Identifikasi Dual Sentimen Terhadap Ulasan Objek Wisata di Daerah Istimewa Yogyakarta

D. B. Prasetyo
Universitas Islam Indonesia
Yogyakarta, Indonesia
16523212@students.uii.ac.id

A. F. Hidayatullah
Universitas Islam Indonesia
Yogyakarta, Indonesia
fathan@uii.ac.id

***Abstraksi*—Saat ini review atau ulasan jadi sangat penting karena bisa jadi sumber informasi dan penilain terhadap suatu objek. Di internet kita dapat dengan mudah menemukan ulasan-ulasan terkait banyak hal termasuk objek wisata. Namun dengan banyaknya ulasan yang ada di internet tidak semuanya dapat dipahami dengan mudah. Masih banyak ulasan-ulasan yang memiliki ambigiutas, sehingga sulit untuk menentukan inti sarinya. Salah satu cara dalam menangani masalah ini ialah dengan menggunakan *natural language processing* (NLP). Jenis NLP yang tepat dalam hal ini ialah *sentiment analysis*. Sedangkan algoritma klasifiaksi yang digunakan ialah *Support Vector Machine, Naïve Bayes Classifier, dan Logistic Regression*. Berdasarkan percobaan yang dilakukan dalam memprediksi kalimat *dual sentiment*, algoritma *Support Vector Machine* memiliki performa paling baik dengan tingkat akurasi sebesar 83%.**

Kata kunci*—ulasan,objek wisata,*natural language processing, sentiment analysis

I. PENDAHULUAN

Saat ini Daerah Istimewa Yogyakarta (DIY) merupakan salah satu tujuan wisata terbesar yang ada di Indonesia. Data menunjukkan setiap tahunnya jumlah wisatawan yang berkunjung ke DIY selalu meningkat. Pada tahun 2017 total wisatawan yang berkunjung ke DIY sebanyak 5.229.298 orang, terdiri dari 397.951 orang wisatawan mancanegara dan 4.831.347 wisatawan lokal (nusantara). Dari 131 objek wisata yang ada di DIY tercatat 601.781 kunjungan berasal dari wisatawan mancanegara, sedangkan 25.349.012 kunjungan berasal dari wisatawan lokal [1].

Dengan selalu meningkatnya kunjungan wisatawan ke DIY tentu perlu melihat pendapat atau opini dari para wisatawan. Hal ini perlu dilakukan agar dapat meningkatkan kualitas objek wisata yang ada di DIY. Pendapat atau opini dari para wisatawan saat ini sangat banyak dan dapat dengan mudah kita temukan di situs ulasan ataupun sosial media. Cara terbaik dalam mengelola kumpulan pendapat tersebut adalah menggunakan *sentiment analysis* (analisis sentimen).

Analisis sentimen mengunakan *natural language processing* (NLP), *text analysis* dan *computational techniques* untuk mengotomatisasi ekstraksi atau klasifikasi sentimen dari suatu *reviews* (ulasan) [2]. Kebanyakan penelitian tentang analisis sentimen berfokus pada mengidentifikasi polaritas dari suatu kalimat seperti positif, negatif dan netral [3]. Namun, terkadang dalam suatu ulasan terdapat dua sisi sentimen sekaligus (*dual sentiment*). Dual sentiment pada umumnya berisi sentimen positif dan sentimen negatif. Dengan adanya lebih dari satu sentimen pada suatu kalimat tentu akan menyulitkan dalam menentukan inti sari dari kalimat tersebut.

Oleh karena itu, penelitian ini akan mengindentifikasi sentimen yang ada pada suatu kalimat dan juga dapat menentukan kalimat tersebut termasuk *dual sentiment* atau *single sentiment*. Penelitian ini dilakukan dengan mengambil data pendapat wisatawan tentang objek wisata yang ada di DIY dan sekitarnya melalui situs ulasan Tripadvisor dan Google Reviews.

