Pendekatan berbasis *Machine Learning* dan Leksikal Pada Analisis Sentimen

Elvia Budianita*[1], Eka Pandu Cynthia[2], Anggi Pranata[3], Dicky Abimanyu[4]
[1,2,3,4] Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
Email: [1]elvia.budianita@uin-suska.ac.id, [2]eka.pandu.cynthia@uin-suska.ac.id,
[3]anggi.pranata1@student.uin-suska.ac.id,

Abstrak

Pada analisis sentimen terdapat dua pendekatan. Pertama berbasis *machine learning* dengan melatih data latih pada dataset yang telah dilabelkan secara manual dengan melibatkan seorang pakar atau Annotator. Pendekatan yang kedua adalah berbasis leksikal (*Lexicon Based*) yang tidak memerlukan pelatihan dataset untuk menemukan polaritas sentiment. Data set komentar yang digunakan adalah mengenai penyedia jasa transportasi online local Maxim di media social *Twitter*. Data set komentar yang dilabel secara manual akan diklasifikasikan ke dalam kelas positif netral, dan negatif menggunakan metode *Support Vector Machine* (SVM). Berdasarkan hasil pengujian diperoleh kesimpulan bahwa analisis sentimen untuk kasus Maxim menggunakan pelabelan manual yang dilatih menggunakan metode SVM adalah lebih banyak mengandung kalimat positif sedangkan jika menggunakan pelabelan *Lexicon based* lebih banyak mengandung kalimat netral.

Kata kunci: *Analisis sentimen, Lexicon based, machine learning, support vector machine*

Abstract

There are two approaches to sentiment analysis. The first is based on machine learning by training data on datasets that have been manually labeled by involving an expert or annotator. The second approach is lexical-based which does not require dataset training to find sentiment polarity. The collection of comment data used is about local online transportation service provider Maxim on Twitter social media. The manually labeled comment data set will be classified into neutral positive, and negative classes using the Support Vector Machine (SVM) method. Based on the test results, it can be concluded that sentiment analysis for the Maxim case using manual labeling using the SVM method is more positive sentences if using Lexicon-based labeling contains more sentences.

Keywords*: Lexicon based, machine learning, sentiment analysis, support vector machine*

1. Pendahuluan

Analisis sentimen merupakan proses untuk memahami dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Penelitian di bidang Analisis sentimen atau yang dikenal dengan *opinion mining* mulai marak dilakukan pada tahun 2002. Analisis sentimen digunakan untuk menilai suatu produk. *Opinion mining* bisa dianggap sebagai kombinasi antara *text mining* dan *natural language processing*. Konsep dasar dalam analisis sentimen adalah melakukan klasifikasi atau mengelompokkan teks, kalimat, atau pendapat yang dikemukakan dalam dokumen kedalam polaritas opini positif, negatif, atau netral [1]. Penerapan analisis sentimen selalu berkaitan dengan pengkategorian kelas sentiment atau yang disebut dengan tahap data labeling.

Berbagai penelitian terkait analisis sentimen telah banyak dilakukan. Terdapat dua pendekatan untuk melakukan analisis sentimen, pendekatan yang pertama adalah berbasis *machine learning* yaitu dengan melatih data latih pada dataset yang telah dilabelkan secara manual dengan melibatkan seorang pakar atau Annotator. Pendekatan yang kedua adalah berbasis leksikal (*Lexicon Based*) yang tidak memerlukan pelatihan dataset untuk menemukan polaritas sentiment [2]. Penelitian yang pernah dilakukan dengan judul *Analisis sentimen in Tourism: Capitalizing on Big Data* membandingkan beberapa metode analisis sentimen dalam topik pariwisata. Penelitian ini memperoleh kesimpulan bahwa menggunakan metode *lexicon* memiliki peningkatan hasil sentimen paling tinggi dibandingkan dengan metode lainnya [3]. Sedangkan pada penelitian yang berjudul Implementasi Lexicon Based Dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019 memberikan hasil dengan tingkat akurasi antara sentimen prediksi dan sentimen aktual dengan Lexicon Based sebesar

64,49% pada data uji dan pada data latih sebanyak 94,2% serta dengan menggunakan Labelisasi dan Naive Bayes Classifier sebesar 86,53% pada data uji dan data latih sebesar 94,08% [4].

