OPEN ACCESS
J. OF INISTA, VOL. 3, NO. 2, PP.021-029, MEI 2021

ISSN : 2622-8106 (ONLINE)
DOI: 10.20895/INISTA.V3I2

Journal of Informatics, Information System, Software Engineering and Applications

Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR)

Boma Bayu Baskoro[#1], Irwan Susanto[#2], Siti Khomsah[#3]

#Prodi Informatika[1,2], Prodi Sains Data[3], Insitut Teknologi Telkom Purwokerto
Jl. D.I Panjaitan 128 Purwokerto, Jawa Tengah, Indonesia

[1]bomabayu789@gmail.com
[2]irwansusanto_yk@ittelkom-pwt.ac.id
[3]siti@ittelkom-pwt.ac.id

accepted on 10-06-2021

Abstract

Aplikasi e-tourism di Indonesia sudah banyak diterapkan terutama untuk layanan akomodasi wisata seperti hotel atau penginapan. Salah satu aplikasi e-tourism yang terkenal adalah tripadvisor.co.id. Aplikasi tersebut memudahkan masyarakat memesan hotel secara online karena lebih cepat, praktis dan mudah. Salah satu faktor penting dalam memilih hotel terbaik dengan harga terjangkau ialah pendapat para pelanggan hotel dari ulasan pada kolom komentar dari para pelanggan hotel sebelumnya. Banyaknya data ulasan pelanggan membutuhkan waktu yang lama untuk mengetahui polaritas ulasan positif dan mana ulasan negatif secara manual. Oleh karena itu diperlukan model analisis sentimen yang akurat yang dapat mengklasifikasikan ulasan pelanggan menjadi ulasan positif dan negatif. Pada penelitian ini diusulkan model analisis sentimen pelanggan hotel menggunakan metode *Random Forest Classifier* dan *Term Frequency–Inverse Document Frequency* (TF–IDF). Dataset yang digunakan untuk membangun model sentimen analisis adalah data komentar-komentar pelanggan hotel di Purwokerto yang diunduh dari situs tripadvisor.co.id. Pada *preprocessing* melibatkan proses konversi slangword menjadi kata baku sesuai KBBI, *stemming*, dan menambahkan kata-kata *stopword* baru selain *stopword* dalam library sastrawi. Hasil penelitian menunjukkan akurasi model mencapai akurasi 87,23%. Akan tetapi jika tanpa proses *stemming*, akurasi model hanya 87,01%.

Keywords: *analisis sentimen, ulasan hotel, purwokerto, Random Forest*,TF-IDF

I. PENDAHULUAN

Bisnis perhotelan atau penginapan menggunakan platform digital berkembang pesat dengan berbagai model bisnisnya. Pada tahun 2017, jumlah reservasi kamar hotel melalui jaringan internet telah mencapai lebih dari 65 persen, dan diperkirakan akan terus meningkat. Purwokerto sebagai kota bisnis dan pelajar menjadi barometer perhotelan di wilayah Jawa Tengah bagian barat. Data Badan Pusat Statistik (BPS) menunjukkan terdapat 182 hotel besar dan kecil di Purwokerto[1]. Data jumlah hotel juga dapat diakses

melalui bebagai situs e-tourism seperti situs web www.tripadvisor.co.id. Beberapa alasan kenapa masyarakat lebih menyukai reservasi hotel melalui jaringan internet, karena lebih menghemat waktu dan tenaga, lebih mudah membandingkan harga karena banyak pilihan, dan adanya testimoni pelanggan lain berupa ulasan yang ditinggalkan pada situs seperti www.tripadvisor.co.id.