II. METODOLOGI

A. Data

Dalam penelitian ini data yang akan digunakan adalah kumpulan ulasan terkait objek wisata yang diperoleh dari situs ulasan.

i. Dataset

Dataset yang digunakan pada penelitian kali ini diperoleh menggunakan teknik *web scraper*. *Web scraper* adalah metode yang digunakan untuk mengumpulkan informasi dari seluruh internet. Proses *web scraper* dilakukan dengan cara mengambil ulasan terkait beberapa objek wisata terkenal yang ada di DIY dan sekitarnya melalui situs ulasan Tripadvisor dan Google Reviews menggunakan bantuan *extensions tool* yang ada di Google Chrome yaitu *Data Miner*. Dari proses *web scraping* yang telah dilakukan, diperoleh setidaknya 10.000 ulasan. Dari ulasan-ulasan tersebut terdapat 5 objek wisata dengan jumlah ulasan terbanyak yaitu:

Tabel 1. Objek wisata dengan ulasan terbanyak

Objek Wisata	Jumlah Ulasan
Candi Borobudur	993
Malioboro	881
Pantai	798
Keraton Yogyakarta	742
Gunung Merapi	608

Data-data yang telah diperoleh akan dibagi menjadi dua *dataset* yang selanjutnya akan dilakukan *labelling data*. *Dataset* pertama mengambarkan sentimen dari suatu ulasan. Pada *dataset* ini setiap ulasan diberikan label positif, negatif, atau netral berdasarkan sentimen yang diwakili oleh ulasan tersebut. Untuk ulasan yang memiliki sentimen positif (baik) akan diberikan label positf, begitu pula untuk ulasan yang memiliki sentimen negatif (buruk) akan diberikan label negatif. Sedangkan label netral diberikan kepada ulasan yang tidak memiliki sentimen negatif maupun sentimen positif. Sebagian ulasan yang ada pada *dataset* pertama diberikan label

secara manual, sedangkan sisanya dilakukan otomatisasi menggunakan teknik *pseudo labelling.*

Tabel 2. Contoh *dataset* pertama

Ulasan	Label
Candi Bororbudur sangat indah dan megah	positif
Malioboro tempat yang nyaman buat belanja	positif
Kamar mandi kotor dan bau busuk	negatif
Tiket masuknya sangat mahal	negatif
Saya dan keluaraga pergi ke Borobudur	netral
Malioboro ada ditengah kota Yogyakarta	netral

Pada *dataset* kedua, setiap ulasan akan diberikan label *dual* atau *single*. Label *dual* diberikan untuk ulasan yang memiliki dua sisi sentimen sekaligus didalamnya (positif dan negatif). Sedangkan label *single* untuk ulasan yang hanya memiliki satu sentimen baik itu positif, negatif, ataupun netral. Pemberian label pada *dataset* kedua ini dilakukan secara manual.

Tabel 3. Cotoh *dataset* kedua

Ulasan	Label
Tiket museum murah tapi koleksi banyak yang rusak	dual
Tempatnya bagus, sayang harga makanannya mahal	dual
Kamar mandi kotor dan bau busuk	single
Malioboro tempat yang nyaman buat belanja	single

Adapun komposisi dari kedua dataset yang telah diberikan label dapat dilihat pada tabel 4 dan 5.

Tabel 4. Komposisi *dataset* pertama

Label	Jumlah
negatif	3333
netral	3333
positif	3334
Total	10000

Tabel 5. Komposisi *dataset* kedua

Label	Jumlah
dual	1699
single	1710
Total	3409

Dataset kedua memiliki jumlah data lebih sedikit dibandingkan dengan *dataset* pertama karena dari kesuluruhan data yang ada, hanya dapat ditemukan sekitar 1699 ulasan yang memiliki dua sentimen sekaligus didalamnya. Untuk tetap menjaga keseimbangan data saat dilakukan *training dataset* nantinya, maka ulasan dengan *single sentiment* dipilih mengikuti jumlah ulasan *dual sentiment.*

ii. Pre-processing

Dataset yang ada akan melalui *pre-processing* atau praproses terlebih dahulu sebelum digunakan. Proses ini bertujuan untuk menghindari data yang kurang sempurna, data yang bermasalah, dan data-data yang tidak konsisten[5]. Adapun tahapan-tahapan dalam *pre-processing* antara lain :

1) menghapus karakter yang tidak berguna
2) *case folding*
3) menghapus *stopwords*
4) *stemming*
5) mengubah *slang words*

Contoh dari penerapan *pre-processing* dapat dilihat pada Tabel 4.