Berdasarkan penelitian terkait tersebut, penelitan ini dilakukan menggunakan data mengenai sentimen publik berdasarkan posting Twitter tentang layanan transportasi online Maxim dengan pendekatan *Lexicon based* dan *machine learning* metode *Support Vector Machine* (SVM) untuk mengetahui bagaimana hasil klasifikasi dengan pendekatan yang berbeda.

Layanan transportasi *online* Maxim telah berkembang tidak hanya taxi saja melainkan sudah memiliki transportasi *online* (motor dan mobil), pengiriman barang, pesan antar makanan dan barang, kargo, jasa pembersih, dan laundry bahkan meluncurkan *marketplace*. Setiap pelayanan yang diberikan Maxim sangat berkaitan dengan tingkat kepuasan konsumen atau pelanggan. Kepuasan konsumen dapat diciptakan melalui kualitas pelayanan dan penilaian.

Sedangkan metode *Support Vector Machine* (SVM) diperkenalkan pertama kali oleh Vapnik tahun 1992 sebagai salah satu metode *machine learning* yang bekerja dengan prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space* [5]. (Indri monika parapet). Penelitian mengenai metode SVM salah satunya adalah analisis sentimen tanggapan masyarakat Indonesia terhadap pandemi Covid-19 pada media sosial Twitter menggunakan algoritma *Support Vector Machine* (SVM), *Naïve Bayes*, dan K-*Nearest Neighbor*. Pada penelitian ini ketiga algoritma tersebut dibandingkan untuk mengetahui hasil klasifikasi terbaik pada data tanggapan tersebut. Berdasarkan tingkat rata-rata akurasi dengan menggunakan evaluasi model 10-*Fold Cross Validation*, diperoleh kesimpulan bahwa algoritma SVM memiliki akurasi yang lebih tinggi daripada Naive Bayes dan KNN. Rata-rata akurasi metode SVM dengan kernel Linier sebesar 90,01%, pada Naive Bayes sebesar 79,20% dengan jumlah laplace adalah 1, dan pada KNN sebesar 62,10% dengan jumlah K adalah 20 dan menggunakan kernel optimal [6].

2. Metode Penelitian

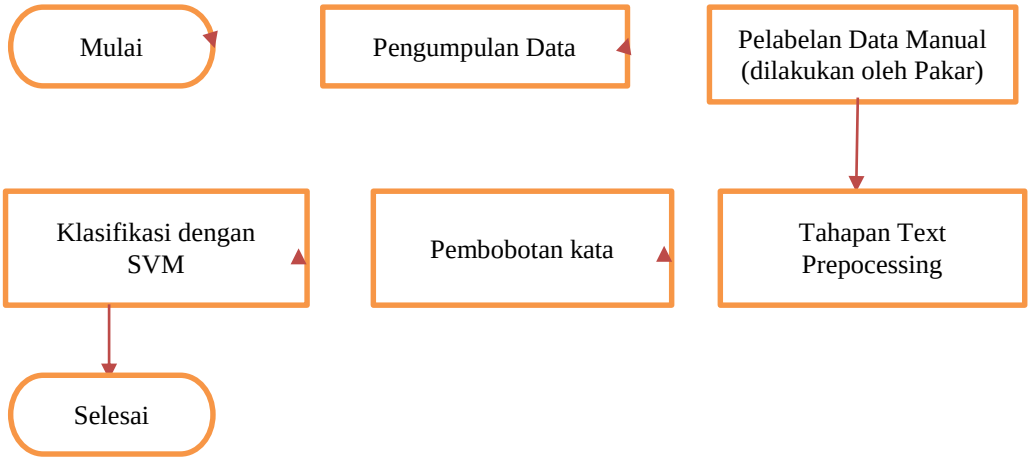
Metode penelitian yang dilakukan terdiri atas dua bagian. Pertama adalah melakukan dengan pendekatan berbasis machine learning menggunakan metode SVM dan yang kedua adalah melakukan dengan pendekatan leksikal yang dapat dilihat pada gambar 1 dan 3.

