

Sentimen Analisis Opini Pembeli pada Aplikasi Shopee Berbasis N-Gram Lexicon

Erlina Halim¹, Arwin Halim², Andriana Sunjaya³, Novresia Sunjaya⁴

^{1,2,3,4}Universitas Mikroskil, Jl. Thamrin No. 112, 124, 140, Telp. (061) 4573767 Fax. (061) 4567789

^{1,2,3,4}Fakultas Informatika, Program Studi Teknik Informatika, Universitas Mikroskil, Medan

¹erlina.halim@mikroskil.ac.id, ²arwin@mikroskil.ac.id, ³191110601@students.mikroskil.ac.id,

⁴191110703@students.mikroskil.ac.id

Abstrak

Perubahan perilaku konsumen dalam melakukan pembelian di *e-commerce*, telah mengubah cara konsumen dalam memutuskan membeli barang. Konsumen akan menilai berdasarkan gambar, *rating* dan opini dari pembeli sebelumnya. Namun konsumen harus membaca semua opini untuk mendapatkan hasil yang akurat. Solusi yang ditawarkan adalah dengan menyediakan fitur pengolahan opini. Namun opini yang ditulis tidak terstruktur, terdapat kata yang tidak lengkap, *emoticon*, gambar, serta menggunakan kata tidak baku menjadi tantangan dalam mengolah opini. Corpus yang digunakan dalam penelitian ini diperoleh dari penelitian sebelumnya yang berjumlah 124.606 dengan atribut yang tersedia yaitu *itemname*, *date*, *starrate* dan opini. Penambahan corpus dilakukan dengan membangun program untuk menarik data opini sebanyak 95.731 sehingga total opini yang diproses sebanyak 220.337. Proses ekstraksi opini menggunakan pendekatan lexicon dengan peningkatan penambahan proses *opinion representation* dan *n-gram lexicon* untuk menentukan arah sentimen. Peningkatan proses *pre-processing* dengan penerapan *opinion representation* berbasis Bahasa Indonesia pada kasus *review e-commerce*. Peningkatan klasifikasi dengan memperhatikan semantik kata dengan menerapkan *n-gram*. Opini hasil klasifikasi menghasilkan kelas positif, netral, dan negatif. Klasifikasi opini akan dibandingkan dengan klasifikasi *rating* untuk mengetahui tingkat keakuratan. Hasil dari evaluasi diperoleh akurasi sebesar 78%. Terjadi peningkatan akurasi sebesar 2% dibanding dengan ekstraksi opini tanpa adanya proses *opinion representation* dan *n-gram*.

Kata kunci: Opini Konsumen, *Opinion Representation*, N-Gram Lexicon, Klasifikasi

Abstract

The way consumers choose to acquire items has altered consumer behavior when engaging in e-commerce. Consumers will evaluate based on product images, user reviews, and customer testimonials. To get reliable results, buyers should read all viewpoints. Offering features for opinion processing is the suggested solution. However, it might be difficult to interpret written comments since they lack structure and contain partial sentences. The 220,337-item corpus utilized in this study was used to process and included the attributes itemname, date, starrate, and opinion. The opinion extraction process uses a lexicon approach with an increase in the process of adding opinion representations and the n-gram lexicon to determine direction sentiment. Applying Indonesian-based opinion representation to e-commerce review instances improved the preparation procedure. To improve classification by paying attention to word semantics by applying n-grams. Opinion classification results produce positive, neutral, and negative classes. The opinion classification will be compared with the rating classification for accuracy. The accuracy of the evaluation's findings was 78%. The accuracy of the evaluation's findings was 78%. Compared to n-gram and opinion extraction without opinion representation, there is a 2% improvement in accuracy. Compared to opinion extraction without opinion representation and n-gram, there is a 2% improvement in accuracy.

Keywords: Consumer Opinion, Opinion Representations, N-Gram Lexicon, Classification

1. PENDAHULUAN

Kondisi COVID-19 yang melanda Indonesia sejak tahun 2019 sangat berpengaruh terhadap kehidupan bermasyarakat khususnya kondisi ekonomi, situasi ini mendorong masyarakat untuk menghindari transaksi secara langsung guna menghindari penyebaran virus. Serta masyarakat diharuskan menghindari kontak fisik (*physical distancing*) sehingga dalam melakukan transaksi pembelian, masyarakat beralih dari langsung menjadi *online* [1]. Peralihan tersebut mengubah perilaku konsumen dari melihat dan mengecek langsung produk yang akan dibeli, menjadi menilai berdasarkan gambar, *rating* dan opini dari pembeli sebelumnya. Namun untuk mendapatkan informasi yang akurat, konsumen harus membaca semua opini [2]. Sehingga *e-commerce* perlu memikirkan cara agar konsumen tidak perlu membaca semua opini dan konsumen dapat percaya serta memiliki niat untuk melakukan transaksi. Menurut model perilaku konsumen, faktor utama konsumen untuk melakukan pembelian adalah kepercayaan [1]. Kepercayaan dapat dibangun dengan mengolah opini. Salah satu pengolahan opini konsumen adalah ekstraksi opini yang akan menghasilkan informasi baru dan penting dari opini yang berisi pernyataan mengenai klaim tentang layanan atau produk yang diberikan oleh *e-commerce* kepada konsumen [3]. Dan menurut Tama dkk, efek opini memberikan pengaruh terhadap pembelian sebesar 73% - 87% [4].