Ulasan online memberikan informasi bagaimana kesan, penilaian dan pengalaman para pelanggan hotel. Penilaian biasanya berbentuk rating dan pengalaman ditulis dalam bentuk komentar maupun saran. Kualitas layanan hotel pada umumnya menggunakan rating dan saran dari pengguna hotel. Meskipun metode rating banyak digunakan dalam pengukuran layanan hotel namun rating tidak mampu menangkap pesan kekurangan yang dikeluhkan oleh para pelanggan. Sebagai contoh pelanggan memberikan rating yang rendah tapi tidak diketahui kekurangan layanannya pada fitur apa saja. Saran merupakan metode pelengkap rating dalam menjaring opini pelanggan atas layanan hotel. Analisis sentimen sudah digunakan untuk mengelola saran melalui polaritas saran/opini menjadi positif atau negatif. Penentuan polaritas sentimen ini dapat dilakukan secara manual tetapi seiring bertambahnya waktu maka meningkat pula data opini dari masyarakat sehingga membutuhkan usaha yang lebih lama untuk memberikan label polaritas secara manual ke opini tersebut.

Random Forest merupakan salah satu metode klasifikasi sudah banyak digunakan untuk mengetahui polaritas komentar pelangan secara otomatis. Random Forest adalah teknik pembelajaran *ensemble* berdasarkan algoritma Decision Tree (D-Tree), yang terdiri dari beberapa pohon keputusan sebagai *classifier*. Kelas yang dihasilkan dari proses klasifikasi diambil dari keptusan kelas terbanyak yang dihasilkan oleh pohon-pohon keputusan yang ada. Dengan melakukan *voting* keputusan dari pohon-pohon keputusan yang tersedia membuat akurasi dari Random Forest meningkat. Random Forest merupakan salah satu algoritma klasifikasi dengan tingkat akurasi yang baik. Random Forest lebih banyak diterapkan untuk klasifikasi dalam beberapa tahun terakhir sejak kinerja dari jenis algoritma ini melampaui SVM, Naïve Bayes dan algoritma pembelajaran mesin lainnya [2].

Penelitian ini dimaksudkan menerapkan metode Random Forest untuk analisis sentimen pelanggan hotel di Purwokerto. Untuk menjawab persoalan yang akan dibahas tersebut maka dibutuhkan metode transformasi data komentar pelanggan dalam bentuk tekstual ke bentuk kuantitas yaitu metode *Term Frequency–Inverse Document Frequency* (TF–IDF). Tahapan pra-pemrosesan teks (*text preprocessing*) merupakan salah satu tahapan penting untuk mendapatkan model sentimen yang akurat karena teks komentar biasanya mengandung banya noise seperti slang word, tanda baca, kata tidak baku, singkatan, emoticon, dan sejenisnya sehingga diperlukan metode prapemrosesan yang tepat[3]. Oleh karena itu, penelitian ini menggunakan proses konversi slang word, memperkaya jumlah *stopword* selain yang sudah ada di library sastrawi, dan proses *stemming*. Keluaran yang direncakan berupa sebuah model analisis sentimen dengan kelas positif dan negatif. dengan pelabelan menurut rating.

