JPPI Vol 7 No 2 (2017) 109-120

Jurnal Penelitian Pos dan Informatika

771/AU1/P2MI-LIPI/08/2017
32a/E/KPT/2017

e-ISSN 2476-9266
p-ISSN: 2088-9402
DOI: 10.17933/jppi.2017.070203

PEMANFAATAN ANALISIS SENTIMEN UNTUK PEMERINGKATAN POPULARITAS TUJUAN WISATA

UTILIZATION OF SENTIMENT ANALYSIS FOR TOURIST DESTINATION POPULARITY RANKING

Murnawan[1], Ardiles Sinaga[2]

[1]Program Studi Sistem Informasi, Fakultas Teknik – Universitas Widyatama
[2]Program Studi Teknik Informatika, Fakultas Teknik – Universitas Widyatama
[1]murnawan@widyatama.ac.id

Naskah Diterima: 29 September 2017; Direvisi : 10 Desember 2017; Disetujui : 10 Desember 2017

Abstrak

Saat ini, hampir seluruh industri tidak terkecuali industri pariwisata pasti bersentuhan dengan teknologi informasi. Peranan teknologi sangat erat kaitannya dalam meningkatkan industri pariwisata di Indonesia. Selama berwisata, wisatawan juga dapat membagikan pengalamannya melalui upload foto maupun aktif berkomentar baik di media sosial maupun di forum diskusi yang khusus membahas tentang pariwisata. Informasi seputar kondisi pariwisata, pengalaman wisatawan, opini serta foto yang di upload dari tempat wisata, dapat diolah menjadi suatu informasi yang sangat bermanfaat, salah satunya adalah dapat dimanfaatkan untuk pemeringkatan popularitas. Untuk itu diperlukan suatu sistem yang dapat memberikan informasi tentang popularitas tujuan wisata. Dalam penelitian ini rancangan sistem yang dibuat ini dapat menentukan peringkat tujuan wisata yang paling populer dengan memanfaatkan analisis sentimen. Ada lima komponen penilaian untuk menghasilkan nilai akhir dari peringkat popularitas yaitu *comment count*, Facebook *likes count*, Facebook *were here count*, Facebook *talking about* dan Instagram *visitor*. Pada penelitian ini, peneliti memutuskan untuk menggunakan strategi klasifikasi berdasarkan pada algorima *naïve bayes* (NB) karena merupakan suatu metode yang sederhana dan intuitif yang kinerjanya mirip dengan pendekatan lain. Selain itu, berdasarkan penelitian-penelitian terdahulu, NB menggabungkan efisiensi waktu kinerja yang optimal yang cukup akurat.

Kata kunci: analisis sentimen, popularitas, pariwisata, tujuan wisata, Naïve Bayes

Abstract

Today, almost all industries are no exception to the tourism industry inevitably in contact with information technology. The role of technology in improving the tourism industry in Indonesia is very closely related. During the tour, tourists can also share their experiences through photo uploads and active comments both in social media and in discussion forums that specifically discuss tourism. Information about the condition of tourism, tourist experiences, opinions and photos uploaded from tourist attractions, can be processed into a very useful information, one of which is to be utilized for popularity ranking. For that, we need a system that can provide information about the popularity of tourist destinations. In this study, a system designed to determine the ranking of the most popular tourist destinations by utilizing sentiment analysis. There are five components of the assessment to produce the final value of the popularity ranking of comment count, Facebook likes count, Facebook was here count, Facebook talking about and Instagram visitor. In this study, researchers decided to use a classification strategy based on the algorithm naïve Bayes (NB) because it is a simple and intuitive method whose performance is similar to other approaches. Also, based on previous studies, NB combines optimum performance time efficiency that is quite accurate.

Keywords*: sentiment analysis, popularity, tourism, tourist destination, Naïve Bayes*

PENDAHULUAN

Saat ini pertumbuhan pengguna aktif di media sosial mengalami perkembangan yang sangat pesat. Menurut data terbaru dari halaman web *We Are Social*, pengguna internet aktif berdasarkan *platform* media sosial di seluruh dunia kini mencapai angka 9,1 miliar. Dari tahun ke tahun, jumlah pengguna internet terus bertumbuh. Menurut laporan yang sama, pengguna media sosial aktif kini mencapai 4,5 miliar, sedangkan pengguna mobile mencapai 4,7 miliar. Sementara itu, *Facebook* masih menjadi media sosial yang paling banyak digunakan dengan angka hampir mencapai 1,6 miliar pengguna.

Dampak positif media sosial dalam perkembangan *information technology* (IT) sebenarnya membawa banyak keuntungan, misalnya saja memudahkan dalam hal komunikasi, mencari dan mengakses informasi. Salah satu cabang riset yang kemudian berkembang dari situasi ledakan informasi di internet adalah *sentiment analysis*. *Sentiment analysis* atau sering disebut juga *opinion mining* adalah studi komputasional dari opini-opini orang, *appraisal* dan emosi melalui entitas, event dan atribut yang dimiliki (Liu, 2012).

