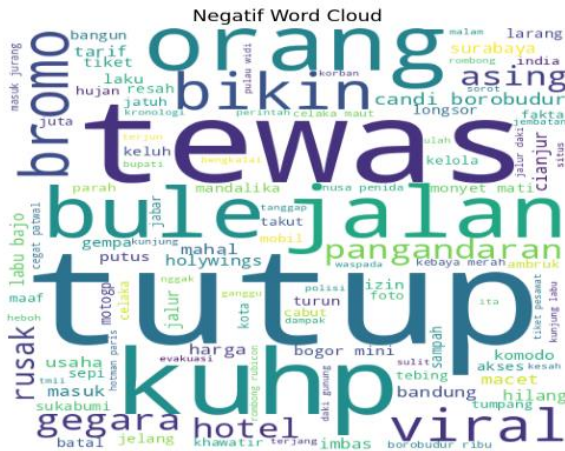
Berdasarkan tabel percobaan, dilakukan pembuatan model klasifikasi sentimen dengan SVM. SVM akan mempelajari karakteristik kata-kata setiap kelas, dan hasil dari *training* adalah model dengan pola yang dapat digunakan untuk klasifikasi pada proses selanjutnya, yaitu proses *testing*. Pembuatan model klasifikasi dengan SVM menggunakan kernel RBF, Polinomial, dan Sigmoid untuk melihat pada kernel mana model mendapatkan akurasi terbaik. Adapun hasil akurasi pembuatan model ditunjukkan pada Tabel 4.

Tabel 4. Hasil *Traning* dan *Testing* SVM

Kernel	K-fold	Training	Testing
RBF	10	80,96%	84,28%
	15	83,18%	85,90%
	20	85,45%	85,45%
	25	87,12%	87,80%
	10	78,91%	69,91%
Polinomial	15	81,36%	70,73%
	20	83,63%	70,73%
	25	84,84%	70,75%
	10	61,02%	60,97%
	15	64,70%	60,16%
Sigmoid	20	65,85%	69,64%
	25	66,66%	69,95%

Berdasarkan Tabel 4, terlihat bahwa terjadi peningkatan akurasi seiring dengan peningkatan nilai *Fold* yang digunakan pada setiap model SVM yang menggunakan kernel tersebut. Namun, dapat diperhatikan bahwa pada hasil *training* dan *testing* kernel RBF menunjukkan performa yang lebih baik. Pada *Fold* 25, model dengan kernel RBF mencapai akurasi *training* sebesar 87,12% dan akurasi *testing* sebesar 87,80, dibandingkan dengan kernel Polinomial dan Sigmoid. Hal ini dapat diartikan bahwa model SVM dengan ekstraksi fitur *Binary Term Presence* menggunakan kernel RBF memiliki kemampuan yang lebih baik dalam mempelajari pola-pola yang ada dalam data, tanpa mengalami *overfitting* maupun *underfitting*. Hasil ini mendukung konsep dalam teori ilmiah bahwa kernel RBF, dengan kemampuannya dalam memetakan data ke dalam ruang fitur yang lebih kompleks, cenderung lebih baik dalam menggeneralisasi dan menangkap pola yang terdapat dalam data. Berikut merupakan confusion matrix dari model terbaik.

jogja” memiliki sentimen netral dikarenakan makna judul berita tersebut bersifat dekritif.



Gambar 7. Wordcloud Sentimen Negatif

Gambar 7 merupakan hasil visualisasi dalam konteks pariwisata Indonesia pada kelas sentimen negatif. Visualisasi tersebut menunjukkan kata-kata yang sering muncul dalam judul berita pariwisata yang memiliki sentimen negatif. Kata-kata yang sering muncul dalam judul berita tersebut meliputi "tutup", "tewas", "kuhp", "bule", "jalan", dan kata-kata lainnya yang memiliki konotasi negatif.