2.1 Pendekatan berbasis Machine learning menggunakan metode SVM.

Pendekatan pertama untuk melakukan analisis sentiment terhadap opini masyarakat mengenai layanan transportasi online Maxim adalah dengan pendekatan berbasis *Machine learning menggunakan metode SVM* yang dapat dilihat pada gambar 1. Tahapan awal adalah pengumpulan data.

Pengumpulan data komentar diperoleh dari tanggal 1 Juni 2019 sampai dengan 27 Maret 2022 menggunakan *tools python* dengan cara menggunakan *package tweepy, csv dan panda as pd*. Data dikumpulkan dari server *Twitter* menggunakan Twitter API (*Application Programming Interface*) berjumlah 1200 data. Langkah pertama adalah dengan meng-import ketiga package tersebut, lalu memasukkan credit (*consumer key, consumer secret, access token* serta access token secret), kemudian memasukkan *sourcecode crawling* data mulai dari autentifikasi hingga crawling data untuk menentukan save as data, keyword, bahasa dan waktu data diambil. Lalu jalan kan programnya. Jika sudah selesai dijalankan akan diperoleh data dalam entuk file Ms.Excel. Data yang digunakan hanya username dan comment text. Data yang dikumpulkan berasal dari *Tweet* pada media sosial Twitter terhadap akun @Maxim_Indonesia serta *keyword* yang berkaitan dengan akun tersebut. Pelabelan dilakukan oleh Rosmiati, S.Pd sebagai Guru Bahasa Indonesia SMK Teknologi Riau yang terdiri atas 3 kelas yakni kalimat komentar positif, negatif, dan netral.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

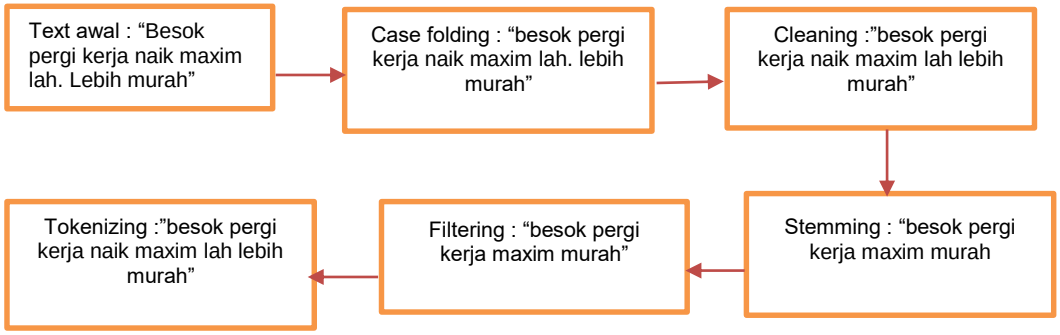


Gambar.1 Tahapan Proses dengan Pendekatan berbasis *Machine learning* menggunakan metode SVM

