Analisis sentimen pada review hotel menggunakan metode pembobotan dan klasifikasi

Aam Shodiqul Munir*[1], Enda Putri Atika[2], and Aziza Devita Indraswari[3]

1,2,3Program Magister Teknik Informatika Program Pascasarjana Universitas Amikom Yogyakarta
Jl. Padjajaran, Ring Road Utara, Kel. Condongcatur, Kec. Depok, Kab.Sleman, Prop. Daerah Istimewa Yogyakarta 55283
aamshodiqulmunir@students.amikom.ac.id;enda.1414@students.amikom.ac.id; aziza.devita@students.amikom.ac.id

Abstrak

Secara global, industri pariwisata memiliki peranan penting dalam kemajuan ekonomi pada suatu wilayah atau negara. perkembangan tersebut dibantu oleh perkembangan teknologi internet seperti sosial media, website portal pariwisata dan lain - lain. Penilaian dari suatu hotel di website portal juga dapat mempengaruhi keinginan konsumen apakah memilih hotel tersebut atau tidak. Analisis sentimen terhadap review yang dikeluarkan oleh konsumen dapat dibagi menjadi review positif atau review negatif. Analisis sentimen dimulai dari pengambilan data yaitu scrapping kemudian diteruskan menuju proses preprocessing sehingga didapat data yang siap untuk dianalis. setelah dilakukan proses preprocessing dilanjutkan dengan proses pembobotan. proses pembobotan menggunakan tiga buah metode yaitu Unigram, bigram dan term *frequency inverse document frequency* (TF-IDF). Setelah dilakukan proses pembobotan dilakukan proses klasifikasi menggunakan dua buah metode yaitu Naïve Bayes dan *support vector machine* (SVM). Hasil dari proses klasifikasi tersebut adalah akurasi tertinggi yang didapat oleh metode pembobotan TF-IDF dan metode SVM sebesar 95% diikuti dengan Metode pembobotan unigram dengan metode SVM sebesar 94%.

Kata Kunci sentimen analisis, review hotel, TF-IDF, unigram, bigram

Digital Object Identifier 10.36802/jnanaloka.v3-no1-33-38

1 Pendahuluan

Secara global, industri yang semakin berkembang khususnya bidang pariwisata memiliki peran penting pada suatu negara [1]. Bersamaan dengan perkembangan internet, beberapa situs web dan media sosial banyak ditemukan opini publik tentang topik tertentu yang dapat mempengaruhi pandangan dan sentiment terhadap topik tersebut [2]. Banyaknya sentiment yang beragam membuat pengunjung akan merasa bingung apakah akan mengunjungi tempat tersebut atau tidak [1]. Analisis sentiment adalah cara sederhana untuk menilai opini atau sentiment pada sebuah kalimat yang bertujuan untuk mengetahui pendapat seseorang terhadap topik tertentu [3]. Salah satu kegunaan analisis sentiment adalah untuk mengetahui kepuasan pelanggan terutama dalam industri perhotelan. Dimana kepuasan pelanggan menjadi kunci keberhasilan dari operasional hotel [4].

Analisis sentimen dalam [1; 3; 5; 6] dapat membantu untuk menentukan apakah suatu kalimat / frase bernilai sentiment positif atau sentimen negatif. Pada penelitian yang

* Corresponding author.

© Munir. A.S, dkk.;
licensed under Creative Commons License CC-BY
Jurnal Open Access
Yayasan Lentera Dua Indonesia

