IMPLEMENTASI ANALISIS CLUSTERING DAN SENTIMEN DATA TWITTER PADA OPINI WISATA PANTAI MENGGUNAKAN METODE K-MEANS

Yan Watequlis Syaifudin[1], Rizki Andi Irawan[2]

Jurusan Teknologi Informasi, Program Studi Teknik Informatika, Politeknik Negeri Malang
[1]qulis@polinema.ac.id, [2]rizki4ndi@gmail.com

Abstrak

Analisis sentimen atau opinion mining merupakan topik riset dari cabang penelitian text mining. Fokus dari analisis sentimen adalah melakukan analisis opini dari suatu dokumen teks, sehingga membantu usaha untuk melakukan riset pasar atas opini publik. Data opini diperoleh dari jejaring sosial Twitter dalam Bahasa Indonesia dengan topik suatu pantai.Klasifikasi opini diperlukan untuk memudahkan pengguna dalam melihat opini positif, negatif, ataupun netral. Algoritma yang digunakan dalam klasifikasi adalah *Support Vector Machine*. Pada penelitian digunakan dataset dari 10 pantai yang ada di Indonesia sebanyak 500 tweet. Hasil akurasi dari klasifikasi menggunakan algoritma *Support Vector Machine* sebesar 74,39%. Selanjutnya data opini dari kuesioner ditambahkan untuk mengelompokkan pantai berdasarkan ketersediaan sumber daya, fasilitas, akses, kesiapan masyarakat, potensi pasar dan posisi pariwisata. Dalam proses pengelompokan data ini digunakan metode *K-Means*.

Kata kunci : analisis sentimen, wisata pantai, *support vector machine*, *k-means clustering*

1. Pendahuluan

Pada era globalisasi, media sosial saat ini sudah sangat umum dan banyak digunakan untuk kepentingan masyarakat. Dalam implementasinya, media sosial lebih banyak digunakan untuk kegiatan jual beli, menyampaikan informasi, bahkan sebagai media untuk mengekspresikan diri. Pertumbuhan media sosial sangat cepat tidak hanya penggunaanya yang terus meningkat, namun semakin banyaknya media sosial yang ditawarkan melalui aplikasi mobile ataupun website. Salah satu media sosial yang banyak digunakan adalah Twitter. Kata yang terkandung dalam Twitter adalah bahasa alami manusia yang merupakan bahasa dengan struktur kompleks.

Indonesia yang dikenal sebagai negara maritim merupakan salah satu negara yang terkenal akan pariwisatanya. Banyak objek wisata yang berpotensi untuk menarik pendatang lokal maupun wisatawan. Salah satu objek wisata adalah pantai yang sangat banyak dijumpai hampir di seluruh wilayah Indonesia. Berkembangnya objek wisata pantai di Indonesia berawal karena banyaknya masyarakat yang menyampaikan opini tentang kunjungan wisata mereka dari mulut ke mulut, bahkan di era globalisasi sekarang opini menjadi lebih mudah disampaikan karena banyaknya media sosial salah satunya seperti Twitter.

Opini merupakan pendapat, pikiran, atau pendirian. Semakin bertambah banyak kalimat kritik dan saran pada media sosial sehingga dapat membentuk opini masyarakat. Opini ini dapat dijadikan masukan terhadap penilaian suatu objek. Salah satu objek yang sering dibahas dalam twitter adalah mengenai keadaan dan pelayanan di suatu tempat pariwisata yang nantinya akan menjadi opini terhadap masyarakat. Analisis sentimen atau opinion mining dapat digunakan untuk memperoleh gambaran umum persepsi masyarakat terhadap kualitas layanan, apakah cenderung positif, negatif atau netral. Opini biasanya bernilai positif atau negatif tetapi dapat dikategorisasikan juga menjadi baik, sangat baik, buruk, dan sangat buruk.

Wisata pantai adalah salah satu tempat wisata yang menjadi pendukung sektor perekononiam di Indonesia. Wisata pantai menjadi trend karena banyaknya opini masyarakat yang berkunjung. Berdasarkan trend tersebut, permasalahan seperti ketersediaan sumber daya, fasilitas, akses, kesiapan masyarakat, potensi pasar dan posisi pariwisata dapat dilihat dari opini masyarakat. Selanjutnya opini tersebut ditambahkan dengan hasil kuesioner dari beberapa aspek.