Adapun penelitian-penelitian terdahulu yang terkait dengan *sentiment analysis*, antara lain adalah penelitian (Gamallo, Pablo, & Fernandez-Lanza, 2013) tentang strategy naïve bayes untuk *sentiment analysis* untuk tweet berbahasa Spanyol. Penelitian (Pak, Alexander and Paroubek, 2010) menganalisis tweeter sebagai linguistik korpus untuk *sentiment analysis*. Penelitian (Vinodhini and Chandrasekaran, 2012) mengembangkan sistem yang dapat mengidentifikasi dan mengklasifikasikan sentimen masyarakat untuk memprediksi produk yang menarik dalam pemasaran.

Pada penelitian ini, *sentiment analysis* akan digunakan untuk menentukan peringkat popularitas tujuan pariwisata berdasarkan data komentar atau opini yang terdapat di forum diskusi serta di media sosial yang spesifik membicarakan tentang suatu tempat pariwisata. Pada penelitian ini juga, peneliti memutuskan untuk menggunakan strategi klasifikasi berdasarkan pada algorima Naif Bayes (NB) karena merupakan suatu metode yang sederhana dan merupakan dan intuitif yang kinerjanya mirip dengan pendekatan lain. Selain itu, berdasarkan penelitian-penelitian terdahulu, NB menggabungkan efisiensi (waktu kinerja yang optimal) dengan cukup akurat.

Agar pembahasan lebih terarah dan sesuai dengan apa yang diharapkan maka pada penelitian ini yang menjadi batasan masalah adalah:

1. Sebagai uji coba aplikasi, peneliti menetapkan *point of interest* (POI) atau spesifik lokasi dari tujuan wisata adalah sebanyak 50 POI di pulau Bali.
2. Data yang dianalisis hanya dalam format bahasa Indonesia dan dengan kata kunci tentang tujuan wisata yang berhubungan dengan POI.
3. Metode klasifikasi yang digunakan untuk mengklasifikasikan komentar-komentar adalah menggunakan *Naïve Bayes Classifier*.
4. Sebagai sumber data, informasi di *crawling* dari beberapa forum diskusi dan dari media sosial *Facebook* dan *Instagram*.

Adapun tujuan dari penelitian ini adalah membangun suatu aplikasi berbasis web yang dapat menghasilkan informasi tentang peringkat tempat pariwisata yang paling populer di Indonesia

berdasarkan hasil *sentiment analysis* terhadap komentar maupun foto yang terdapat pada media sosial.

METODE

Konsep Sentiment Analysis

Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek dan menentukan apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif, negatif atau netral. Lebih lanjut *sentiment analysis* dapat menyatakan emosional sedih, gembira, atau marah (Liu, 2012).

Ekspresi atau sentiment mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subject yang berbeda. Sebagai contoh, adalah hal yang baik untuk mengatakan alur film tidak terprediksi, tapi adalah hal yang tidak baik jika ‘tidak terprediksi' dinyatakan pada kemudi dari kendaraan. Bahkan pada produk tertentu, kata-kata yang sama dapat menggambarkan makna kebalikan, contoh adalah hal yang buruk untuk waktu *start-up* pada kamera digital jika dinyatakan “lama”, namun jika” lama” dinyatakan pada usia batere maka akan menjadi hal positif.

Hal pertama dalam pemrosesan dokumen adalah memecah kumpulan karakter ke dalam kata atau token, sering disebut sebagai tokenisasi. Tokenisasi adalah hal yang kompleks untuk program komputer karena beberapa karakter dapat dapat ditemukan sebagai token delimiters. Delimiter adalah karakter spasi, tab dan baris baru (*newline*), sedangkan karakter ( ) <> ! ? “ kadang dijadikan delimiter namun, kadang juga bukan, tergantung pada lingkungannya (Triawati, 2009).

Konsep Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban & Liang, 2005).

Istilah data mining memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. Data mining, sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Larose, 2005).

Metode Pelatihan Data

Secara garis besar metode pelatihan yang digunakan dalam teknik-teknik data mining dibedakan ke dalam dua pendekatan, yaitu (Santosa, 2007) :

1. *Unsupervised learning*, metode ini diterapkan tanpa adanya latihan (training) dan tanpa ada guru (*teacher*). Guru di sini adalah label dari data.
2. *Supervised learning*, yaitu metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi, digunakan beberapa contoh data yang mempunyai output atau label selama proses training.