Pada pengolahan opini, terdapat beberapa permasalahan, seperti kata yang tidak lengkap, *emoticon*, struktur bahasa yang tidak baku, penggunaan kata informal, gambar dan lain-lain [5] [6]. Untuk dapat memahami setiap kata yang telah dituliskan pada opini, dibutuhkan sebuah proses tambahan sebelum data tersebut digunakan. Salah satu pendekatan yang digunakan adalah memperhatikan semantik kata. Rizka dkk telah mencoba untuk membangun model padanan kata dengan menggunakan pendekatan Word2Vec untuk *review* berbahasa Indonesia [7]. Hasilnya berupa parameter dari Word2Vec untuk Bahasa Indonesia telah diperoleh dan memiliki akurasi yang lebih baik. Pada proses sentimen analisis, terdapat beberapa teknik *text mining* yang dapat digunakan, seperti pendekatan lexicon [2]. Pendekatan Lexicon terus berkembang sampai saat ini. Penelitian Dey dkk telah menerapkan n-gram pada lexicon pada opini pada *dataset movie* [8]. Hasilnya, terjadi peningkatan performa pada klasifikasi dibandingkan dengan pendekatan n-gram lainnya.

Penelitian ini mengusulkan peningkatan proses pada penelitian sebelumnya [2] dengan memperhatikan bobot kata yang memiliki semantik yang sama (*opinion representation*) dan kombinasi n-gram pada lexicon. Proses pemberian bobot kata yang memiliki semantik yang sama (*opinion representation*) akan berbasis pada Bahasa Indonesia dengan pendekatan Word2Vec. Kombinasi n-gram pada lexicon digunakan untuk meningkatkan hasil dari klasifikasi opini. Evaluasi akan menggunakan *confusion matrix* dengan menghitung nilai presisi, *recall*, akurasi dan *f1 score*.

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Penambahan opini melibatkan analisis opini, sentimen atau mentalitas penulis dari teks tertulis. Pendapat *online* memiliki pengaruh tidak langsung pada bisnis situs *e-commerce*. Situs-situs tersebut memasarkan produk mereka dan para pengguna *web* membaca ulasan produk sebelum membeli produk. Banyak organisasi menggunakan sistem penambahan opini untuk melacak ulasan pelanggan tentang produk yang dijual *online* [9].

Penambahan opini adalah sebuah cara untuk mempertahankan fokus pada beberapa tren bisnis terkait dengan administrasi transaksi, manajemen status dan juga iklan. Prediksi pola juga dilakukan dengan menggunakan pendapat pelanggan. Berbagai istilah yang digunakan dalam penambahan opini diberikan di bawah ini [9]:

1. Fakta: Kepastian adalah apa yang telah benar-benar terjadi atau memang demikian.
2. Opini: Perasaan adalah perspektif atau penilaian yang dibingkai tentang sesuatu, tidak begitu banyak dalam terang kebenaran atau informasi.
3. Kalimat Subyektif: Sebuah kalimat atau konten bersifat subjektif atau keras kepala jika itu benar-benar menunjukkan emosi seseorang.

4. Kalimat Target: Kalimat target menunjukkan beberapa aktualisasi dan data yang diketahui tentang dunia.
5. Hal: satu artikel atau unit, terutama yang merupakan bagian dari jadwal, pengumpulan, atau set.
6. Survei: Audit adalah konten yang mengandung susunan kata yang memiliki sentimen klien untuk hal tertentu. Sebuah survei mungkin subjektif atau objektif atau keduanya.
7. Aspek yang Diketahui: Sudut yang dikenal adalah sudut pandang standar yang diberikan oleh situs tertentu tempat klien memberikan penilaian secara independen.

2.2 Pendekatan Lexicon

Dalam pendekatan ini, kamus yang berisi daftar kata dan idiom yang dipakai untuk analisis dikenal dengan *opinion lexicons*. Kamus ini akan mengklasifikasi kata ke dalam kelas positif, negatif dan netral setelah melalui proses analisis. Berdasarkan pendekatan ini, kata yang termasuk sentimen adalah kata sifat sedangkan fiturnya kata benda. Berdasarkan cara pembobotan untuk *opinion lexicon*, terdapat dua pendekatan sebagai berikut [10]:

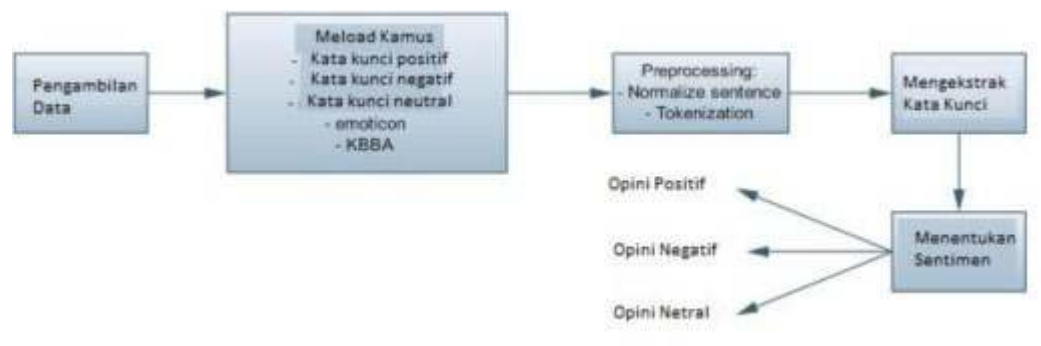
1. Pendekatan berbasis kamus kata

Dalam pendekatan ini daftar kata-kata sebelumnya digunakan untuk memetakan dengan sentimen yang diekstraksi. Daftar terus bertambah dengan mencari sinonim dari kata-kata. Kerugiannya adalah kamus mungkin tidak spesifik konteks atau terkait dengan domain maka kata opini yang diekstraksi mungkin tidak terlalu akurat.

2. Pendekatan berbasis semantik

Skor sentimen diprediksi secara langsung dengan menggunakan kesamaan antara kata-kata yang dihitung oleh algoritma yang berbeda. Kesamaan antara kata-kata dihitung oleh keterkaitan semantiknya, yang memberikan mereka skor yang sama. Misalnya, jika kata sifat memiliki skor dua, menurut lexicon sebagaimana disebutkan di atas kata benda juga mendapat skor.

Dalam analisis sentimen dikenal adanya metode berbasis *machine learning*, berbasis kamus (*lexicon-based*), atau gabungan keduanya. Secara umum alur proses analisis sentimen menggunakan *lexicon-based* ditunjukkan oleh gambar 1 berikut.



Gambar 1. Alur proses analisis sentimen dengan *lexicon based* [11]

2.3 N-Gram Lexicon

Pemodelan N-gram [8] adalah pendekatan identifikasi dan analisis fitur yang populer digunakan dalam pemodelan bahasa dan bidang pemrosesan bahasa alami. N-gram adalah urutan item yang berdekatan dengan panjang n. Itu bisa berupa urutan kata, *byte*, suku kata, atau karakter. Model N-gram yang paling banyak digunakan dalam kategorisasi teks adalah n-gram berbasis Word dan Characters. Model N-gram dalam analisis sentimen membantu menganalisis sentimen teks atau dokumen. Karakteristik pada N-gram sebagai berikut:

- a. Dapat berfungsi dengan baik walaupun terdapat kesalahan tekstual
- b. Dapat berjalan secara efisien, membutuhkan penyimpanan yang sederhana
- c. Dan waktu proses yang cepat.

Jenis N-Gram berdasarkan dari jumlah potongan *gram substring* yaitu *uni-gram*, *bi-gram*, *tri-gram*, *quad-gram* dan seterusnya sesuai dengan jumlah n dalam n-gram. Pada penelitian ini akan menggunakan sampai dengan jenis *Tri-Gram*. Contoh dari proses pemenggalan kalimat menggunakan

N-Gram Word dengan jumlah pemotongan pada gram, jika hasil *input*-an “bagus dan memudahkan dalam mengecek jumlah pajak dan tunggakan” maka:

- a. *Uni-gram* : {“bagus”, “dan”, “memudahkan”, “dalam”, ..., “tunggakan”}
- b. *Bi-gram* : {“bagus”, “bagus_dan”, ..., “dan_tunggakan”, “tunggakan”}
- c. *Tri-gram* : {“bagus”, “bagus_dan”, “bagus_dan_memudahkan”... “tunggakan”}

2.4 Text Preprocessing

Dalam melakukan *text mining*, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu baru dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau *dataset* mentah disebut juga dengan proses *text preprocessing*. *Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur menjadi data yang terstruktur. Secara umum proses yang dilakukan dalam tahapan *preprocessing* adalah sebagai berikut [12]:

1. Remove Punctuation

Tahap penghapusan tanda baca yang tidak dibutuhkan dalam proses ekstraksi opini. Contoh tanda baca yang dihapus [`”#$%&’()*+,-./:;<=>?@[\\]^_`{|}~`]:

2. Case Folding

Case folding adalah proses penyamaan *case* dalam sebuah dokumen. Hal ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu peran *case folding* dibutuhkan dalam mengonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (dalam hal ini huruf kecil atau *lowercase*).

3. Spelling Normalization

Spelling Normalization merupakan proses perbaikan atau substitusi kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Substitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya mempunyai maksud dan arti yang sama tetapi akan dianggap sebagai entitas yang berbeda pada saat proses penyusunan matriks.