Dalam penelitian ini, peneliti mengusulkan metode *Support Vector Machine (SVM)* dalam menganalisis opini yang telah disampaikan oleh beberapa pengguna twitter agar mendapatkan suatu informasi. Metode ini nantinya akan mengklasifikasikan semua opini tentang pariwisata kedalam kategori baik, buruk, maupun netral.

Selanjutnya hasil opini tersebut ditambahkan kuesioner dari masyarakat untuk di c*luster.*

2. Tinjauan Pustaka

2.1 Analisis Sentimen

Analisis sentimen adalah bidang studi yang mengalaisis pendapat, sentiment, evaluasi, penilaian, sikap, dan emosi seseorang terhadap sebuah produk, organisasi, individu, masalah, peristiwa atau topik (Liu, 2012).

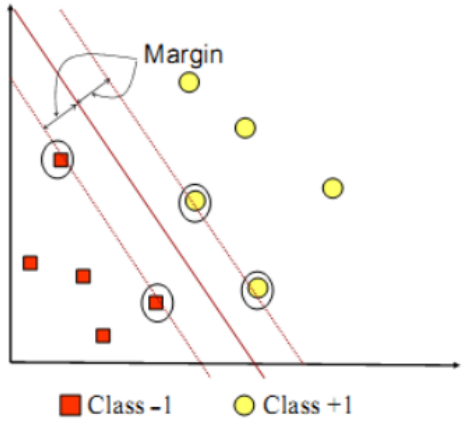
Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek dan menentukan apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif, negatif atau netral. Lebih lanjut sentiment analysis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

Ekspresi atau sentiment mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subjek 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 baterai maka akan menjadi hal positif. Oleh karena itu pada beberapa penelitian, terutama pada review produk, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses opinion mining. *Sentiment analysis* digunakan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang menuju ke opini positif atau negatif (Pang, 2002).

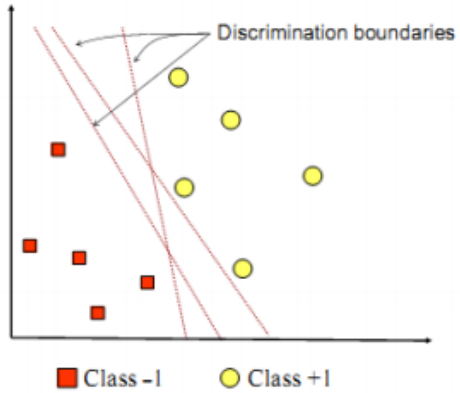
2.2 Support Vector Machine

Support Vector Machine pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep- konsep unggulan dalam bidang pattern recognition (Feldman, 2007). *SVM* adalah algoritma *machine learning* yang bekerja atas prinsip *Structural Risk Minimization (SRM)* dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*.

Gambar 2.1 memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class* : +1 dan –1. *Pattern* yang tergabung pada *class* –1 disimbolkan dengan warna merah (kotak), sedangkan *pattern* pada *class* +1, disimbolkan dengan warna kuning (lingkaran). Masalah klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut (Feldman, 2007).



Gambar 1. *Hyperplane* terbaik yang memisahkan kedua kelas -1 dan +1



Gambar 2 *Hyperplane* terbentuk di antara kelas -1 dan +1

Hyperplane pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur *margin hyperplane* tersebut. dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan pattern terdekat dari masing- masing *class*. *Pattern* yang paling dekat ini disebut sebagai support *vector*. Garis *solid* pada gambar menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah- tengah kedua *class*, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada *SVM*.

2.3 K-Means Clustering

K-Means merupakan metode pengklasteran secara *partitioning* yang memisahkan data ke dalam kelompok yang berbeda. Dengan *partitioning* secara iteratif, *K-Means* mampu meminimalkan rata-rata jarak setiap data ke klasternya. Dalam algoritam *K-Means*, setiap data harus termasuk dalam *cluster* tertentu pada suatu tahapan proses, pada tahapan proses berikutnya dapat berpindah ke *cluster* yang lain. Pada dasarnya penggunaan algoritma *K-Means*

dalam melakukan proses *clsutering* tergantung dari data yang ada dan ada konklusi yang ingin dicapai.