Selanjutnya adalah tahapan *Text processing*. *Text preprocessing* merupakan proses mempersiapkan data agar data siap untuk diolah, tahapan ini adalah tahapan yang paling penting dalam melakukan mining teks. Tahapan preprocessing dalam melakukan penelitian ini terdiri atas :

a. *Case Folding*
Case folding merupakan perubahan huruf yang terdapat pada kata atau kalimat menjadi huruf kecil secara keseluruhan.
b. *Cleaning*
Cleaning merupakan proses penghapusan kata dan karakter yang tidak dibutuhkan. Seperti mention username (@), link, hashtag (#), emoticon maupun simbol (@#$%^&*()+":{}<>?!~/[]) dan angka.
c. *Tokenizing*
Tokenizing merupakan proses pemisahan kalimat menjadi kata. Pemisahan kata berdasarkan karakter angka, huruf dan spasi.
d. *Filtering*
Filtering merupakan proses penghapusan kata yang tidak penting.
e. *Stemming*
Stemming merupakan perubahan kata dasar yang ber-imbuhan. Algoritma *stemming* yang digunakan yaitu Nazief dan Adriani.

Berikut contoh hasil proses Text processing yang dapat dilihat pada gambar 3. Pada kata digambar 3 menunjukan hasil text processing mengubah huruf besar menjadi huruf kecil, menghilangkan tanda titik dan kata “lah”, serta kata “naik”.



Gambar.2 Contoh Hasil Proses *Text processing*

Setelah tahapan *text processing*, maka selanjutnya adalah tahapan pembobotan kata menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF). Pembobotan kata

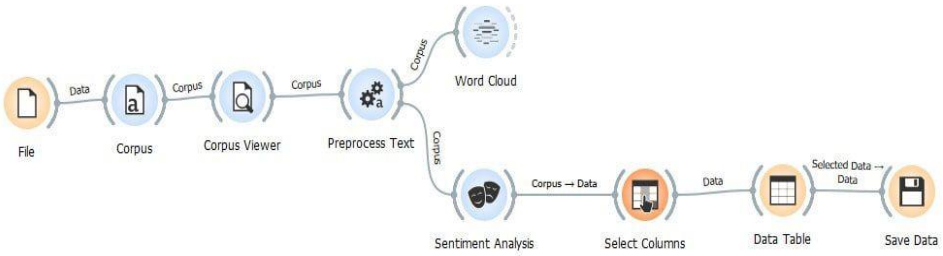
ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

bertujuan agar data berupa teks atau kata tersebut dapat dilatih menggunakan *machine learning*. Metode *Term Frequency Invers Document Frequency* (TF-IDF) merupakan metode yang digunakan untuk menentukan keterhubungan kata (term) terhadap dokumen dengan memberikan bobot setiap kata. Metode TF-IDF ini menggabungkan dua konsep yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen dan inverse frekuensi dokumen yang mengandung kata tersebut [7]. Dalam perhitungan bobot menggunakan TF-IDF, dihitung terlebih dahulu nilai TF perkata dengan bobot masing-masing kata adalah 1.

IDF adalah nilai logaritma dari td per df, dimana td adalah jumlah keseluruhan dokumen yang ada, dan df jumlah kemuculan kata pada semua dokumen. Pembobotan ini dilakukan untuk memberikan bobot pada masing-masing kata dan kemudian akan diklasifikasikan kedalam kelas positif, negatif, dan netral menggunakan metode SVM. Pada metode SVM dicari terlebih dahulu hyperplane atau garis pembatas (decision boundary) yang memisahkan antara suatu kelas dengan kelas lain, garis tersebut berperan memisahkan tweet bersentimen positif (berlabel +1), tweet netral (berlabel 0) dengan tweet bersentimen negatif (berlabel -1). SVM melakukan pencarian nilai hyperplane dengan menggunakan support vector dan nilai margin.

2.2 Pendekatan berbasis leksikal (Lexicon Based).

Pendekatan selanjutnya untuk melakukan analisis sentiment terhadap opini masyarakat mengenai layanan transportasi online Maxim adalah dengan pendekatan berbasis leksikal yang dapat dilihat pada gambar 3 dan 4.



