Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

Analisis Sentimen Ulasan Aplikasi Instagram di Google Play Store: Pendekatan Multinomial Naive Bayes dan Berbasis Leksikon

Novresia Wijaya*, Erwin Setiawan Panjaitan

Fakultas Teknologi Informasi, Program Studi Informatika, Universitas Mikroskil, Medan, Indonesia
Email: novresiawijaya@gmail.com, erwin@mikroskil.ac.id
Email Penulis Korespondensi: novresiawijaya@gmail.com
Submitted: **17/07/2024**; Accepted: **09/09/2024**; Published**: 09/09/2024**

Abstrak−Media sosial seperti Instagram memiliki peran yang signifikan dalam kehidupan sehari-hari banyak individu. Untuk memahami pengalaman pengguna dengan media sosial, kita dapat membaca ulasan dan penilaian oleh pengguna. Namun, seringkali penilaian ini mungkin tidak secara akurat mencerminkan isi dari ulasan mereka. Oleh karena itu, penting untuk melakukan analisis terhadap tanggapan ini untuk memahami secara umum apa yang sering dikeluhkan oleh pengguna. Penelitian ini bertujuan untuk mengembangkan metode analisis sentimen yang akurat untuk ulasan pengguna instagram dengan Naïve bayes yang dipadukan dengan lexicon-based, guna untuk mengatasi ketidaksesuaian antara penilaian bintang dan isi ulasan.masalah utama yang dihadapi adalah bagaimana menganalisis sentimen pengguna instagram secara akurat, mengingat adanya potensi ketidaksesuaian tersebut. Untuk mengatasi masalah ini, penerlitian ini menggunakan metode Naïve bayes yang dipadukan dengan lexicon-based untuk menentukan sentimen positif dan negatif terhadap ulasan pengguna. Hasil pengujian menunjukkan nilai akurasi sebesar 92%, dengan presisi 84%, recall sebesar 91% dan F1-Score 87%.

Kata Kunci: Analisis sentimen; Lexicon-Based; N-Gram; Naïve Bayes; Social media

Abstract−Social media platforms like Instagram play a significant role in the daily lives of many individuals. To understand user experiences with social media, we can read reviews and ratings provided by users. However, these ratings often may not accurately reflect the content of their reviews. Therefore, it is important to analyze these responses to understand the common complaints users have. This study aims to develop an accurate sentiment analysis method for Instagram user reviews by combining Naïve Bayes with a lexicon-based approach to address the discrepancy between star ratings and the content of reviews. The main issue addressed is how to accurately analyze Instagram user sentiment, given the potential discrepancies. To tackle this problem, the study employs a Naïve Bayes method combined with a lexicon-based approach to determine positive and negative sentiments towards user reviews. The testing results show an accuracy of 92%, with a precision of 84%, recall of 91%, and an F1-Score of 87%.

Keywords: Lexicon-Based; Naives Bayes ; N-gram; Social media; Sentiment Analysis

1. PENDAHULUAN

Dalam era digital ini, aplikasi media sosial seluler seperti Instagram memiliki peran yang signifikan dalam kehidupan sehari-hari banyak individu. Instagram adalah platform berbagi foto yang memungkinkan pengguna mengambil foto, mengeditnya dengan filter, dan membagikannya dengan orang lain. Instagram didirikan pada tahun 2010 oleh Mike Krieger dan Kevin Systrom [1]. Berdasarkan data yang dicatat oleh We Are Social, jumlah pengguna Instagram di seluruh dunia mencapai 1,64 miliar pada Oktober 2023, dengan angka ini meningkat sebesar 2,5% dari kuartal ke kuartal dan melonjak sebesar 18,1% dari tahun ke tahun [2].Platform ini telah berkembang menjadi fenomena global dengan miliaran pengguna bulanan, mengubah cara berbagi foto dan video. Fitur-fitur inovatifnya memungkinkan pengguna mendokumentasikan momen penting, berinteraksi dengan orang lain, dan berhubungan dengan merek. Keberhasilan platform ini terletak pada kemampuannya untuk memfasilitasi interaksi digital yang bermakna dan berbagi konten yang beragam, menjadikannya aplikasi media sosial terkemuka di seluruh dunia [3].