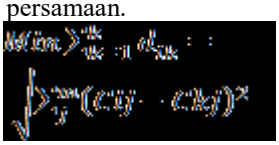
Algoritma *K-Means* pada awalnya mengambil sebagian dari banyaknya komponene dari populasi untuk dijadikan pusat *cluster* awal. Pada step ini pusat *cluster* dipilih secara acak dari sekumpulan populasi data. Berikutnya *K-Means* menguji masing-masing komponen tersebut ke salah satu pusat *cluster* yang telah didefinisikan tergantung dari jarak minimum antar komponen dengan tiap-tiap *cluster.* Posisi pusat *cluster* akan dihitung kembali sampai semua komponen data digolongkan kedalam tiap-tiap *cluster* dan terakhir akan terbentuk posisi *cluster* baru.

Algoritma *K-Means* pada dasarnya melakukan dua proses yakni proses pendeteksian lokasi pusat *cluster* dan proses pencarian anggota dari tiap-tiap *cluster.* Proses *clsutering* dimulai dengan mengidentifikasi data yang akan di*cluster* Xij (i=1,....,n; j=1,...,m) dengan n adalah jumlah data yang akan di*cluster* dan m jumlah variabel. Pada awal iterasi, pusat setiap *cluster* ditetapkan secara bebas, Ckj (k=1,....,n; j=1,...,m). Kemudian dihitung jarak antara setiap data dengan pusat *cluster* ke-k (ck), diberi nama (dik), dapat digunakan formula *euclidean*. Suatu data akan menjadi anggota dari *cluster* ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat *cluster* lain. Proses dasar *algoritma K-Means* antara lain

a. Tentukan k sebagai jumlah *cluster* yang ingin dibentuk. Tetapkan pusat *cluster.*
b. Hitung jarak setiap data ke pusat *cluster* menggunakan persamaan.

$$dik = \sqrt{\sum_j^m (Cij - Ckj)^2}$$

c. Kelompokkan data ke dalam *cluster* yang dengan jarak yang paling pendek menggunakan persamaan.



d. Hitung pusat cluster yang baru menggunakan persamaan

$$Ckj = \frac{\sum_{i=1}^{p} Xij}{p}$$

Dimana:

$$Xij \in cluster\ ke-k$$

P = banyaknya anggota *cluster* ke k.

Ulangi langkah b sampai d hingga sudah tidak ada lagi data yang berpindah ke *cluster* yang lain (Fenty, 2015).

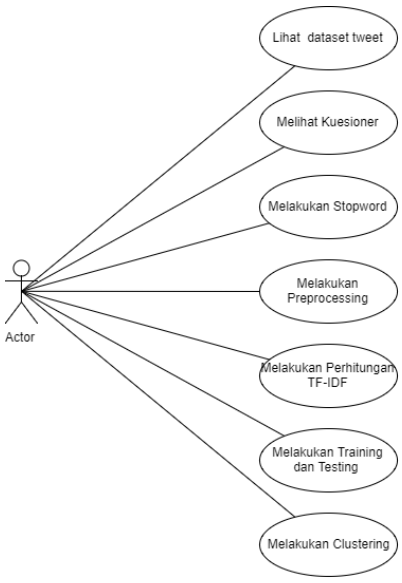
3. Perancangan dan Implementasi

3.1 Perancangan Sistem

Perancangan sistem merupakan suatu desain sistem sebagai penggambaran, perencanaan dan pembuatan sketsa atau pengaturan dari beberapa elemen yang terpisah ke dalam satu kesatuan yang utuh dan berfungsi.

Rancangan akan dibagi menjadi 4 yaitu *use case diagram*, *flowchart* dan perancangan antar muka (*interface*) yang ditampilkan ke dalam bentuk *mockup.*

3.1.1 Use Case Diagram



Gambar 3. Use Case Diagram

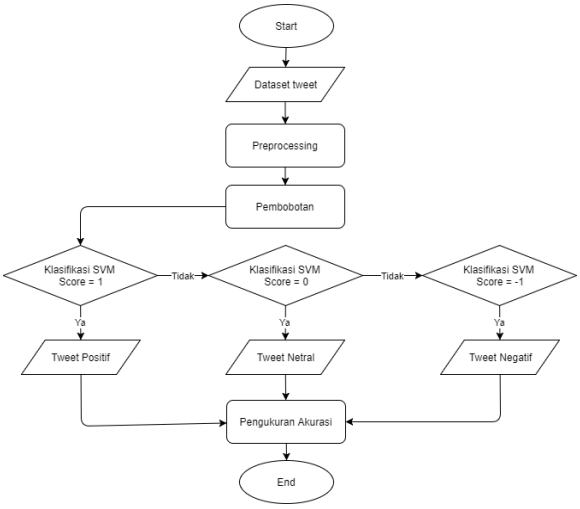
Pengguna dapat melakukan proses klasterisasi yang terdiri dari beberapa fitur yaitu:

a. Pengguna dapat melihat *tweet* kotor yang telah diambil sebelumnya melalui Twitter API.
b. Pengguna dapat mengisi dan melihat hasil kuesioner.
c. Melihat data stopword yang merupakan daftar kata-kata yang nantinya akan dihapus.
d. Proses pembersihan kata-kata dari *tweet* kotor.
e. Pembobotan dengan cara menghitung tingkat kemunculan kata.
f. Membagi 70% data sebagai data pelatihan dan 30% sebagai data uji klasifikasi.
g. Melihat hasil klasterisasi dari hasil kuesioner.

3.1.2 Flowchart

Flowchart adalah cara penulisan algoritma dengan menggunakan notasi garis. *Flowchart* merupakan gambar atau bagan yang memperlihatkan urutan atau langkah-langkah dari suatu program dan hubungan antar proses beserta pernyataannya. Gambaran ini dinyatakan dengan simbol. Dengan demikian setiap simbol menggambarkan proses tertentu. Sedangkan antara proses digambarkan dengan garis penghubung. Dengan menggunakan *flowchart* akan memudahkan kita untuk melakukan pengecekan bagian yang terlupakan dalam analisis masalah.

Pada sistem Implementasi Analisis *Clustering* Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode *K-Means* terdapat *flowchart* yang menjelaskan alur metode *Support Vector Machine (SVM)* dan *K-Means Clustering* yang dapat dilihat pada gambar dibawah ini.

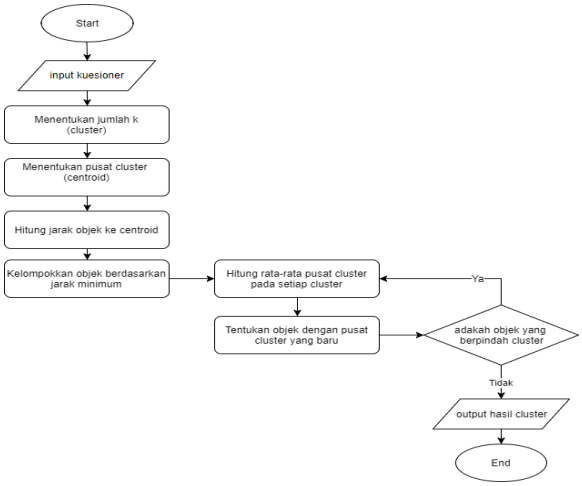


Gambar 4. Flowchart Klasifikasi SVM

SVM adalah metode yang sangat cocok digunakan untuk masalah klasifikasi, tetapi SVM hanya dapat mengklasifikasikan tentang titik dalam ruang. Untuk tujuan ini model ruang *vector* digunakan untuk memberikan setiap kata dalam dokumen sebuah ID (dimensi) dan sebuah bobot berdasarkan seberapa penting keberadaan *term* dalam dokumen (posisi dokumen dalam dimensi itu). SVM mencoba untuk menemukan garis yang terbaik untuk membagi dua kelas, dan kemudian mengklasifikasikan dokumen uji berdasarkan di sisi mana dari garis tersebut akan muncul. SVM akan memisahkan *vector* berdasarkan garis terbaik yang memiliki margin terbesar diantaranya dan contoh titik pelatihan terdekat di kedua sisinya. Oleh karena itu, *vector* contoh (data *training*) berperan besar dalam menentukan margin tersebut adalah yang paling dekat dengan *dividing lines*. Sehingga SVM dapat memberikan keputusan fungsi (kelas atau bukan kelas) untuk classifier.

Berdasarkan *flowchart* pada Gambar 5 menjelasakan bahwa :

1. Memasukkan data kuesioner yang telah diisikan masyarakat
2. Menentukan jumlah k atau jumlah *cluster*
3. Menentukan pusat *cluster* sesuai dengan jumlah *cluster*
4. Htung jarak objek ke pusat *cluster* atau *centroid* menggunakan rumus *euclidean distance*
5. Mengelompokkan data sesuai *cluster* berdasarkan jarak minimun
6. Hitung rata-rata setiap *cluster* untuk mendapatkan pusat *cluster* yang baru
7. Htung jarak objek ke pusat *cluster* atau centroid yang baru dan mengelompokkan berdasarkan jarak minimum
8. Jika masih ada objek yang berpindah *cluster* maka hitung rata-rata setiap *cluster* untuk mendapatkan pusat *cluster* yang baru
9. Jika sudah tidak ada objek yang berpindah maka perhitungan selesai dan akan tampil hasil proses *clustering*.

