ISSN 2460-1160

SENTIMEN ANALISIS TERHADAP OBJEK WISATA ALAM KOTA MALANG DI INSTAGRAM DAN FACEBOOK MENGGUNAKAN METODE *NAIVE BAYES* DAN *SUPPORT VECTOR MACHINE*

Gunawan Budi Prasetyo[1], Habibie Ed Dien[2], Dikta Afif Rahman Prasetyo[3]

[1,2,3] Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1]gunawan.budi@polinema.ac.id, [2]habibie@polinema.ac.id, [3] afifdikta@gmail.com

Abstrak— Perkembangan teknologi sangat cepat di era globlalisasi ini membuat banyaknya muncul seperti Instagram dan Facebook. Dengan adanya media sosial seperti Instagram dan Facebook pengguna dapat saling bertukar informasi. Pengguna media sosial dapat mengunggah informasi tanpa adanya batasan. Dari data informasi yang di unggah di media sosial oleh pengguna Instagram dan Facebook dapat memberikan penilaian dari lokasi objek wisata alam terdapat di kota Malang melalui komentar yang diberikan pengguna. Akan tetapi di menganalisa penilaian objek wisata tersebut sangat tidak mungkin jika di analisa secara manual. Untuk menangani hal tersebut di buatlah sebuah sistem yang dapat menganalisa dari dari komentar objek wisata alam tersebut berdasarkan lokasi wisata.

Dari komentar Instagram dan Facebook pengguna dapat diambil melalui proses *scraping* dan mengolahnya menjadi *text mining.* Setelah diolah komentar tersebut di kelompokkan menggunakan algoritma *Support Vector Machine* dan *Naive Bayes.* Dengan kedua metode tersebut bisa di gunakan untuk perbandingan manakah hasil akurasi yang terbaik untuk penelitian ini. Hasil keluaran dari sistem berupa tingkat akurasi dari kedua metode tersebut berdasarkan ke-dua sosial media tersebut. Data training untuk Instagram sebanyak 725, sedangkan untuk testing sebanyak 273. Data training untuk Facebook sebanyak 682, sedangkan untuk testing sebanyak 231. Berdasarkan Hasil akurasi sangat bergantung dengan keberadaan data pada data training. Semakin sedikit data testing yang tidak memiliki kesamaan dengan data training maka semakin kecil akurasinya. Sebaliknya, semakin banyak data testing yang memiliki kesamaan dengan data training maka semakin besar akurasinya, jadi hasil nilai akurasi tertinggi pada ke-dua metode yaitu *Naive bayes* 75% pada sosial media Instagram dan 60% pada media sosial Facebook sementara *Support Vetor Machine* 60% pada media sosial Instagram dan 55% pada media sosial Facebook.

Kata kunci—Sentimen Analisis, *Naive Bayes* dan *Support Vector Machine* ,Lokasi Wisata Alam

I. PENDAHULUAN

Media sosial telah menjadi trend yang paling banyak dimanfaatkan oleh pengguna Internet untuk berbagi informasi kepada masyarakat luas. Pengguna dapat berbagi berbagai macam konten seperti gambar, video, atau artikel di media sosial. Pengguna membagikan pendapat melalui media sosial, seperti Twitter, Facebook, Instagram dan sebagainya [1]. Facebook dan Instagram adalah media sosial yang paling banyak digunakan ke-tiga dan ke-empat setelah Youtube dan Whatsapp yang digunakan oleh masyarakat di indonesia, yaitu oleh pengguna yang berumur sekitar 18 sampai 34 tahun sebanyak 80.5% [2].

Menganalisis komentar masyarakat dari objek wisata alam tersebut sangat sulit dan membuang banyak waktu jika di lakukan secara manual diakibatkan banyaknya komentar yang di unggah oleh masyarakat di Instagram dan Facebook. Komentar merupakan sebuah pendapat dari masyarakat. Dalam penelitian ini sebuah komentar digunakan untuk menganalisis sebuah objek wisata alam dari berbagai pendapat masyarakat untuk ulasan dari objek wisata alam tersebut. Salah satunya dengan cara memanfaatkan teknologi sentimen analisis terhadap kometar-komentar yang telah di unggah oleh masyarakat. Sentimen analisis, atau disebut juga *opinion mining*, merupakan bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap dan emosi publik terhadap suatu entitas seperti produk, pelayanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya [3].

Berdasarkan latar belakang di atas, pada penelitian ini akan dilakukan sentimen analisis terhadap suatu objek wisata alam dari sebuah komentar yang ada di Instagram dan Facebook. Data yang akan di proses dari penelitian ini yaitu mengambil data komentar Instagram dan Facebook dari masing-masing objek wisata. Pada proses pengolahan data menggunakan metode *Naive Bayes* dan *Support Vector Machine* (SVM) terhadap komentar-komentar yang ada di sosial media tersebut.