dilakukan oleh Afzall dkk [1] menyajikan kerangka klasifikasi sentiment berbasis aspek yang telah diimplementasikan menjadi aplikasi seluler sehingga dapat membantu wisatawan menemukan hotel dan restoran dalam suatu kota. Penelitian oleh Jaman dkk [3] melakukan perbandingan dengan beberapa metode dengan seleksi fitur menggunakan TF-IDF sehingga menghasilkan sentiment terhadap pengguna ojek online berdasarkan komentar. De Godoi dkk [5] menggunakan N-Gram dan TF-IDF untuk ekstraksi Twitter dan menggunakan *support vector machine* (SVM) untuk proses klasifikasi sehingga dapat menghasilkan sentimen dengan akurasi akurasi 63,93% hingga 81,06%. Penelitian yang dilakukan oleh Imamah dkk [6] melakukan analisis sentiment data Covid-19 untuk mengetahui kesehatan mental masyarakat dunia dengan membagi sentimen netral, positif, dan negatif dengan menggunakan metode klasifikasi Logistic Regression dengan pembobotan TF-IDF yang mendapatkan hasil akurasi klasifikasi sentimen tweet covid-19 sebanyak 94,71%.

Berdasarkan dari penelitian yang sudah ada, dalam penelitian ini akan melakukan analisis sentiment menggunakan ekstraksi fitur TF-IDF (*term frequency-inverse document frequency*) berdasarkan komentar atau review pada situs web Traveloka khususnya review hotel. Data review hotel yang diambil antara Juli-Agustus 2020 tersebar dalam tiga kota yaitu Jakarta, Bandung, dan Yogyakarta dengan jumlah data 9.999 review, dimana 5000 review negatif dan 4999 review positif. Dataset yang diambil kemudian diberikan nilai label dengan keterangan 1 untuk positif dan 0 untuk negatif. Selanjutnya, dibandingkan hasil ekstraksi fitur TF-IDF dengan Unigram dan Bigram. Terakhir, menggunakan SVM dan Naïve Bayes untuk proses klasifikasi. Sehingga pada hasil akhir didapatkan sebuah akurasi antara metode yang diusulkan dengan metode pembanding.

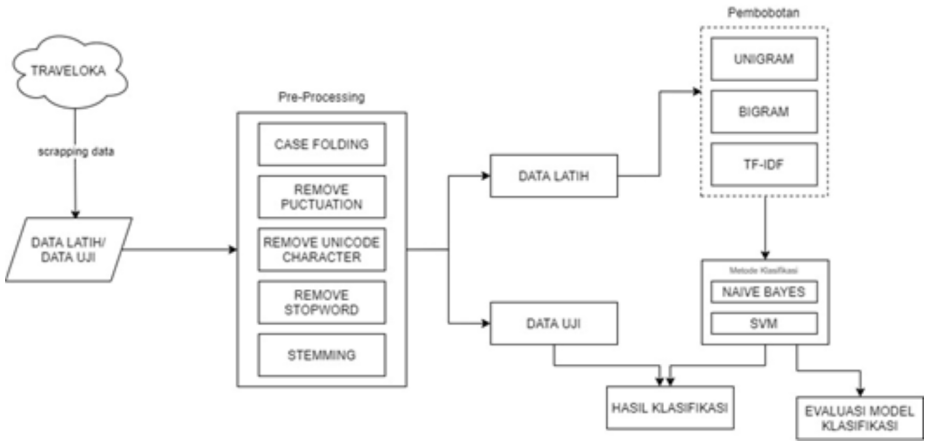
2 Metodologi

Pada penelitian ini dilakukan analisis terhadap perbandingan algoritma pembobotan TF-IDF, Bigram dan Unigram dengan menggunakan algoritma klasifikasi Naïve Bayes dan SVM. Data yang digunakan dalam penelitian ini yaitu data hasil web *scraping* review hotel di Traveloka periode bulan Juli-Agustus 2020. Review hotel hanya diambil dari 3 kota yang berbeda yaitu Jakarta, Bandung dan Yogyakarta. Jumlah data review yang digunakan dalam penelitian ini yaitu 9999 review, yang mana 5000 dari data tersebut bernilai negatif dan 4999 data bernilai positif. Dalam penelitian ini ada 4 tahapan utama yang akan dilakukan, yaitu Scraping data dari situs Traveloka, preprocessing data, pembobotan dan klasifikasi. Proses penelitian dapat dilihat pada Gambar 1.