4. Tokenizing

Tokenizing adalah proses memecah kalimat menjadi kata-kata yang dilakukan untuk menjadikan sebuah kalimat menjadi lebih bermakna. Tahap pertama yang dilakukan adalah normalisasi kata dengan mengubah semua karakter huruf menjadi huruf kecil atau *toLowerCase*. Proses tokenisasi diawali dengan menghilangkan delimiter-delimiter yaitu simbol dan tanda baca yang ada pada teks tersebut seperti @, \$, &, koma (,) tanda tanya (?). Tahap tokenisasi selanjutnya yaitu proses penguraian teks yang semula berupa kalimat-kalimat yang berisi kata-kata. Proses pemotongan *string* berdasarkan tiap kata yang menyusunnya, umumnya setiap kata akan terpisahkan dengan karakter spasi, proses tokenisasi mengandalkan karakter spasi pada dokumen teks untuk melakukan pemisahan. Hasil dari proses ini adalah kumpulan kata saja.

5. Stopword Removal

Tahap mengambil kata-kata penting dari hasil token. Algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata yang penting) dapat digunakan pada tahap ini. *Stopword* adalah kata-kata yang tidak deskriptif dan bukan merupakan kata penting dari suatu dokumen sehingga dapat dibuang. Contoh *stopword* adalah “yang”, “dan”, “di”, “dari” dan seterusnya. Dalam filtrasi ini menggunakan *stoplist/stopword* agar kata-kata yang kurang penting dan sering muncul dalam suatu dokumen dibuang sehingga hanya menyisakan kata-kata yang penting dan mempunyai arti yang diproses ke tahap selanjutnya.

6. Feature Selection

Tahap ini merupakan tahap lanjutan dari pengurangan dimensi. Walaupun ditahap sebelumnya sudah melakukan penghapusan kata-kata yang tidak deskriptif (*stopwords*), tidak semua kata-kata di dalam dokumen memiliki arti penting. Sehingga untuk mengurangi dimensi, pemilihan hanya dilakukan pada kata-kata yang relevan dan yang benar-benar mempresentasikan isi dari suatu dokumen. Kata-kata yang dinilai penting dilihat dari intensitas kemunculan dan yang paling informatif dari keseluruhan.

7. Stemming

Stemming merupakan salah satu tahapan dalam proses *text preprocessing*. *Stemming* bertujuan untuk mentransformasikan kata menjadi kata dasarnya (*root*) dengan menghilangkan semua imbuhan kata (*affixes*) meliputi awalan kata (*prefixes*), sisipan kata (*infixes*), akhiran kata (*suffixes*) dan atau menghilangkan awalan dan akhiran kata (*confixes*) pada kata turunan.

Adapun tahapan algoritma *stemming* sebagai berikut:

- a. Langkah pertama adalah memeriksa apakah kata tersebut merupakan akar kata (*root*) terdapat dalam daftar akar kata (*root*). Jika kata tersebut merupakan akar kata, maka proses dihentikan pada tahap pertama ini.
 - b. Menghilangkan *Inflection Suffixes* (“-lah”, “-kah”, “-ku”, “-mu”, atau “-nya”). Jika kata berupa *particles* (“-lah”, “-kah”, “-tah” atau “-pun”) maka langkah ini diulangi lagi untuk menghapus *Possesive Pronouns* (“-ku”, “-mu”, atau “-nya”).
 - c. Menghilangkan *derivational suffix* (imbuhan turunan). Hilangkan imbuhan -i, -kan, -an.
 - d. Menghilangkan *derivational prefix* (awalan turunan). Hilangkan awalan be-, di-, ke-, me-, pe-, se- dan te-.
 - e. Bila dari langkah iv di atas belum ketemu juga. Maka lakukan analisis apakah kata tersebut masuk dalam tabel diambiguitas kolom terakhir atau tidak.
 - f. Bila dari langkah iv di atas belum ketemu juga. Maka lakukan analisis apakah kata tersebut masuk dalam tabel diambiguitas kolom terakhir atau tidak.
 - g. Bila semua proses di atas gagal, maka algoritma mengembalikan kata aslinya.
8. Word2Vec
- Model Word2Vec yang diusulkan Mikolov dalam bentuk vektor mampu menangkap sintak dan makna semantik. Selain itu, model ini memiliki kinerja terbaik dalam pengelompokan kata yang sama. Arsitektur dari model ini menggunakan *neural network* untuk mendapatkan vektor dan memiliki 3 layer yaitu *input*, *projection (hidden layer)* dan *output* [7].
- Model Word2Vec dapat memproses data teks tidak terstruktur dengan mengambil korpus kata sebagai *input* dan menghasilkan vektor kata sebagai *output*. Salah satu keunggulan utama model Word2Vec adalah model ini merepresentasikan fitur sebagai vektor padat daripada representasi *sparse* konvensional yang secara umum mampu mengatasi masalah sinonim dan homonim yang sering ditemukan dalam tugas NLP [7].