II. LANDASAN TEORI

A. *Text Mining*

Text mining adalah salah satu bidang khusus dari data mining. Hanya saja, yang membedakannya adalah pada sumber datanya, dimana text mining bersumber dari kumpulan dokumen atau teks. Sesuai dengan buku *The Text Mining Handbook*, *Text Mining* dapat didefinisikan sebagai suatu

proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam *data mining* yang salah satunya adalah klasifikasi. Pada proses klasifikasi ini, dokumen akan dikelompokkan ke dalam kategori yang sesuai berdasarkan pola yang dibuat pada proses training.

Sebelum dilakukan proses klasifikasi, perlu dilakukan tahapan *text preprocessing* terlebih dahulu untuk mengubah bentuk *text* yang belum terstruktur menjadi bentuk yang sesuai dengan klasifikasi [3].

B. *Preprocessing*

Metode pengolahan data ini dilakukan dengan cara melakukan *preprocessing* yaitu tahapan awal *dari text mining* untuk mengubah data sesuai kebutuhan [4]. Proses yang dilakukan pada tahap ini seperti *casefolding*, *data cleaning*, *stop removal*, *stemming*, dan *tokenizing.*

C. *TF-IDF*

TF-IDF (*Term Frequency – Inverse Document Frequency*) adalah teknik pembobotan yang sering diterapkan di berbagai permasalahan penggalian informasi. Metode ini merupakan gabungan antara metode *term frequency* (tf) dengan metode *inverse document frequency* (idf). *Term frequency* (tf) merupakan suatu metode yang digunakan untuk mencari bobot suatu kata dalam dokumen kunci di setiap kategori dan mencari kata kunci yang hampir mirip dengan kategori yang tersedia [5].

D. *Naive Bayes*

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan. Algoritma mengunakan *teorema Bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan *Naive Bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Data Training*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. *Naive Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan [6].

Proses *Naive Bayes* dibagi kedalam 2 proses yaitu proses *training* dan *testing*. Proses training digunakan untuk menghasilkan model sentimen analisis yang nantinya akan digunakan sebagai pedoman dalam klasifikasi dengan data testing atau data yang berbeda. Berikut adalah algoritma klasifikasi untuk proses *training* dan *testing* pada algoritma *Nave Bayes.*

1) Proses *training*

Pada proses Naive Bayes menggunakan rumus sebagai berikut:

$$P(word\ class) = (Tct + ב(Nc + ב V) \quad (1)$$

Keterangan :

Tct : Term Frequency dari kata i dalam kategori

Nc : Banyak kata keseluruhan dalam kategori C

V : Jumlah kata

ב : Positive constant, biasanya 1. Untuk menghindari nilai 0

Pada rumus (1) bertujuan untuk menghitung bobot atau nilai probabilitas setiap kata dalam data training di setiap kategori klasifikasi, kemudian setiap nilai probabilitas kata tersebut digunakan dalam *proses testing*.

2) Proses *testing*

Pada proses testing dalam algoritma Naive Bayes menggunakan rumus sebagai berikut:

$$Vmap = \frac{argmax}{V\ j\ e\ V} \prod \quad i = i^{p}\ P(xi, |Vj)P(Vj) \qquad (2)$$

Keterangan :

Vmap : Semua Kategori yang diujikan V

jl : Class Positif

j2 : Class Negatif

j3 : Class Netral

P(Vj) : Probabilitas Xi pada kategori Vj

P(Vj) : Probabilitas dari Vj

Dengan rumus diatas, setiap nilai dari kata pada dokumen testing akan dihitung berdasarkan nilai probabilitas setiap kata yang dihasilkan dari proses training. Perhitungan dengan rumus didatas, dilakukan untuk setiap kategori klasifikasi kemudian dicari Vmap tertinggi.

E. *Support Vector Machine*

SVM merupakan salah satu metode terbaik yang bisa dipakai dalam permasalahan klasifikasi. Konsep *SVM* bermula dari masalah klasifikasi dua kelas sehingga membutuhkan training set positif dan negatif. *SVM* berusaha menemukan *hyperplane* (pemisah) terbaik untuk memisahkan ke dalam dua kelas dan memaksimalkan margin antara dua kelas tersebut. Pada beberapa kasus, data tidak bisa diklasifikasi menggunakan metode linier *SVM*, sehingga dikembangkan fungsi *kernel* untuk mengklasifikasikan data dalam bentuk *nonlinier*. Pada

penelitian menggunakan kernel *Polynomial Degree*. *Sequential Training* memiliki algoritme yang lebih sederhana dan waktu yang diperlukan lebih cepat. Adapun algoritme *Sequential Training* adalah sebagai berikut:

1) Menginisialisasi = 0 αi dan parameter lain, misalnya λ = 0,5, γ = 0,01, c = 1, *IterasiMax* = 100, dan ε = 0,001. Kemudian menghitung *matriks Hessian* dapat dihitung dengan rumus persamaan.

$$D_{ij} = y_i y_j (K(x_i, x_j) + \lambda^2) \quad (3)$$

2) Mulai dari data ke i sampai j , hitung menggunakan persamaan.

a) $E_i = \sum_{j-1}^{n} a_j D_{ij}$ (4)

b) $\delta a_i = min\{max[\gamma(1 - E_i), -a_i], C - a\}$ (5)

c) $a_i = a_i + \delta a_i$ (6)

3) Langkah 2 dilakukan terus-menerus hingga kondisi iterasi maksimum tercapai atau *max* Selanjutnya didapatkan nilai *support vector* (SV), SV = *(Threshold SV)*. Nilai *Threshold SV* didapatkan dari beberapa percobaan, biasanya digunakan *threshold* > 0. Kemudian dilakukan proses testing untuk mendapatkan keputusan di mana fungsi keputusan dapat di hitung dengan persamaan.

a) $f(x) = \sum_{i=1}^{m} a_i y_i K(x_i, x) + b$ (7)

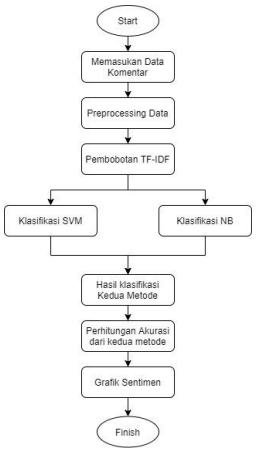
b) $-[\sum_{i=1}^{m} a_i y_i K(x_i, x^+) + \sum_{i=1}^{m} a_i y_i K(x_i, X^-)$ (8)

III. METODELOGI

Metodologi pengembangan sistem yang digunakan pada penelitian ini adalah *waterfall*. Metode ini menyarankan pendekatan pengembangan secara sekuen dan sistematik untuk pengembangan sistem. Metode ini memiliki beberapa tahapan terdiri dari *Requirements analisis and definition, System and software design, Implementation and unit testing, Integration and system testing dan Operation and maintenance*.

A. Perancangan Penelitian

Sistem yang dirancang dan dibangun dalam penelitian ini adalah sistem yang mengimplementasikan kedua metode *support vector machine* dan *naive bayes* untuk mengklasifikasikan sentimen pada data komentar instagram dan facebook dengan menargetkan kata kunci tertentu. sistem akan dapat memasukan data komentar, *preprocessing* data, pembobotan menggunakan *TF IDF*, klasifikasi dan pengujian klasifikasi. Grafik sentimen akan menampilkan persentase jumlah tiap-tiap kategori sentimen berdasarkan kata kunci terkait dan juga menampilkan grafik tren.



Gambar 1 Perancangan Penelitian

B. Pengolahan Data

Data yang di dapatkan yang telah melalui proses scraping yang merupakan data langsung dari komentar pengguna instagram dan facebook berupa file json yang di konversikan menjadikan ke file .xls. Data perlu dilakukan pengolahan agar menjadi data yang mudah digunakan dalam proses sentiment analysis. Beberapa kata yang tidak di gunakan akan di hapus yang disebut dengan proses *preprocessing* untuk memudahakan pada proses pembobotan. Setelah melalui preprocessing, sebelum proses tahapan metode data akan terlebih dahulu diseleksi untuk dikategorikan kedalam positif, netral dan negatif. Setelah proses pengkatogorian selesai maka data yang berbentuk teks akan melaui tahapan proses pembobotan *TF-IDF* digunakan untuk mengolah kata menjadi sebuah angka sebelum memasuki kedua metode. Selanjutnya data akan di klasifikasi menggunakan perbandingan kedua metode yaitu *Naive Bayes* dan *Support Vector Machine* untuk menghasilkan peluang yang berbobot *positif, netral* dan *negatif*.