Web *scraping* data dalam penelitian ini dilakukan menggunakan parsing html yang dibantu dengan aplikasi Webharvy, kemudian data yang didapatkan dari hasil web scraping disimpan ke dalam aplikasi Microsoft excel. Tahapan *preprocessing* yang dilakukan dalam penelitian ini ada 5 tahap. *Case folding*, *remove punctuation*, *remove stopword*, dan *stemming*.

Tahap *preprocessing* yang pertama kali dilakukan yaitu *case folding*. Pada tahap ini teks review yang diketik dengan menggunakan huruf kapital akan diubah ke dalam huruf kecil semua atau biasa disebut dengan *lowercase*. Remove punctuation adalah salah satu tahapan *preprocessing* yang dilakukan dengan penghapusan tanda baca pada teks, seperti tanda tanya, tanya seru, titik, koma, dan lain sebagainya. Penghapusan ini dilakukan karena tanda baca dianggap tidak memberikan pengaruh terhadap pemprosesan kata.

Penghapusan *stopword* adalah tahapan preprocessing yang dilakukan untuk penghapusan kata penghubung yang sering diabaikan dalam pemprosesan text, seperti “dan”, “atau”, “tapi”, dan lainnya yang memiliki frekuensi kemunculan yang tinggi . Kata-kata tersebut nantinya akan disimpan ke dalam stop list. Dalam penelitian ini, tahapan *preprocessing*



Gambar 1 Alur proses penelitian (*flowchart*)

stopword removal dilakukan dengan memanfaatkan *library* Sastrawi yang ada pada Python. Tahap *preprocessing* yang terakhir yaitu *stemming* dilakukan dengan menghapusan seluruh imbuhan kata, meliputi kata depan(*prefixes*), kata sisipan(*infixes*), dan akhiran kata(*suffixes*). Mengubah setiap kata menjadi kata dasarnya. Contoh: "lebersihan" diubah menjadi kata dasarnya yaitu "bersih".

Bigram merupakan salah satu konsep $n-$gram untuk metode pembobotan yang banyak digunakan dalam *natural language processing* (NLP) [7]. Konsep yang digunakan yaitu urutan persebelahan kata-kata n [7]. Contoh kata bigram yang memiliki empat kata "Hotel mewah design elegant". Dari contoh kata tersebut ada $(n-1)$ kombinasi kata bigram yang bisa digunakan, seperti: hotel, mewah, mewah, design, design, elegant, dengan n merupakan jumlah kata. Unigram merupakan salah satu algoritma untuk pembobotan yang termasuk dalam konsep $n-$gram. Perbedaan mendasar antara bigram dan unigram yaitu bigram menggunakan gabungan dua kata sedangkan unigram menggunakan 1 kata. TF-IDF merupakan salah satu teknik pembobotan di dalam NLP yang digunakan untuk melakukan evaluasi terhadap pentingnya suatu kata dalam sebuah dokumen.

Naïve bayes adalah salah satu algoritma klasifikasi supervised learning. Pada penelitian ini algoritma tersebut digunakan untuk mengklasifikasikan review negatif dan positif. Data yang diklasifikasikan menggunakan algoritma ini adalah data dari hasil pembobotan yang telah dilakukan sebelumnya. Naïve Bayes termasuk algoritma pengklasifikasian sederhana dengan menghitung probabilitas keanggotaan suatu kelas [8].

Dalam penelitian ini selain menggunakan algoritma Naïve Bayes tetapi juga menggunakan algoritma SVM, algoritma SVM termasuk ke dalam algoritma pembelajaran yang terawasi [5]. Data yang akan diklasifikasikan oleh algoritma SVM ini adalah data text yang sebelumnya telah dilakukan pembobotan. Output dari klasifikasi ini yaitu review negatif dan positif. Proses kerja algoritma SVM adalah dengan mencari nilai hyperplane terbaik [9].