Meskipun faktor-faktor ini berkontribusi pada popularitas luas Instagram, penting untuk memeriksa bagaimana pengguna sebenarnya mengalami aplikasi ini dalam praktik. Salah satu cara untuk menilai hal ini adalah melalui ulasan dan penilaian yang diberikan oleh pengguna di platform utama seperti Google Play Store. Evaluasi yang dihasilkan pengguna ini menawarkan wawasan berharga tentang kinerja dan tingkat kepuasan aplikasi di dunia nyata.Google Play Store adalah platform yang disediakan oleh Google yang menawarkan berbagai konten digital. Diluncurkan pada tahun 2008 (awalnya sebagai Android Market) [4], ini menyediakan toko online untuk berbagai produk, termasuk aplikasi, game, film, musik, dan buku. Ini memungkinkan pengguna mengakses dan mengunduh konten digital dengan mudah melalui perangkat Android mereka. Dengan jutaan aplikasi dan miliaran pengguna di seluruh dunia, Google Play Store telah menjadi pusat utama distribusi konten seluler.

Sebagai salah satu platform distribusi konten digital terkemuka di dunia, Google Play Store tidak hanya menyediakan berbagai produk yang luas, tetapi juga memfasilitasi umpan balik pengguna melalui sistem penilaiannya. Pengguna dapat memberikan penilaian dari 1 hingga 5 bintang untuk mengevaluasi aplikasi, dengan penilaian bintang yang lebih tinggi umumnya menunjukkan tingkat kepuasan pengguna yang lebih tinggi. Dalam membuat keputusan tentang layanan atau produk online, pengguna sering kali merujuk pada ulasan dan penilaian dari orang lain, dengan banyak yang memilih pilihan yang memiliki penilaian bintang yang lebih tinggi. Namun, penilaian ini tidak selalu mencerminkan isi ulasan mereka [5]. Ketidaksesuaian ini dapat terjadi karena berbagai faktor, seperti penilaian impulsif, kesalahpahaman tentang sistem penilaian, atau penilaian yang sudah usang yang tidak mencerminkan pembaruan terbaru. Akibatnya, penilaian yang diberikan mungkin tidak cukup untuk menggambarkan kualitas

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

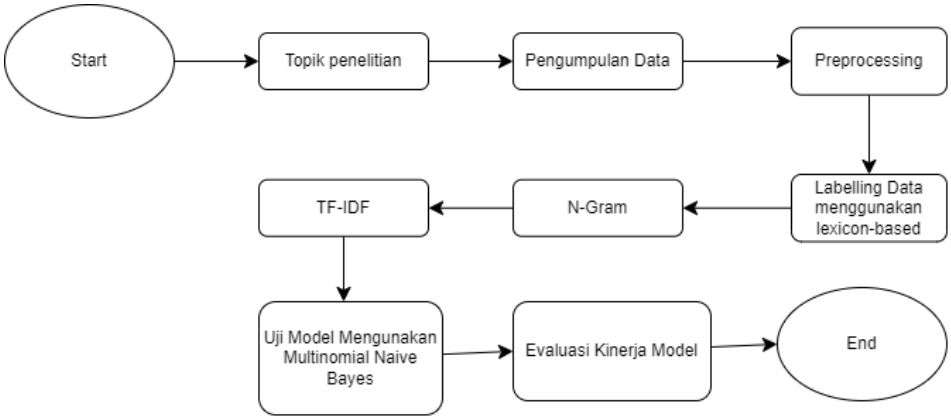
sebenarnya dari aplikasi [6]. Untuk mengatasi keterbatasan ini, metode analisis yang lebih mendalam, seperti analisis sentimen, dapat digunakan untuk memberikan penilaian yang lebih dalam dan akurat tentang kualitas aplikasi di luar apa yang ditangkap oleh penilaian bintang saja.