Pengelompokan Data Mining

Ada beberapa teknik yang dimiliki data mining berdasarkan tugas yang bisa dilakukan, yaitu (Larose, 2005):

1. Deskripsi

   Para peneliti biasanya mencoba menemukan cara untuk mendeskripsikan pola dan *trend* yang tersembunyi dalam data.
2. Estimasi

   Estimasi mirip dengan klasifikasi, kecuali variabel tujuan yang lebih kearah numerik dari pada kategori.
3. Prediksi

   Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi di masa depan).
4. Klasifikasi

   Dalam klasifikasi variabel, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.
5. *Clustering*

   Clustering lebih ke arah pengelompokan *record,* pengamatan, atau kasus dalam kelas yang memiliki kemiripan.
6. Asosiasi

   Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

Konsep *Naïve Bayes Classifier (NBC)*

Naïve Bayes classifier adalah *classifier* probabilistik sederhana berdasarkan penerapan teorema Bayes (dari statistik Bayesian) dengan asumsi independen (naif) yang kuat. Sebuah istilah yang lebih deskriptif untuk model probabilitas yang digaris bawahi adalah "model fitur independen".

Dalam terminologi sederhana, sebuah NBC mengasumsikan bahwa kehadiran (atau ketiadaan) fitur tertentu dari suatu kelas tidak berhubungan dengan kehadiran (atau ketiadaan) fitur lainnya. Sebagai contoh, buah mungkin dianggap apel jika merah, bulat, dan berdiameter sekitar 4 inchi. Bahkan jika fitur ini bergantung satu sama lain atau atas keberadaan fitur lain. Sebuah NBC menganggap bahwa seluruh sifat-sifat berkontribusi mandiri untuk probabilitas bahwa buah ini adalah apel.

Tergantung pada situasi yang tepat dari model probabilitas, NBC dapat dilatih sangat efisien dalam *supervised learning*. Dalam aplikasi praktis, parameter estimasi untuk model NBC menggunakan metode *likely hood* maksimum, dengan kata lain, seseorang dapat bekerja dengan model *Naïve Bayes* tanpa mempercayai probabilitas Bayesian atau menggunakan metode Bayesian lainnya (Sumartini Saraswati, 2011).

Dibalik desain naifnya dan asumsi yang tampaknya terlalu disederhanakan, NBC telah bekerja cukup baik dalam banyak situasi dunia nyata yang kompleks. Pada tahun 2004, analisis masalah klasifikasi Bayesian telah menunjukkan bahwa ada beberapa alasan teoritis untuk keberhasilan yang tampaknya tidak masuk akal dari NBC. Selain itu, perbandingan yang komprehensif dengan metode klasifikasi lainnya pada tahun 2006 menunjukkan bahwa klasifikasi Bayes mengungguli pendekatan terbaru, seperti *boosted tree* atau *random forest.*

Sebuah keuntungan dari NBC adalah hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata-rata dan varian dari variabel) yang diperlukan untuk klasifikasi. Karena variabel diasumsikan independen, hanya varian dari variabel untuk setiap kelas yang perlu ditentukan dan bukan keseluruhan *covariance matrix.*

Pang, B. and Lee (2008), menyatakan bahwa salah satu *core* dalam *sentiment analysis* adalah *problem* klasifikasi opini. Untuk konteks klasifikasi dokumen teks secara umum (bukan teks opini), misalnya teks berita, metode NBC telah diterapkan oleh beberapa peneliti. Menurut Liu (2012) *term sentiment analysis* sering digunakan di dunia industri (misalnya *review* produk untuk mengetahui sentimen pasar). Selanjutnya Liu mendefinisikan opini dalam suatu dokumen sebagai *quantuple* :

$$(e_j, a_{jk}, s_{ijkl}, h_i, t_l) \quad (1)$$

dengan makna simbol:

- e_j adalah entitas target opini
- a_{jk} adalah aspek /feature dari entitas ej
- s_{ijkl} adalah nilai sentimen dari pemilik opini (h_i) pada aspek ajk dari entitas ej pada waktu t_l(opini positif misalnya "bagus", "tepat waktu" ; opini negatif misalnya ("terlambat", "parah")
- h_i adalah pemilik opini
- t_l adalah waktu kapan opini dikeluarkan

Jika opini, pesan atau komentar dianggap sebagai dokumen d, dan diasumsikan dimiliki koleksi dokumen D = $\{d_i|i=1,2,\ldots|D|\} = \{d_1,d_2,\ldots,d_{|D|}\}$ dan koleksi kategori V = $\{v_j|j=1,2,\ldots|V|\} = \{v_1,v_2,\ldots,v_{|V|}\}$. Klasifikasi NBC dilakukan dengan cara mencari probabilitas $P(V=v_j \mid D=d_i)$, yaitu probabilitas kategori v_j jika diketahui dokumen d_i. Dokumen d_i dipandang sebagai *tuple* dari kata-kata dalam dokumen, yaitu $<a_1, a_2,\ldots,a_n>$, yang frekuensi kemunculannya diasumsikan sebagai variable random dengan distribusi probabilitas Bernoulli (McCallum and Nigam, 1998). Selanjutnya klasifikasi dokumen adalah mencari nilai maksimum dari :

$$V_{MAP} = arg_{vj \in V} max\, P(a_1, a_2, \ldots, a_n \mid v_j) \quad (2)$$

Dengan menerapkan teorema *Bayes* persamaan (2) dapat ditulis :

$$V_{MAP} = arg_{vj \in V} max \frac{P(a_1,a_2,\ldots,a_n \mid v_j) P(v_j)}{P(a_1,a_2,\ldots,a_n \mid v_j)} \quad (3)$$