Dalam konteks pariwisata Indonesia, kata-kata ini mencerminkan masalah-masalah atau situasi yang merugikan atau menimbulkan ketidaknyamanan dalam industri pariwisata. Kata "tutup" mengacu pada penutupan objek wisata atau layanan yang berdampak negatif pada wisatawan yang berkunjung pada tempat tersebut. Kata-kata seperti "tewas" atau "kuhp" mengindikasikan insiden atau kejadian yang melibatkan kecelakaan, tindakan kriminal, atau peraturan terkait pariwisata yang dirasa dapat merugikan pariwisata. Sedangkan kata "bule" sering digunakan dalam konteks negatif untuk mengacu pada turis asing atau orang non-Indonesia. Kata-kata ini memberikan gambaran tentang topik dan isu-isu yang sering dikaitkan dengan sentimen negatif dalam berita pariwisata di Indonesia.

4. KESIMPULAN

Penelitian ini telah menjawab permasalahan dengan membangun sebuah *classifier* untuk melakukan klasifikasi berdasarkan sentimen pada judul berita pariwisata Indonesia. Data yang dikumpulkan dari portal berita online terkait dengan pariwisata Indonesia, di mana judul-judul berita tersebut dilabeli secara manual sesuai dengan sentimen yang terkandung. Melalui proses *preprocessing* teks, data judul berita pariwisata Indonesia siap untuk diekstraksi fitur. Ekstraksi fitur dilakukan dengan pendekatan *Binary Term Presence*, yang menghasilkan representasi vektor numerik dari setiap judul berita. Model klasifikasi yang digunakan adalah *Support Vector Machine*

(SVM), yang dilatih dan diuji dengan data yang telah diproses. Evaluasi menunjukkan bahwa model, mampu melakukan klasifikasi sentimen dengan baik yaitu akurasi training 87,12% dan akurasi testing sebesar 87,80% berdasarkan judul berita pariwisata Indonesia. Penelitian ini menghasilkan *classifier* yang cukup efektif dan akurat dalam menentukan sentimen terhadap judul berita. Dengan demikian hasil klasifikasi nantinya dapat digunakan sebagai pembantu dalam upaya pemahaman respon publik terhadap topik dan isu pariwisata yang dipublikasi oleh media.

DAFTAR PUSTAKA

- AMRIZAL, V., 2018. Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *Jurnal Teknik Informatika*, 11(2), hal.149–164. <https://doi.org/10.15408/jti.v11i2.8623>.
- ASIYAH, S.N. dan FITHRIASARI, K., 2016. Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K-Nearest Neighbor. *Jurnal Sains dan Seni ITS*, 5(2), hal.317–322.
- AWALINA, A., BACHTIAR, F.A. dan INDRIATI, I., 2022. Klasifikasi Ulasan Palsu Menggunakan Borderline Over Sampling (BOS) dan Support Vector Machine (SVM) (Studi Kasus: Ulasan Tempat Makan). *Jurnal Teknologi Informasi dan Ilmu Komputer*, 9(2), hal.419. <https://doi.org/10.25126/jtiik.2022925692>.
- BUDIMAN, A.E. dan WIDJAJA, A., 2020. Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir. *Jurnal Teknik Informatika dan Sistem Informasi*, 6(3), hal.475–488. <https://doi.org/10.28932/jutisi.v6i3.2892>.
- DOR, D., 2003. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5), hal.695–721. [https://doi.org/10.1016/S0378-2166\(02\)00134-0](https://doi.org/10.1016/S0378-2166(02)00134-0).
- HIDAYAT, T.F.T., GARNO, G. dan RIDHA, A.A., 2021. Analisis Sentimen Opini Pemindahan Ibu Kota Pada Twitter Dengan Metode Support Vector Machine. *Jurnal Ilmu Komputer*, 14(1), hal.49. <https://doi.org/10.24843/jik.2021.v14.i01.p06>.