Gambar 3. Proses Analisis Sentimen Menggunakan *Lexicon based*

Pada pendekatan berbasis leksikal, proses pengumpulan data sama dengan pendekatan menggunakan metode SVM. Perbedaannya terletak pada proses pelabelan. Pada *orange data mining*, *lexicon based* bekerja berdasarkan *syntax phyton* dengan memanggil kamus lexicon terlebih dahulu. Pada penelitian ini menggunakan *multilingual sentiment lexicons* dari *the Data Science Lab* yang dapat ditunjukan pada gambar 4.

Selanjutnya data kalimat yang telah dipisah menjadi per kata, akan dihubungkan dengan kamus *lexicon* untuk menghitung nilai polaritas. Setelah nilai polarity nya diperoleh, maka selanjutnya nilai tersebut akan dijumlahkan dan dinormalisasi menggunakan rumus normalisasi *Hutto* untuk mendapatkan nilai *compound* dari kalimat hasil *pre-processing* tersebut. Nilai *compound* digunakan untuk menentukan apakah kalimat tersebut bernilai positif, negatif, atau netral. Setelah melalui proses sentiment analisis, selanjutnya data tersebut akan dipisahkan kolomnya menggunakan *widget select column*. *Widget* ini digunakan untuk memisahkan data yang dibutuhkan pada *widget data table*. Widget data table ini digunakan untuk menampilkan hasil tweet, score positif, negatif, netral, dan *compound* dari hasil analisis sentiment menggunakan metode *lexicon based*. Jika nilai *compound* >=0,05 , maka sentiment bernilai positif, jika nilai compound <0,05 dan > -0,05 , maka sentiment bernilai netral. Sedangkan jika nilai compound < - 0,05 , maka sentiment bernilai negatif.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Gambar 4. Atribut Analisis Sentimen pada *Orange Data mining*

3. Hasil dan Analisa

Hasil klasifikasi Kalimat tentang opini masyarakat mengenai Maxim menggunakan *Lexicon based* dan metode SVM pada 10 data uji dapat ditunjukan pada Tabel 1.

Tabel 1. Hasil klasifikasi Kalimat menggunakan *Lexicon based* dan metode SVM

No	Kalimat	Hasil Uji Lexicon Based	Hasil SVM
1	@hafidznoor @anang_kur Semenjak Noor pake gojek gue jadi takut pake gojek. Mungkin itu kenapa gue masih pake maxim	Nilai : -76.923.076.923.076.900 Statement : Negatif	Positif
2	Dari tadi scroll gojek maxim doang, bingung:(gini ini enanya makan apa si ges , rekom dongg ??	Nilai : -76.923.076.923.076.900 Statement : Negatif	Positif
3	Emang ko pake maxim karena murah wkwk\nTapi kenapa yaa ini driver malah ngejelek jelekin ? Emng salah nyari ojek onl\xe2\x80\xa6 https://t.co/ZFOesKNRbz	Nilai : -43.478.260.869.565.200 Statement : Negatif	Netral
4	Lelet banget ini maxim @maxim__indo	Nilai : 0.0 Statement : Netral	Negatif
5	Gunakan Maxim Transportasi dalam setiap aktifitas anda dan dapatkan berbagai kemudahan, harga yang tidak menguras kantong. Kirim barang hanya 4000an...Bike hanya 7900an...Car hny 5000an..ayo tunggu apa lagi https://t.co/y9ZPNx60ud download segera teman2 @lnfomalang @UB_Official https://t.co/niDoAAdtUU	Nilai : -4.545.454.545.454.540 Statement : Negatif	Positif
6	Daripada bingung karena nggak ada promo gojek atau grab, mending pake Maxim guys! Aku biasa pake dan lebih murah da\xe2\x80\xa6 https://t.co/Zsu03chBkc	Nilai : -5.555.555.555.555.550 Statement : Negatif	Positif
7	berdosa kah kalau naik maxim time ni tapi driver dia lelaki \xf0\x9f\x98\xb5	Nilai : -7.142.857.142.857.140 Statement : Negatif	Netral
8	paling ujungnya maxim lagi emang. https://t.co/a7TsYQ8w2E	Nilai : 0.0 Statement : Netral	Positif
9	Naik maxim aja hehe ongkirnya masih murah https://t.co/Yf3uOLkpkC	Nilai : 0.0 Statement : Netral	Positif
10	liat orang pake jaket kuning kirain anak UI, ternyata abang maxim.	Nilai : -10.0 Statement : Negatif	Positif