Karena nilai $P(a_1,a_2,\ldots,a_n)$ untuk semua v_j besarnya sama maka nilainya dapat diabaikan, sehingga persamaan (3) menjadi :

$$V_{MAP} = arg_{vj \in V} max\, P(a_1, a_2, \ldots, a_n|v_j)\, P(v_j) \quad (4)$$

Dengan mengasumsikan bahwa setiap kata dalam $<a_1, a_2,\ldots,a_n>$ adalah *independent*, maka $P(a_1,a_2,\ldots,a_n \mid v_j)$ dalam persamaan (4) dapat ditulis sebagai :

$$P(v_j|a_1, a_2, \ldots, a_n) = \prod_i P(a_i|v_j) \quad (5)$$

Sehingga persamaan (4) dapat ditulis :

$$V_{MAP} = \arg \max_{vj \in V} P(v_j) \prod_i P(a_i|v_j) \quad (6)$$

Nilai $P(v_j)$ ditentukan pada saat pelatihan, yang nilainya didekati dengan :

$$P(v_j) = \frac{|doc_j|}{|Contoh|} \quad (7)$$

dimana $|doc_j|$ adalah banyaknya dokumen yang memiliki kategori j dalam pelatihan, sedangkan |*Contoh*| banyaknya dokumen dalam contoh yang digunakan untuk pelatihan.

Untuk nilai $P(w_k| v_j)$, yaitu probabilitas kata wk dalam kategori j ditentukan dengan :

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |vocabulary|} \quad (8)$$

Dimana n_k adalah frekuensi munculnya kata w_k dalam dokumen yang ber kategori v_j , sedangkan nilai n adalah banyaknya seluruh kata dalam dokumen berkategori v_j, dan |*vocabulary*| adalah banyaknya kata dalam contoh pelatihan.

Naive bayes classifier (NBC) membutuhkan jumlah *record* data yang sangat besar untuk mendapatkan hasil yang baik. Jika kategori

prediktor tidak ada dalam data training, maka *naive bayes classifier* mengasumsikan bahwa record baru dengan kategori *predictor* memiliki probabilitas nol.

Cara Kerja Algoritma *Naive Bayes*

Ada dua tahapan yang dilakukan dalam implementasi algoritma *Naïve Bayes*, yaitu tahapan pelatihan dan tahapan klasifikasi.

1. Pelatihan

   Naïve Bayes merupakan algoritma yang termasuk dalam kategori *supervised learning*, sehingga dibutuhkan pengetahuan awal untuk dapat mengambil keputusan. Adapaun langkah-langkah yang dilakukan adalah:

   a. Bentuk *vocabulary* pada setiap dokumen data training.
   b. Hitung probabilitas pada setiap kategori P(vj).
   c. Tentukan frekuensi setiap kata wk pada setiap kategori P(wk | vj).

2. Klasifikasi

   Adapun langkah-langkah dalam melakukan klasifikasi adalah:

   a. Hitung P(vj)Π P(wk | vj) untuk setiap kategori.
   b. Tentukan kategori dengan nilai P(vj)Π P(wk | vj) maksimal.

Pengumpulan Data

Secara garis besar, ada dua tahapan dalam proses pengumpulan data yaitu pengumpulan data tujuan wisata serta pengumpulan data sumber melalui proses *crawling*.

Pengumpulan Data Tujuan Wisata

Langkah awal yang dilakukan adalah mengumpulkan data *point of interest* (POI) dari tujuan wisata yang akan dirangking. Pada penelitian ini ditetapkan 50 POI untuk tujuan wisata di pulau Bali berikut dengan koordinat latitude dan longitudenya. Sebagai contoh, data tersebut dapat dilihat pada Tabel 1.

Tabel 1. Contoh Data POI

Point of Interest	Koordinat
Kuta Theater	-8.729992,115.168106
Ground Zero	-8.717369,115.174468
Bali Shell Museum	-8.714265,115.186778
Pura Uluwatu	-8.817124,115.08854
Pantai Dreamland	-8.817124,115.08854
Tanjung Benoa	-8.757806,115.219992
Pantai Legian dan Seminyak	-8.780953,115.162895
Pantai Jimbaran	-8.723639,115.169588
Garuda Wisnu Kencana Cultural Park	-8.806919,115.164052
Pantai Kuta	-8.722314,115.169591

Pengumpulan Data Source

Pengumpulan data source dilakukan melalui proses *crawling* secara online dari beberapa forum diskusi yang telah ditetapkan, dari Facebook dan Instagram.

Ada lima area komponen yang akan menjadi perhitungan perangkingan peringkat popularitas tujuan pariwisata, yaitu:

a. *Comment Count*

   Merupakan kumpulan dari komentar yang ada di forum diskusi di Indonesia yang khusus membahas tentang satu tempat pariwisata.