Tabel 6. Contoh penerapan *pre-processing*

Sebelum	Setelah
Tiket masuknya mahall bgt	tiket mahal banget
Borobudur sangat indah dan megah, lebih lebih ukirannya sangat luar biasa.	borobudur indah megah ukir
saya akan selalu kembali ke kota ini….. penuh dengan kenangan	kota penuh kenang
Harga parkirnya terlalu mahal sampai 15.000	harga parkir mahal

iv. Ekstraksi Fitur

Setelah melalui *pre-processing* akan dilakukan ekstraksi fitur pada dataset. Salah satu tahapan dalam ekstraksi fitur adalah tokenisasi. Tokenisasi bertujuan untuk membagi teks baik itu berupa sebuah kalimat, paragraf, ataupun dokumen menjadi bagian-bagian yang lebih kecil. Berdasarkan hasil dari tokenisasi yang telah dilakukan, maka dapat diketahui frekuensi kata yang paling sering muncul.

Tabel 7. Kata dengan jumlah kemunculan terbanyak

Kata	Jumlah kemunculan
Jalan	3814
Pantai	1903
Candi	1570
Museum	1473
Bagus	1451
Indah	1361
Wisata	1320
Malioboro	1119
Foto	1115
Sejarah	1099

Tahapan selanjutnya dalam ekstraksi fitur ialah mengubah data yang sebelumnya merupakan fitur teks menjadi sebuah representasi *vector*. Data yang telah menjadi *vector* kemudian akan dilakukan perhitungan menggunakan *TF-IDF* untuk mendapatkan nilai yang berbobot untuk suatu kata.

B. Metode Klasifikasi

Metode *machine learning* yang digunakan dalam peneitian ini adalah *supervised learning*. Dalam melakukan prediksi menggunakan *supervised learning* dibutuhkan *training dataset* sebagai dasar pembelajaran. Algoritma supervised learning yang akan digunakan antara lain *Support Vector Machine* (SVM), *Naïve Bayes Classifier* (NBC), dan *Logistic Regression*.

i. *Support Vector Machine*

Tujuan dari algoritma *Support Vector Machine* adalah untuk mencari *hyperplane* terbaik. *Hyperplane* adalah garis yang memisahkan tiap-tiap kelompok atau kelas sebuah data.

ii. *Naïve Bayes Classifier*

Naïve Bayes Classifier adalah model pembelajaran mesin yang menggunakan metode probabilistik dalam melakukan klasifkasi. Algoritma klasifikasi ini berdasarkan *Bayes' Theorem* yang dikemukakan oleh Thomas Bayes.

iii. *Logistic Regression*

Logistic Regression sangat cocok digunakan untuk memprediksi ketika variabel dependen atau output suatu data bersifat biner. Sedangkan untuk memprediksi data yang memiliki lebih dari dua kemungkinan maka akan digunakan *multinomial logistic regression.*

III. PENELITIAN TERDAHULU

Penelitian terkait *sentiment analysis* sudah pernah dilakukan sebelumnya, baik itu oleh peneliti dalam negeri maupun luar negeri. Penelitian-penelitian tersebutlah yang menjadi contoh dan acuan dalam penelitian ini.

Penelitian yang dilakukan oleh Paulina Aliandu [6], ia menggunakan tempat wisata yang ada di kota Kupang seperti hotel, tempat makan, dan tempat belanja sebagai objek penelitiannya. Dalam penelitian ini ia menggunakan metode *supervised learning* dalam mengklasifikasikan data ulasan yang diperolehnya dari Foursquare.

Nur Azizah Vidya [7] melakukan penelitian tentang *sentiment analysis* terhadap reputasi merek dari suatu perusahaan operator telekomunikasi seluler di Indonesia. Ada beberapa aspek yang menjadi bahan penilain dalam penelitian ini yaitu layanan 3G, layanan 4G, layanan pesang singkat, layanan telepon, dan layanan data selular. Metode klasifikasi yang digunakan adalah *Naïve Bayes Classifier*, *Support Vector Machine* , dan *Decision Tree*.

Sedikit berbeda dari dua penelitian sebelumnya, Peter D. Turney [3] dalam penelitiannya menggunakan metode unsupervised learning dalam melakukan klasifikasi ulasan.