II. KAJIAN PUSTAKA

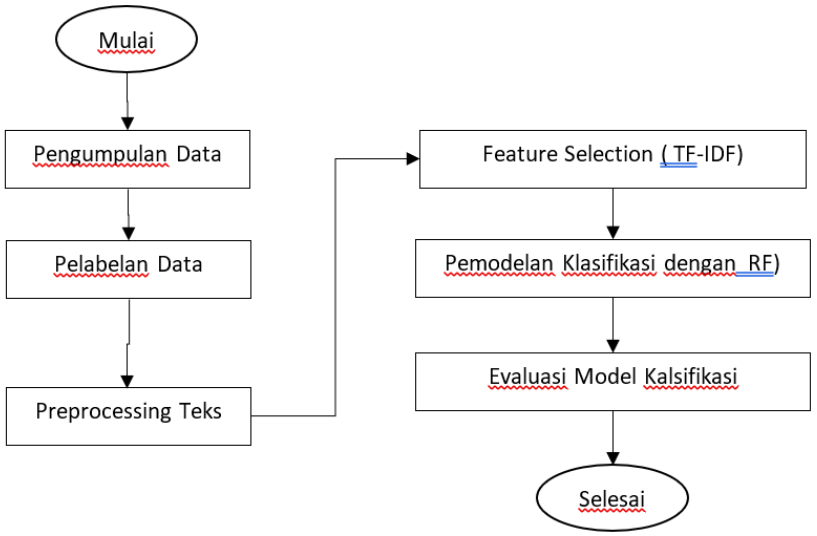
Beberapa penelitian sebelumnya yang terkait analisis sentimen yang berasal dari komentar pelanggan. Sujata Khedka menggunakan pendekatan berbasis fitur Linguistik untuk menyeleksi kalimat komentar menggunakan metode pemilihan fitur hybrid, tujuannya dalam mengklasifikasikan kalimat komentar sebagai pujian atau keluhan. Kalimat pujian dan keluhan ini dapat dianalisis lebih lanjut oleh bisnis organisasi untuk mengidentifikasi alasan kepuasan atau ketidakpuasan pelanggan [4]. Data- data ulasan pelanggan sangat mudah didapatkan dari internet. Pada 2016, Sipayung, dkk meneliti model analisis sentimen pelanggan hotel dari data 675 komentar terhadap Grand Royal Panghegar Bandung di situs Tripadvisor. Metode yang yang digunakan adalah Naïve Bayes Classifier, dan hasil akurasi model tersebut 75.42% [5]. Naïve Bayes bekerja dengan cara menghitung probabilitas kata unik yang muncul dalam komentar. Pada 2020, Pramitha, dkk membuat penelitian model sentimen pelanggan hotel di Purwokerto menggunakan Naïve Bayes yang dioptimasi menggunakan Particle Swarm [6]. Model sentimen yang dibangun dari 874 data komentar di www.tripadvisor.co.id dan www.agoda.com. Model yang dibangun berhasil mencapai akurasi 81,14%. Lamiaa menganalisis sentimen pengguna dari 111.986 komentar terhadap hotel-hotel di dua kota besar di China. Data diambil dari Tripadvisor, Booking, Expedia, dan Triage. Metode SVM, Naïve Bayes, dan

decision tree (D-Tree) digunakan secara terpisah tujuannya untuk melihat metode yang paling akurat. Hasilnya, SVM mencapai akurasi 79%, Naïve Bayes 85%, dan D-Tree 71 % dengan masin-masing presisi berturut- turut 61 %, 76%, dan 51 % [7]. Penelitian Lamiaa ini menunjukkan D-Tree sangat rendah akurasinya. D-Tree merupakan machine learning yang bekerja dengan cara pohon keputusan biner. Ada banyak cara untuk meningkatkan akurasi D-Tree salah satunya dengan metode *ensemble*. Random Forest (RF) adalah metode *classifier ensemble* berbasis D-Tree.

Random Forest bekerja dengan cara berbeda dari D-Tree. D-Tree hanya membangun model klasifikasi dengan satu pohon saja sehingga akan menghasilkan pohon keputusan yang cukup besar. Berbeda dengan D-Tree, RF membangun banyak pohon keputusan sehingga menghasilkan pohon yang kecil- kecil namun banyak, hal ini membuat kinerja RF lebih baik dalam hal akurasi [8]. Hedge dan Padma menerapkan RF untuk meningkatkan *sentiment analyzer* (SA) pada dokumen Kannada yaitu komentar produk online [9]. RF berhasil meningkatkan akuras SA sebelumnya meskipun nilai akurasinya berkisar 65 -72%. Kedua penelitian tersebut menggunakan TF-IDF untuk vektorisasi teks komentar. Namun mengkombinasikan RF dapat juga dikombinasikan dengan word2Vec digunakan untuk model analisis sentimen, contohnya untuk model analisis sentimen pada komentar YouTube[10], hasil percobaan menunjukkan akurasi cukup tinggi yaitu 88.77% sampai 89.05%. Model tersebut dibangun dengan 31947 data komentar terhadap acara debat calon presiden di 2019. Namun, Word2Vec merupakan deep learning sehingga mempunya cara kerja berbeda dengan vektorisasi TF-IDF. Berdasarkan penelitian-penelitian tersebut maka peneliti mengajukan model analisis sentimen dengan kombinasi metode Random Forest dan vektorisasi TF-IDF. Tujuannya adalah melihat berapa akurasi RF dengan vektorisasi fitur menggunakan TF-IDF pada model sentimen analisis pelanggan hotel di Purwokerto.