   Tujuannya adalah untuk melihat apakah satu tempat cukup populer untuk dibahas dalam suatu forum diskusi, dan juga apakah komentar tersebut bermakna positif atau negatif.

b. Instagram *Visitor*

   Merupakan kumpulan data dari pengunggah foto pada *website* Instagram, dimana aplikasi Instagram telah menyediakan koordinat

latitude dan longitude dalam meta foto yang diunggah melalui telepon seluler ke *website* mereka.

Tujuannya untuk mengetahui berapa banyak orang yang mengambil foto di tempat pariwisata berdasarkan latitude dan longitude yang disediakan oleh Instagram.

c. Facebook *Likes Count*

Merupakan jumlah *like* dari *location* yang tersedia di laman Facebook

Tujuannya untuk mendapatkan berapa banyak orang yang menyukai suatu tempat pariwisata tertentu secara massal.

d. Facebook *Were Here Count*

Merupakan jumlah "*were here count*" untuk suatu tempat pariwisata.

Tujuannya untuk mendapatkan jumlah banyaknya orang yang pernah mengunjungi suatu tempat pariwisata berdasarkan update status Facebook dari tempat pariwisata tersebut.

e. **Facebook *Talking About***

Merupakan jumlah orang yang membicarakan suatu tempat pariwisata pada status Facebook mereka.

Tujuannya untuk mendapatkan berapa banyak orang yang melakukan *mention* tempat pariwisata.

Proses *Text Preprocessing*

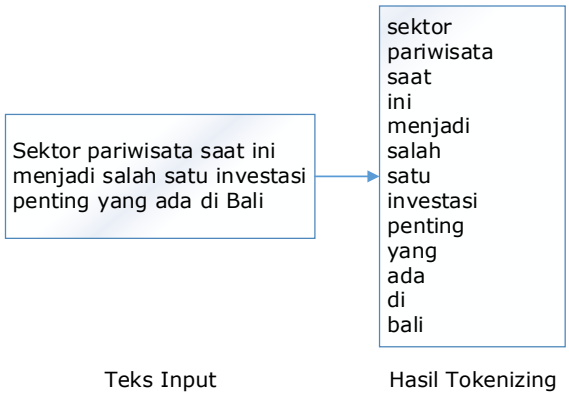
Proses *text prepocessing* merupakan implementasi dari *text mining* untuk memproses suatu teks yang tidak terstruktur menjadi lebih tersruktur atau dengan kata lain mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Ada empat tahapan *text prepocessing* yang meliputi *case folding*, *tokenizing*, *filtering*, dan *stemming*.

1. *Case folding*

Tahapan *case folding* adalah tahapan untuk mengkonversi keseluruhan teks dalam dokumen menjadi seluruhnya huruf kecil (*lowercase*). Sebagai contoh, teks "PARIWISATA", "Pariwisata", "PariWisata", atau "pariwisata", tetap diberikan hasil retrieval yang sama yakni "pariwisata". Pada tahapan *case folding* ini hanya huruf 'a' sampai dengan 'z' yang diterima, sedangkan karakter selain huruf dihilangkan dan dianggap *delimiter*.

2. *Tokenizing*

Tahapan *tokenizing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Contoh dari tahap ini dapat dilihat pada Gambar 1.



Gambar 1. Proses Tokenizing

Tokenisasi secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata, bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan.

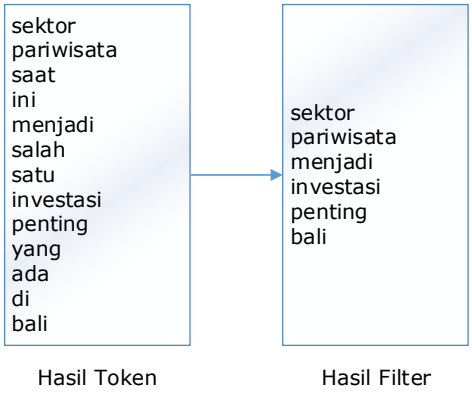
Sebagai contoh karakter *whitespace*, seperti enter, tabulasi, spasi dianggap sebagai pemisah kata. Namun untuk karakter petik tunggal ('), titik (.), semikolon (;), titk dua (:) atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata.

Dalam memperlakukan karakter-karakter dalam teks sangat tergantung pada kontek aplikasi yang dikembangkan. Pekerjaan tokenisasi ini akan semakin sulit jika juga harus memperhatikan struktur bahasa (grammatikal).

3. *Filtering*

Tahapan *filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist*/*stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contohnya dapat dilihat pada Gambar 2.

Contoh *stopword* s adalah "yang", "dan", "di", "dari" dan seterusnya (TALA, 2003).