- HUTAMI, W.P., WIJAYANTO, H. dan SULVIANTI, I.D., 2022. Penerapan Support Vector Machine dengan SMOTE Untuk Klasifikasi Sentimen Pemberitaan Omnibus Law Pada Situs CNNIndonesia.com. *Xplore: Journal of Statistics*, 11(1),hal.2635.<https://doi.org/10.29244/xplore.v11i1.852>.
- INDRAYANI, H. dan DWI WARDANI, A., 2018. Netralitas Konten Berita Online. *Interaksi : Jurnal Ilmu Komunikasi*, 7(1), hal.1–7.
- IRHAM, L.G., ADIWIJAYA, A. dan WISESTY, U.N., 2019. Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine. *Jurnal Media Informatika Budidarma*, 3(4), hal.284. <https://doi.org/10.30865/mib.v3i4.1410>.
- JUDITHA, C., 2017. Sentimen Dan Imparsialitas Isi Berita Tentang Ahok Di Portal Berita Online. *Jurnal Penelitian Komunikasi dan Pembangunan*, 18(1), hal.57. <https://doi.org/10.31346/jpkp.v18i1.839>.
- MURSALINA, H., WIBOWO YUNANTO, P. dan OKTAVIANI, V., 2019. Pengembangan Media Cetak dan Digital untuk Dellsand Kreasi Kaos Nama Sebagai Media Promosi. *PINTER : Jurnal Pendidikan Teknik Informatika dan Komputer*, 3(2), hal.127–131. <https://doi.org/10.21009/pinter.3.2.7>.
- NURROCHMAN, M.M. dan PRASASTI, A.L., 2021. Implementasi Machine Learning Untuk Mendeteksi Unsur Depresi Pada Tweet Menggunakan Metode Naïve Bayes (Machine Learning Implementation for Depression Detection in Tweet Using Naïve Bayes Method). *e-Proceeding of Engineering*, 8(5), hal.6250–6257.
- O'KEEFE, T. dan KOPRINSKA, I., 2009. Feature selection and weighting methods in sentiment analysis. *ADCS 2009 - Proceedings of the Fourteenth Australasian Document Computing Symposium*, (January 2009), hal.67–74.
- PRAMUDITA, Y.D., PUTRO, S.S. dan MAKHMUD, N., 2018. Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 5(3), hal.269. <https://doi.org/10.25126/jtiik.201853810>.
- RAMESHBHAI, C.J. dan PAULOSE, J., 2019. Opinion mining on newspaper headlines using SVM and NLP. *International Journal of Electrical and Computer Engineering*, 9(3), hal.2152–2163. <https://doi.org/10.11591/ijece.v9i3.pp2152-2163>.
- ROFIQI, M.A., FAUZAN, A.C., AGUSTIN, A.P. dan SAPUTRA, A.A., 2019. Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, 1(2), hal.58–64. <https://doi.org/10.28926/ilkomnika.v1i2.18>.
- ROHMATUL, S., GEDE, I.K., PUTRA, D. dan ARYA, I.P., 2022. Analisis Sentimen Berita Universitas Udayana Menggunakan Metode Full Text Search (Natural Language) Dan Rule Based Method. 3(1).
- SETIOWATI, D., CHERIL, R. dan SARY, M.P., 2021. Pengaruh Perkembangan Digitalisasi Media Pada Portal Berita Kompas Terhadap Minat Baca Masyarakat (Studi Kasus Pada Mahasiswa Program Studi Ilmu Komunikasi Universitas Negeri Jakarta). *Mediakom : Jurnal Ilmu Komunikasi*, 5(1), hal.29–38. <https://doi.org/10.35760/mkm.2021.v5i1.3377>.
- UGY SOEBIYANTORO, 2009. Pengaruh Ketersediaan Sarana Prasarana, Sarana Transportasi Terhadap Kepuasan Wisatawan. *Jurnal Manajemen Pemasaran*, [daring] 4(1), hal.16–22. Tersedia pada: <<http://puslit2.petra.ac.id/ejournal/index.php/mar/article/view/18082>>.
- ZALYHATY, L. QOADARY, 2021. Analisis sentiment tanggapan masyarakat terhadap vaksin covid-19 menggunakan algoritma Support Vector Machine.