III. METODOLOGI PENELITIAN

Metodologi penelitian yang digunakan melibatkan beberapa tahapan dan metode, seperti ditunjukkan oleh diagram alir pada Gambar.1.



Gambar 1. Diagram Alir Penelitian

1. Pengumpulan Data

Langkah pertama dalam metodologi penelitian ini adalah pengumpulan data dari situs web tripadvisor menggunakan alat bantuan tool *webharvy* yang secara otomatis dapat mengikis data dari halaman web dan menyimpan konten diekstrak dalam format *.csv. *Webharvy* bekerja dengan mengakses halaman web, memilih elemen data, mengekstrak dan menyimpannya ke dalam dataset terstruktur. Pada penelitian ini menggunakan *webharvy* untuk *crowling* data web www.tripadvisor.co.id dengan elemen keyword “hotel purwokerto”. Dengan mengetik Purwokerto dan enter pada pencarian web tripadvisor dan memilih kolom

hotel, akan memunculkan hotel yang berada di Purwokerto dan sekitarnya. Kemudian crawling data komentar menggunakan *webharvy* dengan memilih komentar pada tiap hotel. Data yang berhasil diunduh sebanyak 1166 komentar dari berbagai jenis hotel.

2. Pelabelan Data

Model analisis sentimen pada penelitian ini merupakan permasalahan *Supervised Learning*. *Supervised Learning* dibentuk dari dataset yang memiliki target atau label. Pada kolom komentar ulasan pelanggan hotel di TripAdvisor.co.id, selain menuliskan ulasan, pelanggan juga dapat memberikan *rating* pelayanan 10 sampai 50. Pelabelan kelas data positif akan dilakukan dengan cara memilah ulasan berdasarkan nilai *rating* 40 dan 50 dan ulasan kelas data negatif berdasarkan *rating* 10, 20, dan 30. Penelitian ini tidak mennggunakan label kelas netral karena ulasan cenderung mengandung kata-kata sentimen positif atau negatif.

Pelabelan data ulasan pelanggan hotel menjadi kelas positif dan kelas negatif dilakukan secara otomatis didalam bagian *subprogram,* yang mengacu pada rating ulasan menggunakan fungsi *if-else* menggunakan sintak program pada Gambar 2.

```
def label_data(dataset):
    rows = dataset
    labels = []
    for cell in rows['Rating']:
        if cell >= 40:
            labels.append('Positif')   #Positif
        else:
            labels.append('Negatif')   #Negatif
    rows['Label'] = labels
return rows
```

Gambar 2 Kode Program untuk Fungsi Pelabelan

3. Preprocessing Teks

Preprocessing Teks *(Pre-processing* dilakukan untuk membentuk dataset yang siap dianalisis. Pada penelitian ini, tahap *pre-processing* dilakukan dengan 3 langkah, yaitu :

Pertama data cleansing adalah membersihkan data dari noise seperti tanda baca dan karakter lain yang tidak penting. Data cleansing dimulai pengubahan komentar menjadi berhuruf kecil semua. Sebuah kata terdiri dari huruf-huruf dengan berbagai kasus seperti huruf besar dan huruf kecil. Untuk membakukan huruf besar huruf, semua huruf dikonversi menjadi huruf kecil (case folding). Mengubah komentar juga meningkatkan konsistensi data.

Kedua *stopword* adalah kosakata yang bukan kata-kata unik dari suatu dokumen. Contohnya adalah "di", "oleh", "pada", "sebuah", "karena", dll. *Stopword*s akan dihapus untuk meningkatkan kinerja analisis sentimen.