Berdasarkan tabel 1 tersebut menunjukan perbedaan hasil klasifikasi komentar mengenai Maxim menggunakan pendekatan *machine learning* metode SVM dengan pelabelan manual dengan pendekatan leksikal. Pelabelan manual yang dilakukan oleh Guru Bahasa Indonesia dengan memahami makna kata sedangkan pelabelan dengan leksikal berdasarkan jumlah polarity kata yang terdapat pada setiap kalimat.

Kemudian pengujian juga dilakukan dengan confusion matrik pada pendekatan *machine learning* metode SVM menggunakan perbandingan data latih dan data uji 90 : 10 dari 1200 data. Akurasi yang diperoleh dapat ditunjukan pada tabel 2. Hasil akurasi yang diperoleh adalah 85 % dengan jumlah kalimat positif lebih banyak dari kalimat negative dan netral.

Tabel 2. Hasil klasifikasi menggunakan metode SVM dengan perbandingan data latih dan data uji 90 : 10 dari 1200 data

Hasil klasifikasi SVM / **Kelas sebenarnya**	**Positif**	**Negatif**	**Netral**
Positif	100	1	0
Negatif	10	0	0
Netral	7	0	2

Sedangkan hasil menggunakan *lexicon based* pada klasifikasi sentiment dengan jumlah data 1200 adalah sebagai berikut :

Positif = 334/1200 x 100% = 27,83%
Netral = 602/1200 x 100% = 50,16%
Negatif = 264/1200 x 100% = 22%
Jumlah komentar terhadap Maxim yang bernilai netral sebesar 50,16%.

4. Kesimpulan

Berdasarkan hasil pengujian diperoleh kesimpulan bahwa analisis sentimen untuk kasus Maxim menggunakan pelabelan manual yang dilatih menggunakan metode SVM adalah lebih banyak mengandung kalimat positif sedangkan jika menggunakan pelabelan Lexicon based lebih banyak mengandung kalimat netral. Pelabelan manual pada metode SVM dilakukan dengan memahami makna kata sedangkan pelabelan dengan leksikal berdasarkan jumlah polarity kata yang terdapat pada setiap kalimat. Penelitian ini dapat dikembangkan dengan pendekatan semantic yakni mengacu pada makna yang disampaikan oleh sebuah teks.

Referensi

[1] Ardiani.L., Sujaini. H., dan Tursina, “Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak,” Jurnal Sistem dan Teknologi Informasi (*Justin)*, vol. 8, no. 2, April 2020.

[2]-[3] Prasetya.Y.N., Winarso.D., dan Syahril, “Penerapan Lexicon Based Untuk Analisis Sentimen Pada Twiter Terhadap Isu Covid-19,” Jurnal Fasilkom, vol. 11, no. 2, Agustus 2021.

[4] Aulia.G.N dan Patriya. E., “Implementasi Lexicon Based Dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019,” Jurnal Ilmiah Informatika komputer, vol. 24, no. 2, Agustus 2019.

[5] Parapat. I.M., Furqon.M.T., dan Sutrisno., “Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak,” Jurnal Pengembangan Teknologi Informasi dan ilmu komputer, vol. 2, no. 10, Oktober 2018.

[6] Pamungkas. F.S. dan Kharisudin.I., “Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter,” Prosiding Seminar Nasional Matematika (PRISMA), Februari 2021.

[7] Deolika. A., Kusrini., dan Luthfi. E., T, “Analisis Pembobotan Kata Pada Klasifikasi Text Mining,” Jurnal Teknologi Informasi, vol.3, no.2, Desember 2019.