2.5 Ukuran Evaluasi Model Klasifikasi

Evaluasi terhadap suatu klasifikasi umumnya dilakukan menggunakan sebuah himpunan data uji, yang tidak digunakan dalam pelatihan klasifikasi tersebut, dengan suatu ukuran tertentu. Terdapat sejumlah ukuran yang dapat digunakan untuk menilai atau mengevaluasi model klasifikasi, diantaranya adalah: *accuracy* atau tingkat pengenalan, *error rate* atau tingkat kesalahan atau kekeliruan klasifikasi, *recall* atau *sensitivity* atau *true positive*, *specificity* atau *true negative* dan *precision*. Model klasifikasi yang dibuat adalah pemetaan dari suatu baris data dengan keluaran sebuah hasil prediksi kelas/target dari data tersebut. Klasifikasi yang memiliki dua kelas sebagai keluarannya disebut dengan klasifikasi biner. Kedua kelas tersebut biasa direpresentasikan dalam $\{0,1\}$, $\{+1,-1\}$ atau $\{positive; negative\}$ [13]. Empat istilah penting untuk memahami semua ukuran evaluasi adalah sebagai berikut:

1. TP atau *True Positive* adalah jumlah *tuple* positif yang dilabeli dengan benar oleh *classifier*. Yang dimaksud *tuple* positif adalah *tuple* aktual yang berlabel positif, seperti *tuple* dengan label Bonus='Ya'.
2. TN atau *True Negatives* adalah jumlah *tuple* negatif yang dilabeli dengan benar oleh *classifier*. Yang dimaksud *tuple* negatif adalah *tuple* aktual yang berlabel negatif, seperti *tuple* dengan label Bonus='Tidak'.
3. FP atau *False Positive* adalah jumlah *tuple* negatif yang salah dilabeli oleh *classifier*. Misalnya, sebuah *tuple* pelanggan yang berlabel Bonus='Tidak' tetapi oleh *classifier* dilabeli Bonus='Ya'.
4. FN atau *False Negatives* adalah jumlah *tuple* positif yang salah dilabeli oleh *classifier*. Misalnya sebuah *tuple* pelanggan yang berlabel Bonus='Ya', tetapi oleh *classifier* dilabeli Bonus='Tidak'.

Keempat istilah di atas dapat direpresentasikan sebagai matriks 2 x 2 yang disebut *confusion matrix* yang ditunjukkan oleh gambar 2.

		Aktual	
		Class	Positive
Prediksi	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Gambar 2. Confusion Matrix

Confusion matrix berguna untuk menganalisis kualitas *classifier* dalam mengenali *tuple-tuple* dari kelas yang ada. TP dan TN menyatakan bahwa *classifier* mengenali *tuple* dengan benar, artinya *tuple* positif dikenali sebagai positif dan *tuple* negatif dikenali sebagai *negative*. Sebaliknya, FP dan FN menyatakan bahwa *classifier* salah dalam mengenali *tuple*, *tuple* positif dikenali sebagai negatif dan *tuple* negatif dikenali sebagai positif. Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai akurasi, presisi, *recall* dan *f1 score* biasa ditampilkan dalam persentase.

1. Accuracy

Akurasi adalah jumlah proporsi prediksi yang benar. Adapun rumus penghitungan akurasi.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. Precision

Precision adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks yang terpilih oleh sistem. Rumus *precision*.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

3. Recall

Recall adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks relevan yang ada pada koleksi. Rumus *recall*.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

4. f1 score

f1 score adalah perbandingan rata-rata presisi dan *recall* yang dibobotkan. Rumus *f1 score*.

$$f1 \text{ score} = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}} \quad (4)$$

Hasil pengukuran *accuracy*, *precision* dan *recall* pada penelitian sebelumnya dengan menggunakan pendekatan lexicon [2] diperoleh *precision* dan *recall* untuk kelas negatif secara terurut sebesar 61.71% dan 33.51%, untuk kelas netral sebesar 59.85% dan 15.67%, dan untuk kelas positif sebesar 81.98% dan 98.63%. Serta untuk akurasinya diperoleh 80.34%.

3. METODE PENELITIAN

Penelitian ini akan dilakukan dengan menerapkan langkah yang sistematis sesuai dengan gambar

3. Langkah-langkah penelitian dapat dikategorikan menjadi beberapa bagian utama antara lain:

1. Pengumpulan data.

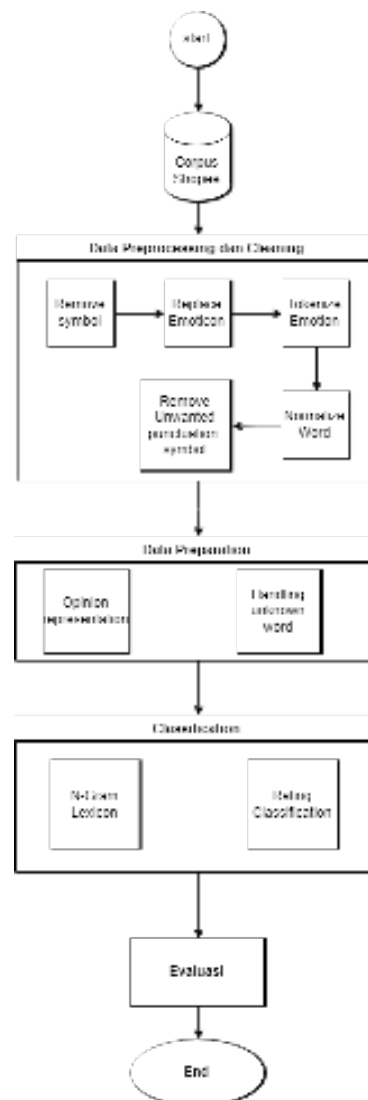
Data yang digunakan pada penelitian ini merupakan data corpus opini pembeli pada aplikasi Shopee (*e-commerce*). Data diperoleh dari penelitian sebelumnya [2] yang terdiri dari 124.606 opini dengan atribut yang tersedia yaitu *itemname*, *date*, *starrate* dan opini. Data corpus juga akan ditambah dengan opini pembeli pada masa COVID-19 dengan menerapkan proses *crawling* sehingga data yang diproses menjadi 220.337 opini. Proses *crawling* dilakukan dengan membuat program dengan memanfaatkan API yang disediakan Shopee.