Ketiga Tokenizing adalah proses pemotongan sebuah dokumen menjadi bagian-bagian, yang disebut dengan token (memisahkan kata, simbol, frase, dan entitas penting lainnya).

4. Feature Selection

Algoritma yang digunakan dalam seleksi fitur yaitu Term frequency inverse document frequency (TF-IDF). Metode TF-IDF merupakan metode pembobotan *term* secara statistik yang banyak digunakan sebagai vektorisasi pada teks analisis. Cara kerja TF-IDF yaitu mencari kata-kata unik atau relevan dalam komentar. TF-IDF akann mengukur berapa kali sebuah kata muncul dalam sebuah komentar dan keseluruhan komentar yang dijadikan dataset[2]. Proses seleksi fitur akan menghasilan token atau satuan kecil kalimat yaitu token-token beserta bobot TF-IDF-nya.

5. Model Klasifikasi

Setelah dilakukan *preprocessing* maka sebuah input algoritma Random Forest diperoleh hasil TF-IDF dimana term yang dihasilkan merupakan atribut dari algoritma. Random forest sendiri merupakan algoritma ensenmble learning yang mana untuk memperoleh keputusan akhir akan dilakukan votting mayority. Membangun model klasifikasi untuk menentukan sebuah kalimat ulasan sebagai anggota kelas positif atau kelas negatif berdasarkan nilai perhitungan keputusan seacra *voting* dari rumus Random Forest (RF). Jika hasil *voting* kalimat tersebut positif, maka kalimat tersebut termasuk ke dalam positif. Namun jika hasil *voting* akhir pohon-pohon keputusan RF adalah adalah negatif, maka kalimat tersebut masuk ke dalam kelas negatif [2].

6. Evaluasi dan Hasil

Untuk mengetahui akurasi model klasifikasi sentimen maka diguankan *Confusion Matrix*. Pemodelan diawali dengan pembagian dataset menjadi data *training* dan data *testing*. *Training* bertujuan untuk membuat RF belajar berdasarkan data dan menghasilkan pola mana komentar yang masuk kelas positif atau negatif. Data *testing* digunakan untuk menguji seberapa besar akurasi model. Hasil pengujian dengan data *testing* akan menghasilkan matrik konfusi yaitu berupa nila True Positif (TP), True Negatif (TN), False Positif (FP), dan False Negatif (FN) [6], sehingga akurasi RF pada pemodelan analisis sentimen menggunakan data hotel di Purwokerto dapat diketahui.

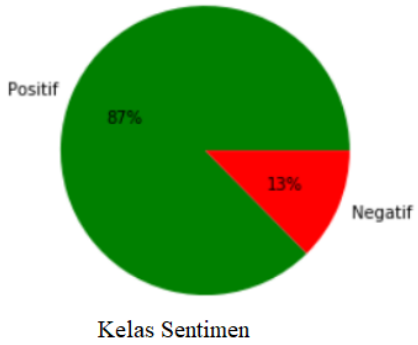
IV. HASIL DAN PEMBAHASAN

Data yang telah diberikan label akan mempunyai kelas positif atau negatif. Contoh hasil pelabelan terdapat pada Gambar 2.

	Description	Rating	Label
0	Selama perjalanan liburan keluarga yang paling...	50	Positif
1	Terletak di Lokasi yang strategis dan mudah di...	50	Positif
2	pelayanan sangat baik,ramah dan nyaman nginap ...	50	Positif
3	Pertama kali minep di aston pelayanannya ramah...	40	Positif
4	sangat bagus, rekomendasi, pemandangan bagus p...	50	Positif
...	...	...	...
1161	Hotel lumayan... hotel tua.. perlu peremajaan....	30	Negatif
1162	Lokasi yang strategis di puncak bukit dengan r...	30	Negatif
1163	Hotel ini memiliki pemandangan yang indah gunu...	40	Positif

Gambar 2 Hasil Pelabelan Data

Proporsi jumlah data yang masuk kelas positif dan negatif ditunjukkan oleh Gambar 3. Jika dilihat dari jumlah kelas, terlihat kedua kelas sentimen termasuk *imbalanced* (tidak seimbang).