C. Pengujian

Pengujian perangakat lunak adalah elemen kritis dari jaminan kualitas perangkat lunak dan mepresentasikan kajian pokok dari spesifikasi, desain, dan pengkodean. Pengujian yang digunakan dalam penelitian kali ini menggunakan akurasi. Akurasi adalah hasil pengukuran seberapa dekat suatu angka hasil terhadap angka sebenarnya (*true value or reference value*). Dalam penelitian ini akurasi hasil dihitung dari jumlah hasil yang tepat dibagi dengan jumlah data. Tingkat akurasi diperoleh dengan perhitungan sesuai dengan persamaan 5.
Pengujian akurasi sistem dilakukan dengan menghitung jumlah data yang benar (memiliki kategori yang sama dengan kategori data sebenarnya), dibagi dengan jumlah data keseluruhan dan dikalikan 100 [7].

$$Akurasi = \frac{Jumlah\ data\ yang\ sesuai}{Jumlah\ data\ keseluruhan} x100\% \quad (9)$$

1) Pengujian *Naive Bayes*

Berdasarkan pengujian yang telah dilakukan, diambil rata-rata dari hasil setiap percobaan tersebut. Berikut adalah hasil representasi akurasi:

Tabel 1 Hasil pengujian *Naive Bayes*

IG		FB	
Jumlah data training	725	Jumlah data training	682
Jumlah data testing	273	Jumlah data testing	231
Akurasi	75%	Akurasi	60%

2) Pengujian *Support Vector machine*

Berdasarkan pengujian yang telah dilakukan, diambil rata-rata dari hasil setiap percobaan tersebut. Berikut adalah hasil representasi akurasi:

Tabel 2 Hasil Pengujian *Support Vector Machine*

IG		FB	
Jumlah data training	725	Jumlah data training	682
Jumlah data testing	273	Jumlah data testing	231
Akurasi	60%	Akurasi	55%

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan dapat ditarik kesimpulan sebagai berikut :

1) Algoritma *Naive Bayes* dan *Support Vector Machine* dapat digunakan untuk mengklasifikasikan komentar tentang pariwisata pada sosial media.
2) Hasil pengujian akurasi dari *Naive Bayes* untuk data Instagram adalah 75%, sementara untuk data Facebook adalah 60%. Hasil pengujian akurasi dari *Support Vector Machine* untuk data Instagram adalah 60%, sementara untuk data Facebook adalah 55%. Hasil akurasi sangat bergantung dengan keberadaan data pada data *training*. Semakin sedikit data testing yang tidak memiliki kesamaan dengan data training maka semakin kecil akurasinya. Sebaliknya, semakin banyak data testing yang memiliki kesamaan dengan data *training* maka semakin besar akurasinya.
3) Hasil pengujian akurasi dengan *Naive Bayes* memiliki nilai yang lebih tinggi dari pada *Support Vector Machine*, karena *Naive Bayes* melakukan proses *testing* dengan menghitung *probabilitas* adanya sebuah kata di dalam data *training*, data *testing* pada *Naive Bayes* juga tidak melalui proses *TF-IDF*. Sedangkan *Support Vector Machine* sudah melakukan proses testing diawali dengan menghitung *TF-IDF* sehingga terlalu banyak proses yang dilalui menyebabkan akurasi dengan metode ini tidak seakurat *Naive Bayes*.

B. Saran

Saran yang dapat diberikan dari hasil penelitian untuk pengembangan sistem ini kedepan sebagai berikut:

1) Dapat dilakukan pengambilan data secara *real time*.
2) Dapat dibandingkan menggunakan metode yang lain.
3) Untuk penggunaan data yg besar, tidak disarankan menggunakan metode *SVM* karena memakan waktu yang lama.

V. DAFTAR PUSTAKA

[1] belladina fahmi, A. Wibowo, and D. Hajar, "Analisa Kepribadian Pengguna Facebook Menggunakan Algoritma Support Vector Machine," *J. Komput. Terap.*, vol. 5, no. 1, pp. 28–35, 2019, doi: 10.35143/jkt.v5i1.2259.

[2] Websindo, "Social Platforms: Active User Accounts," 2019. https://websindo.com/indonesia-digital-2019-media-sosial.

[3] Y. Puspitarani, "Sentimen Analysis Terhadap Nilai Kepercayaan Sebuah Online Shop Di Instagram," *J. Ilm. Teknol. Inf. Terap.*, vol. II, no. 1, pp. 76–81, 2015.

[4] W. A. Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 11, pp. 4704–4713, 2018.

[5] P. H. Saputro, M. Aristin, and Dy. L. Tyas, "Berdasarkan Lirik Menggunakan Metode Tf-," *J. Teknoloi Inform. dan Terap.*, vol. 4, no. 1, pp. 45–50, 2017.

[6] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Creat. Inf. Technol. J.*, vol. 2, no. 3, pp. 207–217, 2015.

[7] A. A. Indra Wiratmaka, I. F. Rozi, and R. A. Asmara,

"Klasifikasi Kualitas Tanaman Cabai Menggunakan Metode Fuzzy K-Nearest Neighbor (Fknn)," *J. Inform. Polinema*, vol. 3, no. 3, p. 1, 2017, doi: 10.33795/jip.v3i3.25.