2. Data Preprocessing dan Cleaning

Tahapan ini terdiri dari beberapa tahapan antara lain:

- a. *remove symbol*. Tahapan ini akan membersihkan opini dari simbol dan angka yang tidak diperlukan atau berlebihan.

- b. *replace emoticon*. Tahapan ini akan mengganti *emoticon* yang disisipkan pengguna pada opini dengan sentimen yang merepresentasikan *emoticon* tersebut. Tahap ini akan menghasilkan sebuah kamus data *emoticon* yang sesuai.
- c. *tokenize emotion*. Tahapan ini akan memisahkan emosi dengan kata-kata di sekitarnya supaya bisa diproses lebih lanjut.
- d. *normalize word*. Tahapan ini memperbaiki kata-kata pada opini seperti kata dengan karakter berulang lebih dari 2 karakter seperti kata sukaaaaaaaaaaaaaaaaa menjadi suka, kata bagggguuuuuussssssss menjadi bagus, baaaaaanngggggggeeeeeeetttttt menjadi banget. Memperbaiki kata singkatan maupun kata *slang* berdasarkan kamus data seperti kata bgt menjadi banget, serta menerapkan *stemming* untuk memperoleh kata dasar seperti menyerap menjadi serap, bajunya menjadi baju.
- e. *remove unwanted punctuation symbol*. Tahapan ini menghapus simbol-simbol lain yang masih ditemukan setelah menerapkan *remove symbol*.



Gambar 3. Usulan Sistematis Penelitian

3. *Data Preparation*

Tahapan ini terdiri dari beberapa tahapan antara lain:

- a. *Opinion representation*. Tahapan ini akan memproses kata-kata yang memiliki kedekatan/kemiripan dari segi semantik. Teknik yang digunakan akan mengadopsi pendekatan Word2Vec berbahasa Indonesia.
 - b. *Handling unknown word*. Tahapan ini akan menghapus kata yang tidak memiliki makna dalam mempengaruhi analisis sentimen.
4. *Classification*
Tahapan ini akan mengklasifikasikan opini dan *rating* ke dalam 3 kategori antara lain positif, negatif dan netral.
 5. *Evaluation*
Hasil klasifikasi opini dan *rating* akan dihitung *confusion matrix* untuk mendapatkan *accuracy*, presisi, dan *recall*.

4. HASIL DAN PEMBAHASAN

Data yang digunakan pada penelitian ini merupakan data corpus opini pembeli pada aplikasi Shopee (*e-commerce*). Data diperoleh dari penelitian sebelumnya yang terdiri dari 124.606 opini dengan atribut yang tersedia yaitu *itemname*, *date*, *rating* dan opini. Data corpus sudah ditambah dengan opini pembeli pada masa COVID-19 dengan menerapkan proses *crawling* sehingga total opini yang diolah menjadi 220.337. Opini hasil *crawling* akan dilakukan *preprocessing* untuk membersihkan opini dari data-data yang tidak diperlukan untuk proses sentimen dan untuk mempercepat proses. Tabel 1 menunjukkan sampel data opini dan *rating* hasil *crawling*.

Tabel 1. Sampel Corpus

No	Komen	Rating
1	whoaaa~~ cepat banget sampenyaa~~ meskipun respon admin rada lama, tapi overall PUAS~!!^^ terima kasih~~^^	5
2	Terimakasih kak,, sukaaaa bangeett..	5
3	Agak panas bahan nya. Gak begitu menyerap keringat :(sadd.... .	5
4	Sukaaaaaaaaaaaaaaaaaaaaa bbbbbbbaaaaaannnnnnnggggggeeeeeeeyttt pengiriman super diper ceoet, respon admin ceoet, recomended seller ♡♡♡♡.	5
5	Sukaaa banget bajunya ♡♡	5
6	Memuaskan sesuai yang diinginkan real juga pengirimannya cepat apalagi bisa cod juga.	5
7	Bagguuuuuusssssss baaaaannggggggeeeeeeetttttt.	5
8	Sukaaaaaaa, barang oke, pengiriman cepat, admin ramaaaaaha 📦	5
9	Kok Sobekkkkk Sichhh kak Huhuhuuuuuuuu.....Pengemasannn Lumayannn Lamaaaa Pengirimannnnn nyaaaa Jugaaaa.....biasa nyaaaa Gerecepppppppp .	4
10	Dipake sekali langsung melar.	2
11	kecewa bgt ..ini si keliatan bgt bekasnyaa kucel bgt ☹ ga bsa dpke .	1
12	Bahan terlalu tipis.	3