Gambar 3. Proporsi Kelas Sentimen Hasil Pelabelan

Tahapan *text-preprocessing* dilakukan untuk perbaikan data dari *noise*. *Text pre-processing* memiliki peran penting pada saat melakukan pemodelan analisis sentimen dikarenakan keadaan dari teks mempengaruhi hasil akurasi. Terdapat 3 tahapan *text-preprocessing* yang dilakukan yaitu *cleansing*, *stopword-removing*, dan *tokenizing*. Ketiga tahapan ini dilakukan secara berurutan.

Proses *cleansing* dilakukan untuk mengubah seluruh teks data ulasan menjadi huruf kecil atau *lowercase*, membuang karakter selain huruf, dikarenakan data ulasan pelanggan tidak konsisten dalam penggunaan huruf kapital dan terdapat angka, tanda baca, maupun *emoticon* seperti contoh Tabel 1. Tahap kedua *stopword-removing* menggunakan library Sastrawi 1.0.1. *Stopword-removing* yaitu membuang kata-kata yang tidak mempunyai sentimen seperti contoh Tabel 1. Misalnya kata sambung dan, dengan, yang, dan sejenisnya. Selain itu juga membuang kata bukan sambung yang tidak bersentimen seperti kata benda, nama orang, nama jalan, nama hotel, kata ganti waktu, atau keterangan tempat. Tahap terakhir *tokenizing* yaitu mengurai kalimat komentar menajdi kata per kata, seperti contoh Tabel 2.

Tabel 1. Contoh Hasil *Cleansing* dan *Stopword*-Removing

Ulasan	Hasil *Cleansing* dan *Stopword*
Selama perjalanan liburan keluarga yang paling berkesan di Purwokerto,saya dan keluarga selalu setia dengan Aston Imperium Purwokerto,dari hotel hotel lain yang pernah kita singgahi,over all aston the best. dipagi hari sy dan keluarga bisa berenang dan melihat view Purwokerto yang bagus dari swimingpool. dilanjut breakfast yang sangat varian dan rasanya si bikin susah dilupain enaknya. kamarnya juga oke banget,bersih very comfortable. pelayanan yang super duper ramah.	selama perjalanan liburan keluarga paling berkesan keluarga selalu setia imperium hotel hotel pernah hari keluarga berenang melihat pemandangan bagus kolam renang sarapan sangat varian rasanya bikin susah enaknya banget bersih pelayanan super ramah
Terletak di Lokasi yang strategis dan mudah dijangkau... Hotel skala Internasional dengan standard fasilitas yang sesuai dengan kelasnya. Memiliki aneka pilihan meeting room yang dilengkapi venue yang sejuk manakala harus rehat sejenak... Hingga acara meeting tidak lagi membosankan...	terletak lokasi strategis mudah hotel skala internasional fasilitas sesuai memiliki aneka pilihan ruang rapat sejuk manakala rehat sejenak hingga acara membosankan

Tabel 2. Contoh Hasil *Tokenizing*

Hasil *Cleansing* dan *Stopword*	Hasil *Tokenizing*
selama perjalanan liburan keluarga paling berkesan keluarga selalu setia imperium hotel hotel pernah hari keluarga berenang melihat	'selama', 'perjalanan', 'liburan', 'keluarga', 'paling', 'berkesan', 'keluarga', 'selalu', 'setia', 'imperium', 'hotel', 'hotel', 'pernah', 'hari', 'keluarga', 'berenang',