Gambar 2. Proses *Filtering*

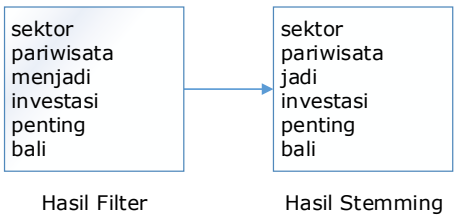
Kata-kata seperti "dari", "yang", "di", dan "ke" adalah beberapa contoh kata-kata yang berfrekuensi tinggi dan dapat ditemukan hampir dalam setiap dokumen (disebut sebagai *stopword*). Penghilangan *stopword* ini dapat mengurangi ukuran index dan waktu pemrosesan. Selain itu, juga dapat mengurangi level *noise*.

Namun terkadang *stopping* tidak selalu meningkatkan nilai retrieval. Pembangunan daftar *stopword* (disebut *stoplist*) yang kurang hati-hati dapat memperburuk kinerja sistem *Information Retrieval* (IR). Belum ada suatu kesimpulan pasti bahwa penggunaan stopping akan selalu meningkatkan nilai retrieval, karena pada beberapa penelitian, hasil yang didapatkan cenderung bervariasi.

4. *Stemming*

Pembuatan indeks dilakukan karena suatu dokumen tidak dapat dikenali langsung oleh suatu sistem temu kembali informasi atau *Information Retrieval System* (IRS). Oleh karena itu, dokumen tersebut terlebih dahulu perlu dipetakan ke dalam suatu representasi dengan menggunakan teks yang berada di dalamnya.

Teknik s*temming* diperlukan selain untuk memperkecil jumlah indeks yang berbeda dari suatu dokumen, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau form yang berbeda karena mendapatkan imbuhan yang berbeda.



Gambar 3. Proses *Stemming*

Sebagai contoh kata bersama, kebersamaan, menyamai, akan distem ke *root word*-nya yaitu "sama". Namun, seperti halnya *stopping*, kinerja *stemming* juga bervariasi dan sering tergantung pada domain bahasa yang digunakan.

Proses *stemming* pada teks berbahasa Indonesia berbeda dengan *stemming* pada teks berbahasa Inggris. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses

menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia semua kata imbuhan baik itu sufiks dan prefiks juga dihilangkan.

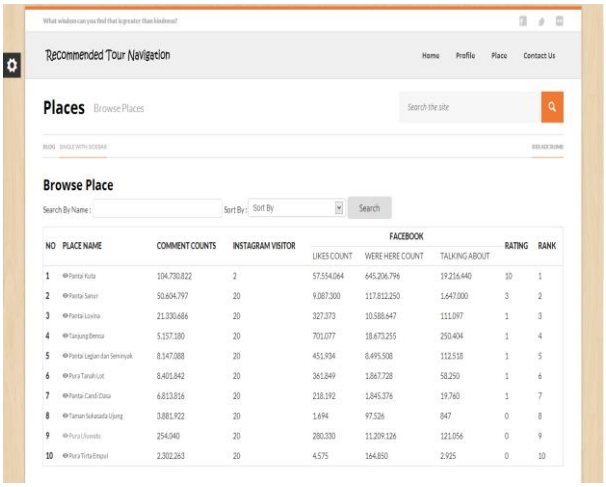
Proses Klasifikasi *Naïve Bayes*

Tujuan dari proses klasifikasi adalah untuk menentukan sebuah kalimat apakah termasuk sebagai anggota kelas opini positif atau sebagai anggota kelas opini negatif yang ditentukan berdasarkan nilai perhitungan probabilitas Bayes yang lebih besar. Jika hasil probabilitas Bayes kalimat tersebut untuk kelas opini positif lebih besar maka kalimat tersebut masuk kategori opini positif demikian juga sebaliknya.

PEMBAHASAN DAN HASIL

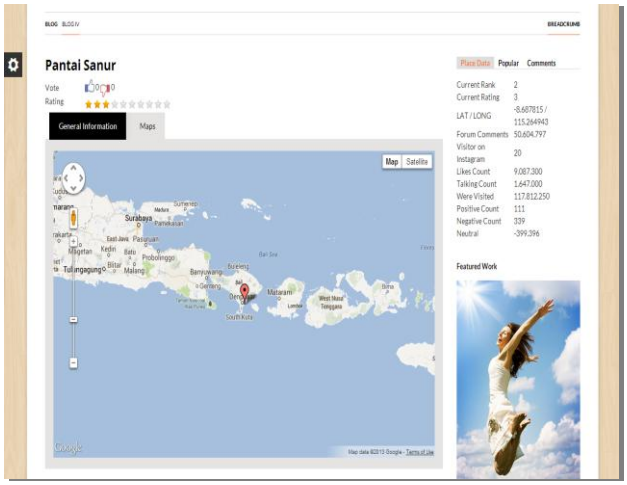
Implementasi Pada *Platform Web*

Untuk menghasilkan informasi popularitas tujuan wisata, hasil *sentiment analysis* diimplemetasikan secara *web based* yang dapat dilihat secara publik di alamat http://poi.iniristi.com. Ada dua halaman utama yang paling penting, yaitu halaman informasi peringkat seperti terlihat pada Gambar 6, serta informasi detail dari tujuan wisata seperti terlihat pada Gambar 7.