IV. PEMBAHASAN

Hasil *dataset* yang telah diperoleh melalui tahap *pre-processing* dan ekstraksi fitur, selanjutnya akan dilakukan uji coba klasifikasi menggunakan algoritma yang telah ditentukan sebelumnya. Percobaan akan dilakukan sebanyak dua kali. Pada percobaan pertama akan memprediksi sisi sentimen suatu data. Percobaan kedua akan memprediksi suatu data termasuk kedalam kelompok *dual* sentimen atau *single* sentimen.

A. Traininng Dataset

Seperti yang telah dijelaskan diawal bahwa penelitian kali ini akan menggunakan metode *supervised learning*, jadi perlu dilakukan *training dataset* terlebih dahulu. Dalam melakukan *training dataset*, data yang ada akan dibagi menjadi dua yaitu *train data* dan *test data*. *Train data* akan digunakan untuk melatih algortima *machine learning*. Sedangkan *test data* akan digunakan untuk mengevaluasi atau menguji algoritma yang dilatih sebelumnya. Adapun pembagian *train data* dan *test data* pada *dataset* pertama dan *dataset* kedua adalah sebagai berikut.

Tabel 8. Pembagian *train data* dan *test data*

	Train data	***Test data***	**Total**
***dataset* pertama**	6.700	3.300	10.000
dataset kedua	2.284	1.125	3.409

B. Pengujian Menggunakan Dataset Pertama

Pengujian pertama dilakukan untuk mengetahui performa SVM, NBC, dan *Logistic Regression* dalam memprediksi sentimen pada suatu ulasan. Pengujian ini menggunakan dataset pertama yang memprediksi suatu ulasan termasuk sentimen positif, negatif atau netral. Berdasarkan pengujian yang telah dilakukan, maka didapatkan nilai akurasi untuk setiap algoritma seperti yang ditampilkan pada tabel 9.

Tabel 9. Nilai akurasi pengujian pertama

	Akurasi
SVM	0,8876
NBC	0,7358
Logisttic Regression	0,8730

Nilai akurasi tersebut diperoleh berdasarkan rasio prediksi yang bernilai benar dari keseluruhan prediksi. Adapun informasi yang lebih lengkap terkait prediksi setiap algoritma dapat dilihat pada tabel-tabel dibawah ini.

Tabel 10. Hasil pengujian pertama menggunakan SVM

Nilai Sebenarnya	**Prediksi**		
	negatif	**netral**	**positif**
negatif	952	84	50
netral	87	967	45
Positif	38	67	1010

Tabel 11. Hasil pengujian pertama menggunakan NBC

Nilai Sebenarnya	**Prediksi**		
	negatif	**netral**	**positif**
negatif	845	121	120
netral	151	770	178
Positif	146	156	813

Tabel 12. Hasil pengujian pertama menggunakan *Logistic Regression*

Nilai Sebenarnya	**Prediksi**		
	negatif	**netral**	**positif**
negatif	948	83	55

netral	112	950	37
Positif	50	82	983

C. Pengujian Menggunakan Dataset Kedua

Pengujian kedua dilakukan dengan cara dan metode yang sama dengan pengujian sebelumnya, hanya saja pada pengujian kedua data yang diuji adalah *dataset* kedua yang bertujuan memprediksi suatu ulasan termasuk kedalam *dual sentiment* atau *single sentiment*. Dari pengujian ini didapatkan nilai akurasi untuk setiap algoritma yang digunakan sebagai berikut.

Tabel 13. Nilai akurasi pengujian kedua

	Akurasi
SVM	0,8267
NBC	0,7413
Logisttic Regression	0,8204

Hasil prediksi pengujian kedua untuk setiap algoritma yang digunakan dapat dilhat pada tabel-tabel dibawah ini.

Tabel 14. Hasil pengujian kedua menggunakan SVM

Nilai Sebenarnya	**Prediksi**	
	dual	***single***
dual	481	109
single	86	4449

Tabel 15. Hasil pengujian kedua menggunakan NBC

Nilai Sebenarnya	**Prediksi**	
	dual	***single***
dual	420	170
single	121	414

Tabel 16. Hasil pengujian kedua menggunakan *Logistic Regression*

Nilai Sebenarnya	**Prediksi**	
	dual	***single***
dual	480	110
single	92	443

D. Eksperimen dan Hasil

Pada tahapan ini akan diberikan *raw data* berupa ulasan tempat objek wisata yang nantinya akan dilakukan pengujian menggunakan model yang telah dibuat.