Hasil *Cleansing* dan *Stopword*	Hasil *Tokenizing*
pemandangan bagus kolam renang sarapan sangat varian rasanya bikin susah enaknya banget bersih pelayanan super ramah	'melihat', 'pemandangan', 'bagus', 'kolam', 'renang', 'sarapan', 'sangat', 'varian', 'rasanya', 'bikin,' 'susah', 'enaknya', 'banget', 'bersih', 'pelayanan', 'super', 'ramah'
terletak lokasi strategis mudah hotel skala internasional fasilitas sesuai memiliki aneka pilihan ruang rapat sejuk manakala rehat sejenak hingga acara membosankan	'terletak', 'lokasi', 'strategis', 'mudah', 'hotel', 'skala', 'internasional', 'fasilitas', 'sesuai', 'memiliki', 'aneka', 'pilihan', 'ruang', 'rapat', 'sejuk', 'manakala', 'rehat', 'sejenak', 'hingga', 'acara', 'membosankan'

Tahapan vektorisasi ini mengubah data berupa kata(token) menjadi numerik dengan menggunakan TF-IDF. Dari perhitungan TF-IDF menghasilkan sejumlah fitur-fitur seperti pada Tabel 3 yang diekstrak dari 1166 komentar. Setelah pembobotan kata kedalam bentuk numerik, selanjutnya dapat dilakukan pemodelan analisis sentimen dengan algoritma klasifikasi RF. Model klasifikas analisis sentimen menggunakan algoritma Random Forest dibangun dengan tiga sekenario percobaan yaitu:

- Sekenario 1, dilakukan tahapan proses *stopword-removing* menggunakan sastrawi dan dilakukan proses *stemming* dan konversi *slang word*. *Stemming* adalah mengubah kata (token) menjadi kata dasar.
- Sekenario 2, dilakukan mengoreksi secara manual komentar negatif pada rating 30 dari rating 10-50 dan konversi *slang-word*, proses *stopword-removing* sastrawi, ditambah *stopword* yang dibuat sendiri berisi 2528 kata seperti pada lampiran daftar *stopword* tambahan yang ditentukan oleh peneliti. Pada percobaan dua tidak dilakukan *stemming*.
- Sekenario 3, dilakukan mengoreksi secara manual komentar negatif pada rating 30 dari rating 10-50, konversi *slang-word*, dan proses *stopword-removing* sastrawi ditambah *stopword* yang dibuat sendiri berisi kata lebih banyak dari pada sekenario dua sebesar 3628 kata, seperti pada lampiran *stopword* buatan. Pada sekenario tiga juga tidak dilakukan *stemming*.

Tahap pemodelan menggunakan proporsi data training 80% dan data *testing* 20%, hasil ketiga sekenario percobaan tersebut dapat dilihat pada Tabel 3.

Tabel 3. Akurasi Model Klasifikasi Random Forest

Keterangan	Sekenario 1	Sekenario 2	Sekenario 3
Jumlah data	1166	1166	1166
Data train	932	932	932
Data test	234	234	234
Jumlah fitur	4957	4134	3035
Akurasi	76,07%	87,01%	87,23%
Presisi	76%	87%	87%
Recall	100%	100%	100%

Metode klasifikasi kelas sentimen menggunakan algoritma Random Forest Classifier ini dianggap cukup baik. Tabel 3 menunjukkan akurasi pada algoritma Random Forest Classifier dipengaruhi oleh seberapa bersih dataset yang digunakan. Algoritma Random Forest bekerja dengan cara membangun beberapa pohon keputusan dan menggabungkannya untuk mendapatkan prediksi yang lebih akurat dan stabil. Dari hasil tiga sekenario percobaan menunjukkan bahwa *preprocessing* yang berbeda menghasilkan jumlah fitur kata unik berbeda-beda.

Pada tahap *text-preprocessing* khususnya pada tahap *stopword* yang menggunakan *library* Sastrawi masih menghasilkan hasil data kata yang ambigu dan belum bersih. Terdapat beberapa kata yang penting menjadi hilang seperti kata ‘tidak’ dan ‘kurang’. Perbaikan *stopword* kedalam daftar *stopword default* yang dimiliki sastrawi, juga berhasil meningkatkan akurasi, terlihat pada hasil sekenario 2. Bahkan semakin banyak *stopword* yang dikenali oleh model maka akurasi berpeluang semakin baik. Pengaruh jumlah *stopword* yang dikenali oleh model ditunjukkan oleh sekenario 3 dimana jumlah *stopword* lebih banyak dari sekenario 2.

