Makalah Kolokium

by Dimas Bintang Prasetyo

Identifikasi Dual Sentimen Terhadap Ulasan Objek Wisata di Daerah Istimewa Yogyakarta

***Abstraksi*—Saat ini review atau ulasan jadi sangat penting karena bisa jadi sumber informasi dan penilain terhadap suatu objek. Tidak jarang sebuah ulasan dapat merubah pandangan dan keputusan seseorang terhadap objek tersebut. 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*.**

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 senitimen 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. METEDOLOGI

A. Data

Secara umum, data adalah serangkaian karakter yang dikumpulkan dan diterjemahkan untuk suatu tujuan yang memuat suatu informasi. Data bisa berbentuk apa saja, baik itu teks dan angka, gambar, suara, atau video. Dalam penelitian ini data yang akan digunakan adalah kumpulan ulasan terkait objek wisata yang di peroleh dari situs ulasan.

i. Data mining

Proses data mining 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 Scraper*. Dari proses data mining yang telah dilakukan, diperoleh setidaknya 10.000 ulasan. Dari ulasan-ulasan tersebut terdapat 5 objek wisata dengan jumlah ulasan terbanyak yaitu:

Table 1. Objek wisata dengan ulasan terbanyak

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

ii. Dataset

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.

Table 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). Sedangakan label *single* untuk ulasan yang hanya memiliki satu sentiment baik itu positif, negatif, ataupun netral.

Table 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

iii. 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[4]. 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 table 4.

Table 4. 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. Ekstraksi fitur adalah proses pengambilan ciri sebuah objek yang dapat menggambarkan karakteristik dari objek tersebut [5]. 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.

Table 5. 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*. *Supervised learning* sendiri adalah metode *machine learning* yang mengklasifikasikan sebuah data kekelompok tertentu berdasarkan data yang sudah ada sebelumnya. Dengan kata lain, dalam melakukan prediksi *supervised learning* membutuhkan *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. 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.

Table 6. 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 7.

Table 7. 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.

Table 8. Hasil pengujian pertama menggunakan SVM

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

Table 9. Hasil pengujian pertama menggunakan NBC

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

Table 10. 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* sentimen atau *single* sentimen. Dari pengujian ini didapatkan nilai akurasi untuk setiap algoritma yang digunakan sebagai berikut.

Table 11. 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.

Table 12. Hasil pengujian kedua menggunakan SVM

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

Table 13. Hasil pengujian kedua menggunakan NBC

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

Table 14. 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.

Table 15. 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 15 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.

Table16. Hasil prediksi

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

IV. KESIMPULAN

Dalam penelitian ini dapat diperoleh kesimpulan bahwa identifikasi dual sentimen pada data berupa teks baik itu kata, kalimat maupun dokumen dapat menggunakan salah satu cabang ilmu AI yaitu *natural language processing* (NLP). Jenis NLP yang digunakan dalam penelitian ini sendiri adalah *sentiment analysis*. Berdasarkan pengujian yang telah dilakukan dapat diketahui bahwa algoritma klasifikasi *support vector machine* memiliki kemampuan lebih baik dalam melakukan sentimen analisis dibandingkan dengan *naïve bayes classifier* dan *logistic regression*.

REFERENCES

[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] 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.

[5] S. Dony, Mushthofa, Perbandingan Metode Ekstraksi Ciri Histogram dan PCA untuk Mendeteksi Stoma pada Citra Penampang Daun Freycinetia, vol. 2, pp. 20-28, 2013

Makalah Kolokium

ORIGINALITY REPORT

12%	8%	2%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Fan Sun, Ammar Belatreche, Sonya Coleman, T. M. McGinnity, Yuhua Li. "Pre-processing online financial text for sentiment classification: A natural language processing approach", 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), 2014 Publication	2%
2	Submitted to Politeknik Negeri Bandung Student Paper	1%
3	Submitted to Universitas Islam Indonesia Student Paper	1%
4	jurnal.poltekba.ac.id Internet Source	1%
5	www.scribd.com Internet Source	1%
6	media.neliti.com Internet Source	1%
7	www.telefonanlage-shop.de	

	Internet Source	1%
8	link.springer.com Internet Source	1%
9	kasmudody.blogspot.com Internet Source	1%
10	www.slideshare.net Internet Source	<1%
11	pt.scribd.com Internet Source	<1%
12	Submitted to Sriwijaya University Student Paper	<1%
13	chalvjr.blogspot.com Internet Source	<1%
14	mafiadoc.com Internet Source	<1%
15	Doaa Mohey El-Din Mohamed Hussein. "A survey on sentiment analysis challenges", Journal of King Saud University - Engineering Sciences, 2016 Publication	<1%
16	Submitted to Sekolah Tinggi Pariwisata Bandung Student Paper	<1%

17	www.muslimsources.com Internet Source	<1%
18	Submitted to Universitas Negeri Jakarta Student Paper	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		