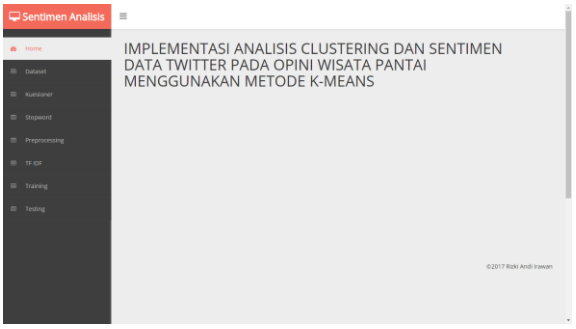


Gambar 5. Flowchart K-Means Clustering

3.2 Implementasi

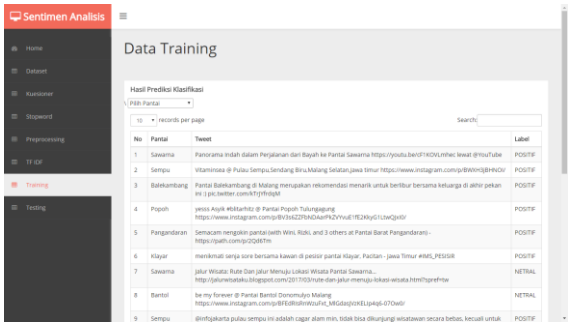
3.2.1 Pembuatan AntarMuka

Tampilan pada menu utama berisi judul penelitian yang dilakukan dan pada sebelah kiri terdapat 8 menu utama, yaitu home, dataset, kuesioner, stopword, preprocessing, TF-IDF, Training, dan Testing, dapat dilihat pada Gambar 6.



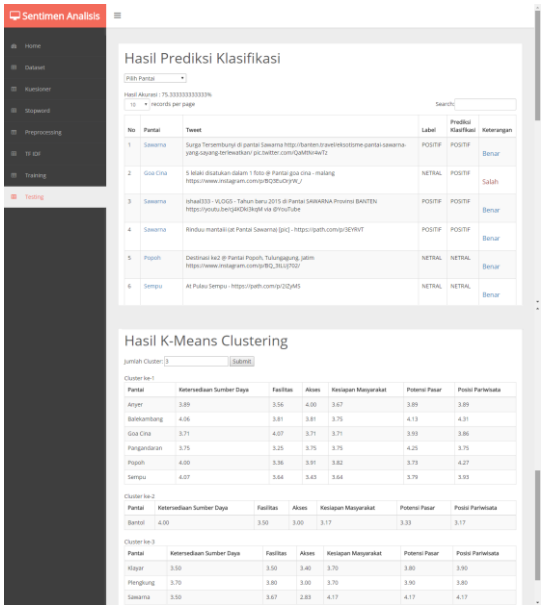
Gambar 6. Antarmuka Menu Utama

Pada antarmuka training berisi data yang digunakan sebagai data pelatihan untuk *Support Vector Machine*. Jumlah data untuk pelatihan adalah 70% dari keseluruhan dataset. Gambar antarmuka training dapat dilihat pada Gambar 7.



Gambar 7. Antarmuka Menu Training

Pada antarmuka testing berisi hasil klasifikasi dari total 30% dataset yang dilakukan oleh *Support Vector Machine* sebagai data uji sistem. Hasil klasifikasi tadi akan menunjukkan tingkat akurasi berdasarkan benar atau salahnya sistem mengklasifikasi data uji tersebut. Terdapat juga hasil *K-Means Clustering* dari Kuesioner yang mana dapat disesuaikan jumlah cluster yang dipilih. Gambar antarmuka testing dapat dilihat pada gambar 8.