Justru proses *stemming* pada dataset ini tidak berpengaruh pada peningkatan akurasi, hal ini ditunjukkan oleh hasil sekenario 1 dengan akurasi 87,23 %, lebih tinggi dari sekenario 1 dan 2.

V. KESIMPULAN

Dari penelitian yang sudah dilakukan, peneliti dapat menarik kesimpulan sebagai berikut :

1. Algoritma Random Forest dapat digunakan untuk membuat model klasifikasi sentimen pelanggan hotel di Purwokerto dengan memanfaatkan data komentar yang didapat dari situs online. Akurasi model mencapai 87,23% dengan tahapan mengoreksi secara manual komentar negatif pada rating 30 dari rating 10-50 dan preprocessing pembenaran/konversi kata slang word sesuai kata dalam KBBI dan penambahan daftar *stopword* yang ditambah *stopword* buatan sendiri. Akurasi ini didapat tanpa proses *stemming*.
2. *Stemming* justru tidak berkontibusi untuk menningkatkan akurasi.

DAFTAR PUSTAKA

[1] BPS, "Banyaknya Hotel Di Wilayah Kabupaten Banyumas Dirinci Per Kecamatan Tahun 2014," 2016. https://banyumaskab.bps.go.id/statictable/2016/03/28/73/banyaknya-hotel-di-wilayah-kabupaten-banyumas-dirinci-per-kecamatan-tahun-2014.html (accessed Jan. 01, 2020).

[2] S. Khomsah, A. F. Hidayatullah, and A. S. Aribowo, "Comparison of the Effects of Feature Selection and Tree-Based Ensemble Machine Learning for Sentiment Analysis on Indonesian YouTube Comments," 2021, pp. 161–172, doi: 10.1007/978-981-33-6926-9_15.

[3] S. Khomsah and A. S. Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Rekayasa Sistem dan Teknologi Informasi*, *RESTI*, vol. 4, no. 4, pp. 648–654, 2020, doi: 10.29207/resti.v4i4.2035.

[4] S. Khedkar and S. Shinde, "Linguistic Feature-Based Praise or Complaint Classification from Customer Komentars," in *International Conference on Intelligent Computing, Information and Control Systems (ICICCS 2019)*, 2019, pp. 470–481, doi: 10.1007/978-3-030-30465-2_52.

[5] I. Z. Evasaria M. Sipayung, Herastia Maharani, "Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem Informasi (JSI)*, vol. 8, no. 1, pp. 104–126, 2016.

[6] Elin Hanjani Pramitha; Siti Khomsah; Amalia Beladinna Arifa;, "Analisis Sentimen Pelanggan Hotel Di Purwokerto Menggunakan Naive Bayes Classifier dan Particle Swarm Optimization (Studi Kasus: Ulasan Pelanggan pada Situs Agoda dan Tripadvisor)," Banyumas, 2020.

[7] L. Mostafa, "Machine Learning-Based Sentiment Analysis for Analyzing the Travelers Komentars on Egyptian Hotels," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, 2020, pp. 405–413, doi: 10.1007/978-3-030-44289-7_38.

[8] A. S. Aribowo, H. Basiron, N. S. Herman, and S. Khomsah, "Fanaticism Category Generation Using Tree-Based Machine Learning Method," in *International Conference on Science & Technology (ICoST 2019)*, 2020, vol. 1501, no. 1, doi: 10.1088/1742-6596/1501/1/012021.

[9] Y. Hegde and S. K. Padma, "Sentiment Analysis Using Random Forest Ensemble for Mobile Product Komentars in Kannada," 2017, doi: 10.1109/IACC.2017.151.

[10] S. Khomsah, P. Studi, S. Data, J. Tengah, and I. Artikel, "Sentiment Analysis On YouTube Comments Using Word2Vec and Random Forest," vol. 18, no. 1, pp. 61–72, 2021, doi: 10.31515/telematika.v18i1.4493.