Gambar 6. Peringkat Popularitas Tujuan Wisata

Dari halaman seperti yang terlihat pada gambar 6, kemudian dapat dipilih salah satu tempat tujuan wisata jika ingin melihat detail informasi seperti yang terlihat pada Gambar 7.



Gambar 7. Detail Informasi

Adapun detail informasi yang ditampilkan adalah posisi tujuan wisata yang terletak pada peta (*map*) serta informasi dari hasil *crawling* dan *sentiment analysis*. Berikut ini data yang ditampilkan:

- *Current Rank* : 2
- *Current Rating* : 3
- LAT / LONG : -8.687815 / 115.264943
- *Forum Comments* : 50.604.797
- *Visitor on* Instagram : 20
- *Likes Count* : 9.087.300
- *Talking Count* : 1.647.000
- Were Visited : 117.812.250
- *Positive Count* : 111
- *Negative Count* : 339
- *Neutral* : -399.396

Metode Pemeringkatan

Metode pemeringkatan secara umum yang dilakukan masih sangat sederhana, dimana tidak ada pembobotan dari masing-masing komponen penilaian.

a. Skala yang digunakan: 0 – 10
b. Masing-masing komponen penilaian dihitung persentase nilai positif dari total data training yang ada.
c. Rating dihitung berdarkan jumlah seluruh data training dibagi jumlah seluruh persentase nilai positif.
d. Kemudian dibuat perangkingannya.

Hasil Pemeringkatan

Berdasarkan hasil *crawling* data yang dilakukan terhadap 50 POI, maka untuk sepuluh besar urutan teratas dapat dilihat pada tabel 2.

Tabel 2. Sepuluh Peringkat Tujuan Wisata Terbaik

Rank	Place	CC	IV	Facebook		
				LC	WH	TA
1	Pantai Kuta	104.73	2	57.554	645.206	19.216
2	Pantai Sanur	50.604	20	9.087	117.812	1.647
3	Pantai Lovina	21.33	20	327	10.588	111
4	Tanjung Benoa	5.157	20	701	18.673	250
5	Pantai Legian dan Seminyak	8.147	20	451	8.495	112
6	Pura Tanah Lot	8.401	20	361	1.867	58
7	Pantai Candi Dasa	6.813	20	218	1.845	19
8	Taman Sukasada Ujung	3.881	20	1	97	8
9	Pura Uluwatu	254	20	280	11.209	121
10	Pura Tirta Empul	2.302	20	4	164	2

Keterangan tabel:

- Data dalam ribu.
- *Rank* = Peringkat tujuan wisata
- *Place* = Tempat tujuan wisata (POI)
- CC = *Comment Count*
- IV = Instagram *Visitor*
- LC = *Like Count*
- WH = *Where Here Count*
- TA = *Talking About*

Secara keseluruhan, hasil pemeringkatan dapat dilihat pada alamat http://poi.iniristi.com.

Hasil Analisis Validitas Data

Berdasarkan data-data yang dikumpulkan dari beberapa sumber hasil *crawling*, yaitu dari Facebook, Instagram dan forum diskusi, data konten yang dihasilkan secara jumlah dapat secara langsung dapat dilihat tempat mana yang paling populer, tanpa harus memakai metode yang khusus. Berikut ini salah satu contoh yang didapatkan dari Facebook:

1. Pantai Kuta :
   a. Facebook *Likes Count* : 57.554.064
   b. *Were Here Count* : 645.206.796
   c. *Talking About Count* : 19.216.440
2. Pantai Sanur :
   a. Facebook *Likes Count* : 9.087.300
   b. *Were Here Count* : 117.812.250
   c. *Talking About Count* : 1.647.000

Dan seperti yang telah dijelaskan sebelumnya, masing masing komponen penilaian memiliki peranannya masing-masing. Khusus untuk komponen penilaian *Were Here Count*, dimana data langsung didapatkan dari Facebook berdasarkan update status dari setiap orang yang kemudian secara otomatis sistem Facebook menetapkan koordinat *latitude* dan *longitude* dari posisi orang tersebut pada saat melakukan update statusnya. Jadi secara tidak langsung, bisa diketahui berapa banyak orang yang pernah mengunjungi POI tersebut.

Bagaimanapun, data dari satu sumber saja tidak dapat langsung membuat sebuah POI menduduki peringkat teratas. Seperti contoh dibawah ini :

1. Taman Sukasada Ujung :
   a. Facebook *Likes Count* : 1.694
   b. *Were Here Count* : 97.526
   c. *Talking About Count* : 847
   d. *Comment Count* : 3.881.922

2. Pura Uluwatu :
   a. Facebook *Likes Count* : 280.330
   b. *Were Here Count* : 11.209.126
   c. *Talking About Count* : 121.056
   d. *Comment Count* : 254.040

Berdasarkan contoh diatas, jika menggunakan data yang diambil dari Facebook saja, sudah jelas bahwa Pura Uluwatu seharusnya berada jauh diatas Taman Sukasada Ujung. Tetapi dengan adanya komponen penilaian *Comment Count* maka hasil perhitungan akhir mengalami perubahan. Taman Sukasada Ujung memiliki *Comment Count* yang jauh lebih banyak dibanding Pura Uluwatu. Ini menyatakan bahwa banyak orang yang antusias membahas Taman Sukasada Ujung dibanding Pura Uluwatu.