Berdasarkan data sampel di atas, opini yang ditulis konsumen tidak terstruktur dan tidak baku sehingga perlu proses *preprocessing* dan *cleansing*. Proses *preprocessing* dan *cleansing* telah dilakukan terhadap corpus. Tahapan yang dilakukan adalah *case folding*, *replace emoticon*, *remove punctuation*, *tokenize emotion*, dan *normalize word*. *Case folding* merupakan proses mengubah opini menjadi huruf kecil semua. Proses ini membantu dalam proses pengolahan data tanpa perlu membedakan huruf besar


```

stad_lare = (s_kata,ku_mest_adm_lar('kecewa'))
len_stad_lare = len(stad_lare)
print(' | ' + '-' * (len(stad_lare[a],stad_lare[a]))

ipocvawin 0.811241180932779
bawak 0.777244244471934
kepat 0.77712504172163
ipocvawin 0.775001220753131
kecewa 0.773179793761727
pasa 0.771687124471932
ipocvawin 0.885269410075034
kepat 0.870085911386111
ipocvawin 0.871175217527777
ipocvawin 0.870085911386111

```

Gambar 4. Hasil *Similarity* Kata Kecewa

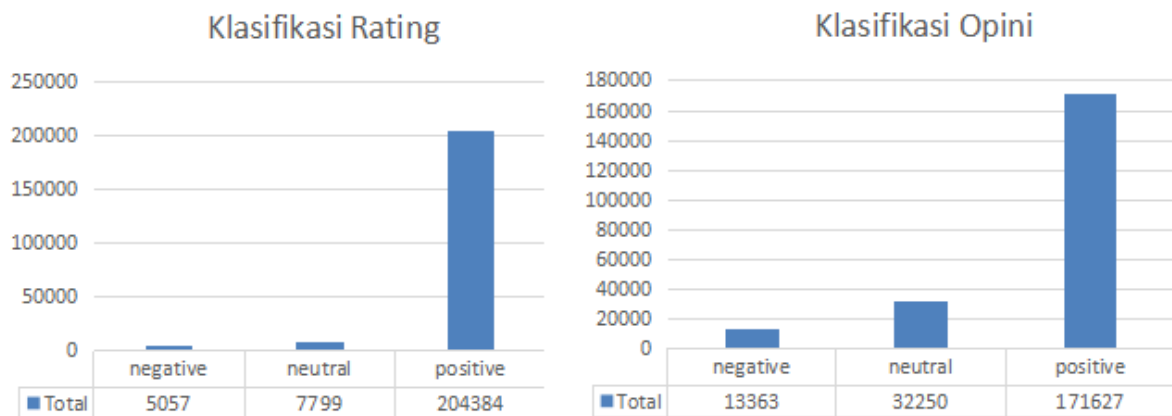
Pada contoh di atas, maka semua kata pada kolom *similar* akan memiliki bobot yang sama dengan kata “kecewa” yang dilanjutkan dengan proses klasifikasi dengan menggunakan pendekatan lexicon. Penentuan klasifikasi menggunakan perbandingan maksimum bobot opini positif dengan maksimum bobot opini negatif. Klasifikasi opini yang dihasilkan adalah positif, negatif atau netral. Klasifikasi juga dilakukan terhadap *rating* pada opini. Hasil klasifikasi untuk opini dapat dilihat pada tabel 3.

Tabel 3. Hasil Klasifikasi Opini

No	<i>Preprocessing</i>	<i>k_senti</i>	<i>kategori_senti</i>
1	whoa cepet banget sampenyaa meski respon admin rada lama overall puas terima kasih	whoa cepet banget sampenyaa meski [2] respon [2] admin rada [-2] lama [-2] overall puas [3] terima kasih [4]	positive
2	terimakassihh kak suka bangeet	terimakassihh kak suka [3] bangeet	positive
3	panas bahan nya begitu serap keringat sadd	panas [-2] bahan nya begitu [-3] serap keringat sadd	negative
4	suka bangeyt kirim super per ceoet respon admin ceoet recomended seller	suka [3] bangeyt kirim [2] super [4] per [-2] ceoet respon [2] admin ceoet recomended seller	positive
5	suka banget baju	suka [3] banget baju	positive
6	muas sesuai dingin real kirim cepat bisa cod juga	muas [3] sesuai [2] dingin [-2] real kirim [2] cepat [1] bisa [-1] cod juga	positive
7	bagus bannget	bagus [2] bannget	positive
8	suka barang oke kirim cepat admin ramaha	suka [3] barang oke [1] kirim [2] cepat [2] admin ramaha	positive
9	kok sobek sich kak huhuhuhu emas lumayan lama kirim nya juga biasa nya gerecep	kok sobek [-2] sich [-2] kak huhuhuhu emas lumayan [1] lama [-2] kirim [2] nya juga biasa [1] nya gerecep	neutral
10	dipake sekali langsung melar	dipake sekali langsung melar [-2]	negative
11	kecewa banget ini si liat banget bekasnyaa kucel banget bsa dpke	kecewa [-2] banget ini si liat [-2] banget bekasnyaa kucel banget bsa [-2] dpke	negative
12	bahan terlalu tipis	bahan terlalu [-3] tipis [-2]	negative

Gambar 5 menunjukkan grafik hasil klasifikasi opini dan klasifikasi *rating*. Pada grafik diketahui bahwa data opini konsumen yang diolah memiliki kelas positif lebih banyak dibanding kelas lainnya. Sedangkan klasifikasi opini juga mendapatkan hasil yang sama yaitu kelas positif lebih banyak

dibanding kelas lainnya. Namun terdapat perbedaan jumlah opini pada masing-masing kelas. Hal tersebut menunjukkan bahwa terdapat *rating* yang tidak sesuai dengan isi *review* atau sebaliknya.