3 Hasil dan pembahasan

Pada penelitian ini data yang digunakan didapatkan dari metode scraping website dari website traveloka.com. Data yang diambil adalah data review hotel dari 3 buah kota yang berbeda yaitu Jakarta, Bandung dan Yogyakarta. Data tersebut kemudian dilakukan proses pelabelan manual. label 1 untuk review yang bernilai positif dan label 0 untuk review yang

bersifat negatif. Data yang sudah melewati proses pelabelan kemudian dilakukan proses preprocessing yang mengacu seperti yang dijelaskan pada Gambar 1. Kemudian setelah dilakukan proses *preprocessing*, data selanjutnya diklasifikasikan menggunakan metode - metode klasifikasi.

Proses *scraping* data dilakukan dengan menggunakan bantuan aplikasi Webharvy. Proses ini menghasilkan data sebanyak 9999 data yang kemudian disimpan ke dalam format Excel sehingga dapat diproses secara luring. Data diambil dari review hotel di 3 kota yaitu Jakarta, Bandung dan Yogyakarta. Untuk menghapus bagian - bagian data yang tidak penting maka dilakukan proses teks *preprocessing*. *Preprocessing* ini dilakukan untuk menghapus karakter atau kata- kata yang tidak memiliki arti yang tidak signifikan pada bidang NLP. Proses *case folding* dilakukan dengan menggunakan bahasa Python. Proses *case Folding* menggunakan *function lower* yang terdapat di dalam bahasa Python tersebut. Simulasi dan hasil case folding dapat dilihat pada Tabel 1.

Tabel 1 Simulasi proses *case folding*

Sebelum *case folding*	Sesudah *case folding*
Sangat baik untuk penginapan keluarga dengan fasilitas kolam renang dan taman bermain anak dengan menu sarapan yang beragam.	sangat baik untuk penginapan keluarga dengan fasilitas kolam renang dan taman bermain anak dengan menu sarapan yang beragam

Proses *remove punctuation* dilakukan dengan cara menghapus semua tanda baca yang ada pada kalimat atau dokumen. Hal ini dimaksudkan untuk membuang tanda baca yang kurang berpengaruh terhadap pemrosesan kalimat pada penelitian. Proses selanjutnya adalah *remove stopword* dilakukan dengan menggunakan library Sastrawi yang telah mendukung pemrosesan Bahasa Indonesia. Simulasi dari proses *remove stopword* dapat dilihat pada Tabel 2.

Tabel 2 Simulasi proses *remove stopword*

Sebelum *remove stopword*	Sesudah *remove stopword*
sangat baik untuk penginapan keluarga dengan fasilitas kolam renang dan taman bermain anak dengan menu sarapan yang beragam	sangat baik penginapan keluarga fasilitas kolam renang taman bermain anak menu sarapan beragam

Seperti tampak pada Tabel 2, kata “dan” dan “untuk” dihapus karna kata tersebut dikategorikan sebagai kata penghubung yang tidak memiliki arti. Library Sastrawi yang juga mendukung proses *stemming* bahasa Indonesia digunakan dalam penelitian ini. Proses *stemming* dilakukan untuk membuat kata yang terdapat pada suatu kalimat atau dokumen menjadi kata dasar. Kata dasar didapatkan dengan menghilangkan semua tambahan-tambahan imbuhan yang ada pada kata tersebut seperti meng-, me-, kan-, di- , i, pe, peng-, a-, dan lain - lain. Simulasi proses *stemming* dapat dilihat pada Tabel 3. Pada Tabel 3 terlihat bahwa proses *stemming* mengembalikan kata yang memiliki imbuhan menjadi kata dasarnya seperti kata “penginapan” menjadi “inap”, kata “bermain” menjadi “main” dan lain sebagainya.