Algoritma *sentiment analyst* yang digunakan pada penelitian ini, yaitu algoritma yang mengharuskan penggunanya untuk mengumpulkan frase sebanyak-banyaknya, sehingga keakurasian sentiment yang didapat bisa menghasilkan keakuratan yang semaksimal mungkin. Dari algoritma yang digunakan, berdasarkan hasil uji coba yang dilakukan ada beberapa data tingkat keakuratan *sentiment* yang didapat:

a. 100 Phrase,
   Accuracy : 0.65653495440729 (65.65%)
b. 5000 Phrase,
   Accuracy : 0.82674772036474 (82.67%)

Dari data diatas dapat lihat bahwa semakin banyak frase yang dimiliki sebagai *core* dari algoritma tersebut, maka semakin akurat *sentiment analysis* yang disajikan.

Perbandingan Rating Secara Umum

Berdasarkan kenyataan yang ada, diketahui bahwa POI yang paling popular di Pulau Bali merupakan Pantai Kuta Bali. Banyak yang berkunjung ke Bali yang mengatakan bahwa "Bukan ke Bali kalau tidak berkunjung ke Pantai Kuta", dan slogan tersebut sangat popular dikalangan para wisatawan.

Sebagai perbandingan, hasil pada penelitian ini dicoba dibandingkan dengan salah satu situs *web* pariwisata yang cukup terkenal yaitu Tripadvisor. Pada Tripadvisor, Pantai Kuta lebih popular dibanding Pantai Sanur. Sedangkan pada hasil penelitian yang kami lakukan, peringkat tersebut sering berganti-ganti. Hal ini bisa disebabkan karena masalah konsistensi proses *crawling* data serta jumlah data dan sumber data yang berbeda. Selain itu, perbedaan metodologi perhitungan pemeringkatan yang digunakan juga sangat mempengaruhi hasil pemeringkatan.

PENUTUP

Berdasarkan hasil analisis uji validitas data, algoritma yang digunakan sudah cukup memiliki akurasi yang baik dan dapat diterima sebagai suatu metode perhitungan pemeringkatan.

Khususnya dalam domain pariwisata, komponen-komponen penilaian yang digunakan sudah dapat memberikan penilaian yang lebih objektif yang dihasilkan berdasarkan konten opini atau ulasan dari media sosial maupun situs pariwisata lainnya. Dengan mengambil data uji terhadap 50 *point of interest* (POI) tujuan wisata di Pulau Bali, hasil ekstraksi dan klasifikasi opini maupun ulasan yang berkitan dengan destinasi wisata di pulau Bali tersebut menghasilkan peringkat popularitas yang dapat menjadi bahan pertimbangan bagi wisatawan sebagai tujuan wisata mereka.

Ada beberapa saran agar hasil penelitian ini dapat lebih baik lagi, yaitu:

1. Penelitian selanjutnya dapat dikembangkan dengan menambah beberapa fitur pada aplikasi
2. Perlu ada perbaikan dalam metodologi pemeringkatan yang digunakan, agar hasil pemeringkatan dapat lebih akurat dan lebih baik lagi.
3. Menambah metodologi *business intelegent* (BI) yang dapat digunakan sebagai alat bantu pengambilan keputusan bagi yang berkepentingan, dalam hal ini khususnya pemerintah sebagai *stakeholder* destinasi wisata.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada pihak-pihak telah membantu untuk menyelesaikan penelitian ini.

DAFTAR PUSTAKA

Gamallo, Pablo, & Fernandez-Lanza. (2013). TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. In *In Workshop on Sentiment Analysis at SEPLN (TASS2013)* (pp. 126–132). Madrid, Spain.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* John Willey & Sons, Inc.

Liu, B. (2012). *Opinion Mining*. Chicago, United States of America.

Pak, Alexander and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *In LREC-2010*. Valletta, Malta.

Pang, B. and Lee, L. (2008). *Opinion mining and sentiment analysis in Foundations and Trends in Information Retrieval*.

Santosa, B. (2007). *Data mining teknik pemanfaatan data untuk keperluan bisnis.* Yogyakarta: Graha Ilmu.

Sumartini Saraswati, N. W. (2011). *Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis*. Denpasar, Bali, Indonesia.

TALA, F. Z. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Universiteit van Amsterdam.

Triawati, C. (2009). *Text Mining*. Bandung, Jawa Barat, Indonesia.

Turban, E., & Liang, J. E. A. and T. P. (2005). Decision Support System and Intelligent Systems (7th ed). Pearson Education, Inc.