Gambar 5. Grafik Klasifikasi *Rating* dan Klasifikasi Opini

Klasifikasi *rating* dan opini dilanjutkan ke tahap evaluasi dengan menggunakan *confusion matrix*. Tabel 4 menampilkan informasi hasil klasifikasi *rating* dengan klasifikasi opini.

Tabel 4. *Confusion Matrix rating vs opini*

<i>Confusion</i>	<i>Negative</i>	<i>Neutral</i>	<i>Positive</i>
Negative	1,170	1,801	2,086
Neutral	1,446	2,173	4,180
Positive	10,747	28,276	165,361

Hasil evaluasi diperoleh akurasi sebesar 78%, serta perhitungan presisi 44%, *recall* 37% dan *f1 score* 40%. Dengan menggunakan corpus pada penelitian ini, proses klasifikasi dijalankan dengan menggunakan model pada penelitian sebelumnya dan diperoleh evaluasi akurasi sebesar 76%, presisi 57%, *recall* 41% dan *f1 score* 48%.

5. KESIMPULAN

Penelitian ini telah mencoba menambahkan fitur *opinion representation* dan pengecekan semantik kata pada *dataset* opini di Shopee. Berdasarkan pengujian, penambahan 2 fitur tersebut dapat meningkatkan akurasi pada sentimen analisis.

6. SARAN

Untuk pengembangan penelitian ini lebih lanjut, maka saran yang dapat disampaikan adalah memperhatikan *emoticon* dalam bentuk grafik maupun gabungan tanda baca dengan menggantikan *emoticon* yang disisipkan pengguna pada opini dengan kata emosi yang merepresentasikan *emoticon* tersebut.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Mikroskil yang telah memberi dukungan finansial terhadap penelitian ini.

DAFTAR PUSTAKA

- [1] E. Mardiani, N. Rahmansyah and F. A. Rizky, "Perilaku Konsumen Terhadap E-Commerce Disaat Pandemi Covid 19 Di Shop And Travel," *Jurnal Informatik*, vol. 3, no. 16, Desember 2020.
- [2] E. Halim, R. Purba and Andri. , "Consumer Opinion Extraction Using Text Mining For Product Recommendations On E-Commerce," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 4, no. 1, pp. 19-28, Maret 2021.
- [3] Mars and M. S. Goudier, "Big Data Analysis to Features Opinions Extraction od Customer," *International Conference on Knowledge Based and Intelegent Information and Engineering Systems, KES2017*, 6-8 September 2017, Marseille, France, p. 906–917, 2017.
- [4] T. V. O, S. Y and A. , "Labeling Analysis in the Classification of Product Review Sentiments by using Multinomial Naive Bayes Algorithm," *Journal of Physics: Conference Series*, 1192, p. 012036, 2019.
- [5] M. Yassine and H. Hajj, "A Framework for Emotion Mining from Text in Online Social Networks," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2010.
- [6] H. e. a. Al-Rubaiee, "Techniques for Improving the Labelling Process of Sentiment Analysis in the Saudi Stock Market," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 3, 2018.
- [7] R. P. Nawangsari, R. Kusumaningrum and A. Wibowo, "Word 2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study," *International Conference on Computer Science and Computational Intelligence*, vol. 4, pp. 360-366, 2019.
- [8] A. Dey, M. Jenamani and J. J. Thakkar, "Senti-N-Gram: An n-gram lexicon for sentiment analysis," *Expert Systems With Applications*, 2018.
- [9] S. R. De, "Sentiment analysis on product purchase through e commerce," *International Journal of Scientific Research and Management*, vol. 5, no. 6, pp. 5441-5444, 2017.
- [10] S. L. e. a. Addepalli, "A Proposed Framework for Measuring Customer Satisfaction and Product Recommendation for Ecommerce," *International Journal of Computer Applications*, vol. 138, no. 3, pp. 30-35, 2016.
- [11] A. Nurfalah and A. A. Suryani, "Analisis Sentimen Berbahasa Indonesia dengan Pendekatan Lexicon-Based Pada Media Sosial," *Jurnal Masyarakat Informatika Indonesia*, vol. 2, no. 1, pp. 1-8, 2017.
- [12] Informatikalogi, "Data Science/Text Mining," 2017. [Online]. Available: <https://informatikalogi.com/text-preprocessing/>. [Accessed Mei 2019].
- [13] S. Data Mining Untuk Klasifikasi dan Klasterisasi Data, Bandung: Informatika, 2019.