Proses pembobotan dilakukan dengan menggunakan tiga buah metode yaitu bigram, unigram dan TF-IDF. Metode tersebut di implementasikan kepada data dengan menggunakan *library* Scikit-learn pada bahasa Python. Proses ini dilakukan untuk memberikan nilai atau

Tabel 3 Simulasi proses *stemming*

Sebelum *stemming*	Sesudah *stemming*
sangat baik penginapan keluarga fasilitas kolam renang taman bermain anak menu sarapan beragam	sangat baik inap keluarga fasilitas kolam renang taman main anak menu sarap agam

bobot kepada kata atau kumpulan kata sesuai dengan perhitungan dari setiap metode. Proses klasifikasi dilakukan menggunakan bahasa Python menggunakan dua buah metode yaitu Naïve Bayes dan SVM. Pembagian datanya adalah 70 % dari data digunakan untuk proses training dan 30 % digunakan untuk proses testing. Hasil dari proses klasifikasi tersebut dapat dilihat pada Tabel 4.

Tabel 4 Sebaran dataset

Metode	bigram	unigram	TF-IDF
Naïve Bayes	88%	88%	88%
SVM	88%	94%	95%

Hasil akurasi tertinggi didapatkan oleh metode pembobotan TF-IDF yang digabungkan dengan metode klasifikasi SVM sebesar 95%. Akurasi terendah didapatkan oleh metode bigram dan unigram yang digabungkan dengan metode Naïve Bayes dan juga metode Bigram dengan metode klasifikasi SVM yaitu sebesar 88%.

4 Kesimpulan dan saran

Berdasarkan penelitian di atas, dapat disimpulkan bahwa metode TF-IDF digabungkan dengan metode SVM menghasilkan akurasi terbaik sebesar 95% diikuti dengan metode unigram digabungkan dengan metode SVM yaitu sebesar 94%. Metode yang digunakan pada penelitian ini dapat digunakan juga untuk data sentimen lainnya seperti data Twitter, data review e-commerce dan lain sebagainya.

Pustaka

1 M. Afzaal, M. Usman, and A. Fong, "Tourism mobile app with aspect-based sentiment classification framework for tourist reviews," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 2, pp. 233–242, 2019.

2 A. N. Farhan and M. L. Khodra, "Sentiment-specific word embedding for indonesian sentiment analysis," in *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. IEEE, 2017, pp. 1–5.

3 J. H. Jaman and R. Abdulrohman, "Sentiment analysis of customers on utilizing online motorcycle taxi service at twitter with the support vector machine," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*. IEEE, 2019, pp. 231–234.

4 M. Li, Y. Ma, and P. Cao, "Revealing customer satisfaction with hotels through multi-site online reviews: A method based on the evidence theory," *IEEE Access*, vol. 8, pp. 225 226–225 239, 2020.

5 J. de Godoi Brandão and W. P. Calixto, "N-gram and tf-idf for feature extraction on opinion mining of tweets with svm classifier," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2019, pp. 1–5.

6 F. H. Rachman *et al.*, "Twitter sentiment analysis of covid-19 using term weighting tf-idf and logistic regresion," in *2020 6th Information Technology International Seminar (ITIS)*. IEEE, 2020, pp. 238–242.

7 T. Fahrudin, J. L. Buliali, and C. Fatichah, "Ina-bwr: Indonesian bigram word rule for multi-label student complaints," *Egyptian Informatics Journal*, vol. 20, no. 3, pp. 151–161, 2019.

8 A. Ordonez, R. E. Paje, and R. Naz, "Sms classification method for disaster response using naïve bayes algorithm," in *2018 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, 2018, pp. 233–236.

9 A. Perdana, "Penerapan algoritma support vector machine (svm) pada pengklasifikasian penyakit kejiwaan skizofrenia (studi kasus: Rsj. radjiman wediodiningrat, lawang)," Ph.D. dissertation, Universitas Brawijaya, 2018.