Gambar 8. Antarmuka Menu Testing

4. Hasil Pengujian

Pengujian akurasi sistem dilakukan dengan cara menghitung nilai dari *accuracy*. Rumus untuk menghitung nilai *accuracy* sebagai berikut.

$$\frac{\sum v}{n} * 100$$

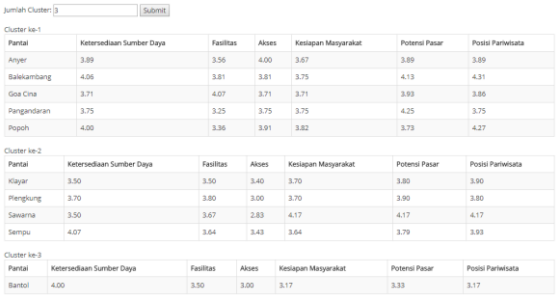
Keterangan
v : Jumlah data benar
n : Jumlah dokumen

Pengujian dilakukan dengan jumlah data *training* yang berbeda. Pada setiap data training akan dilakukan 5 kali pengujian, sehingga tingkat akurasi sistem akan dihitung berdasarkan rata- rata dari setiap jumlah data training. Tabel 1 akan menyajikan hasil pengujian akurasi sistem.

Tabel 1. Pengujian Akurasi Sistem

Pengujian	*Accuracy*
1	76,66 %
2	78,66 %
3	69,33 %
4	72,00 %
5	75,33 %
Rata- rata	**74,39 %**

Analisa hasil clustering dilakukan untuk mengetahui potensi ketersediaan sumber daya, fasilitas, akses, kesiapan masyarakat, potensi pasar, posisi pariwisata yang ada pada pantai di Indonesia.



Gambar 9. Hasil Clustering Pantai

Berdasarkan Gambar 9 telah dilakukan cluster menjadi tiga cluster, dari hasil cluster yang telah dilakukan menampilkan bahwa hasil pada cluster pertama Pantai Anyer, Balekambang, Goa Cina, Pangandaran, dan Popoh menunjukan bahwa pantai tersebut memiliki nilai tiap kriteria yang tinggi. Pada cluster kedua Pantai Klayar, Plengkung, Sawarna, dan Sempu menunjukan pantai tersebut memiliki nilai yang lebih rendah dibandingkan cluster pertama. Pada cluster ketiga Pantai Bantol menunjukan bahwa pantai tersebut memiliki nilai yang rendah dibandingkan cluster pertama dan kedua. Berdasarkan hasil cluster tersebut bahwa cluster pertama dengan nilai tiap kriteria yang tinggi dapat dijadikan sebagai destinasi tujuan wisata.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Adapun kesimpulan yang dapat diambil dari hasil penelitian yang dilakukan mengenai Implementasi Analisis *Clustering* Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode *K-Means* adalah sebagai berikut.

1. Implementasi *Support Vector Machine (SVM)* pada Analisis Sentimen bekerja dengan baik.

2. Algoritma *SVM* memberikan hasil dengan rata-rata *accuracy* yang didapatkan oleh sistem sebesar 74,39%
3. Hasil *K-Means clustering* dipengaruhi dari nilai titik pusat *cluster (centroid)* yang dan jumlah data yang digunakan. Selain itu perbedaan pengambilan data pusat *cluster* awal yang digunakan juga akan mempengaruhi hasil akhir pengelompokkan.

5.2 Saran

Berdasarkan penelitian, ada beberapa hal yang disarankan antara lain sebagai berikut.

1. Aplikasi dapat dijalankan secara realtime guna mendapatkan hasil data yang selalu terbaru setiap waktunya.
2. Pada penelitian ini sistem tidak memiliki fitur *stemming* yang digunakan untuk mencari kata dasar/baku suatu kata, untuk penelitan selanjutnya diharapkan sistem memiliki fitur *stemming* pada proses *preprocessing*.

Daftar Pustaka:

Liu, B. *Sentiment Analysis and Opining Mining*. Morgan & Claypool Publishers, 2012

Pang, Bo and Lee, L, Vaithyanathan, S. 2002. "Sentiment Classification Using Machine Learning Techniques". Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing (EMNLP-02). USA, 2002.

Feldman, R, Sanger, J,. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. New York : Cambridge University Press, 2007.

Fenty Eka, dkk., "Implementasi Algoritma K-Means untuk Menetukan Kelompok Pengayaan Materi Mata Pelajaran Ujian Nasional (Studi Kasus : SMA Negeri 101 Jakarta)", Jurnal Teknik Infomatika, 2015.

Priyanto Hidayatullah dan Jauhari Khairul Kawistara, "Pemrograman WEB", Bandung : Informatika, 2014