Tabel 17. Contoh ulasan tentang objek wisata

	Ulasan
P1	Fasilitas lengkap, tapi sayang kamar mandi kotor dan bau
P2	Tempat favorit saya.. biaya masik juga murah. Cuma 40k perak.
P3	Walau jalannya jauh dari kota yogyakarta... tapi tetap manteeppp... terbayarkan indahnya candi borobudur...
P4	Tempat parkir penuh sesak
P5	Skrg sdh rame jadi kurang bagus lagi
P6	Seru. Banyak spot foto menarik. Hanya akses keluar candi ribet dan buat capek
P7	Serbah mahal..!di negara sendiri aja mahal nya minta ampuun,parkir aja 15.000.apa lagi masuk nya??????????40ribu parah..!!!
P8	Candi adalah Borobudur salah satu dari keajaiban dunia

Selanjutnya data yang ada pada tabel 17 akan melewati tahapan *pre-processing* dan ekstraksi fitur terlebih dahulu sebelum dilakukan prediksi. Algoritma klasifikasi yang digunakan adalah SVM, hal ini dikarenakan SVM memiliki nilai akurasi paling tinggi dibandingkan algoritma lainnya.

Tabel18. Hasil prediksi

	dual sentimen	**single sentimen**		
		positif	**netral**	**negatif**
P1	✓			
P2		✓		
P3	✓			
P4				✓
P5				✓
P6	✓			
P7				✓
P8			✓	

V. KESIMPULAN

Tujuan dari penelitian ini adalah untuk mengindentifikasi sentimen yang ada pada suatu kalimat dan juga dapat menentukan kalimat tersebut termasuk *dual sentiment* atau *single sentiment* menggunakan metode *supervised learning* yaitu *support vector machine, naïve bayes classifier* dan *logistic regression*.

Dari penelitian ini dapat diperoleh kesimpulan bahwa identifikasi dual sentimen pada data berupa teks baik itu kata, maupun kalimat dapat menggunakan salah satu cabang ilmu AI yaitu *natural language processing* (NLP). Berdasarkan pengujian yang telah dilakukan dapat diketahui bahwa algoritma klasifikasi *support vector machine* memiliki kemampuan lebih baik dalam melakukan prediksi sentimen suatu ulasan serta memprediksi kalimat dengan *dual sentiment*. Dalam memprediksi sentimen dalam suatu ulasan algoritma *support vector machine* mendapatkan nilai akurasi sebesar 89%, sedangakan dalam memprediksi kalimat dengan dual sentimen didapat nilai akurasi sebesar 83%.

REFERENSI

[1] Dinas Pariwisata DIY. Buku Statistik Kepariwisataan DIY Tahun 2017.Yogyakarta, 2018.

[2] Basant, A., Namita, M., Pooja, B., Sonal Garg 2. Sentiment Analysis Using Common-Sense and Context Information. Hindawi Publishing Corporation Computational Intelligence and Neuroscience, 2015.

[3] Turney, P.D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (acl) (pp. 417-424). Association for Computational Linguistics, 2002.

[4] G. Scuh, G. Reinhart, J. P. Prote, F. Sauermann, J. Horsthofer, F. Oppolzer, and D. Knoll, Data Mining Definitions and Applications for the Management of Production Complexity. CIRP Conference on Manufacturing Systems, 2019.

[5] Hemalatha, P Saradhi Varma, G. Govardhan, A.Preprocessing the Informal Text for efficient Sentiment Analysis. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol 1 Issue 2,July August 2012.

[6] P. Aliandu, Sentiment Analysis to determine Accommodation, Shopping and Culinary Location on Foursquare in Kupang City. The Third Information Systems International Conference, 2015.

[7] N. A. Vidya, Twitter Sentiment Analysis Terhadap Brand Reputation: Studi Kasus PT XL AXIATA Tbk. 2015.