Analisis sentimen, yang juga dikenal sebagai penambangan opini, adalah bidang penelitian yang menyelidiki sudut pandang, penilaian, evaluasi, perasaan, perspektif, dan emosi mengenai berbagai entitas seperti produk, layanan, organisasi, individu, peristiwa, subjek, dan atribut lainnya [7]. Sebagai teknik yang banyak diterima, *Naive Bayes* menikmati popularitas yang besar dalam klasifikasi teks karena efisiensi komputasionalnya dan prediksi yang efektif [8]. Berdasarkan penelitian sebelumnya yang dilakukan oleh Dhodi Surya Sayogo dan teman-temannya menggunakan algoritma Naive Bayes classifier diperoleh akurasi sebesar 89% [7]. Penelitian yang dilakukan oleh Rahmawan Bagus Trianto dan rekan-rekannya Menggunakan Metode N-gram dan Naïve Bayes" mencapai hasil yang mengesankan. Studi ini melaporkan akurasi tertinggi sebesar 94,06% menggunakan ekstraksi fitur Bi-gram [9].Dalam studi lain yang dilakukan oleh Laila Atikah Sari dan rekan-rekannya, penelitian menggunakan Naive Bayes classifier menunjukkan hasil akurasi yang cukup baik sebesar 85,88% [10]. Selain itu, terdapat juga penelitian yang melakukan analisis sentimen berdasarkan ulasan di Shopee E-commerce menggunakan metode Naive Bayes dan menunjukkan nilai akurasi tinggi sebesar 92% [11].

Berdasarkan penelitian sebelumnya, kontribusi dari studi ini dilakukan pada kombinasi metode *Naive Bayes* dengan berbasis Lexicon termasuk memberikan kontribusi pada pengetahuan akademis di bidang analisis sentimen, melibatkan eksplorasi teori, metodologi, dan temuan baru yang dapat memajukan pemahaman dan penerapan analisis sentimen di berbagai domain. Selain itu, jurnal ini bertujuan untuk mendorong penelitian, kolaborasi dan inovasi terbaru di bidang ini dengan menyebarkan temuan ini melalui publikasi. Tujuan utama dari penelitian ini adalah untuk mengembangkan metode kombinasi yang mengabungan Naïve Bayes dan pendekatan berbasis leksikon untuk meningkatkan akurasi analisis sentimen pada uladan pengguna instagram , mengindentifikasi untuk memberikan wawasan yang lebih mendalam tentang pengalaman pengguna , mengevaluasi efektivitas metode yang disusulakn dalam berbagai konteks ulasan dan membandingkannya dengan metode analisis sentimen yang ada, serta menyediakan kerangka kerja yang dapat digunakan untuk mengubah hasil analisis sentimen menjadi rekomendasi konkret untuk perbaikan platform instagram.

2. METODOLOGI PENELITIAN

Tahapan yang dilakukan dalam penelitian ini akan dimulai dengan pengambilan data menggunakan *Python*. Langkah selanjutnya adalah pemberian label pada ulasan menggunakan metode berbasis lexicon, kemudian akan dilakukan proses *preprocessing*. Setelah proses pra pengolahan, akan dilakukan ekstraksi fitur menggunakan *N-gram* kemudian menggunakan *TF-IDF* untuk mengukur kata dan pengujian pada model Multinomial *Naive Bayes* yang telah dilatih.Gambar 1 berikut menampilkan gambaran mengenai alur metodologi penelitian ini.



Gambar 1. Flowchart Tahapan Penelitian

Berdasarkan gambar 1,penelitian ini akan dilaksanan melalui serangkaian tahapan sistematis sebagai berikut :

a. Penentuan topik penelitian : pada tahap ini , ruang lingkup dan fokus penelitian ditetapkan yaitu analasis sentimen terhadap ulasan pengguna Instagram
b. Pengunpulan data : data berupa ulasan pengguna instagram dikumpulkan menggunakan metode crawling data
c. Preprocessing : data yang terkumpul kemudian melalui tahap preprocessing untuk dibersihkan dan disispakan untuk analisis yang lebih lanjut
d. Labelling data : menggunakan lexicon-based, data diberikan label sentimen positif atau netral berdasarkan Vader library
e. Penerapan N-gram : teknik N-gram diterapkan untuk menganalisis sekuens kata dalam ulasan
f. Pembobotan TF-IDF : digunakan untuk menghitung bobot kata-kata dalam dataset

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

g. Uji Model : model Mulltinomial Naives bayes diimplementasikan dan diuji menggunakan dataset yang telah disiapkan
h. Evaluasi kinerja model : peerforma model diwvaluasi menggunakan matrix.

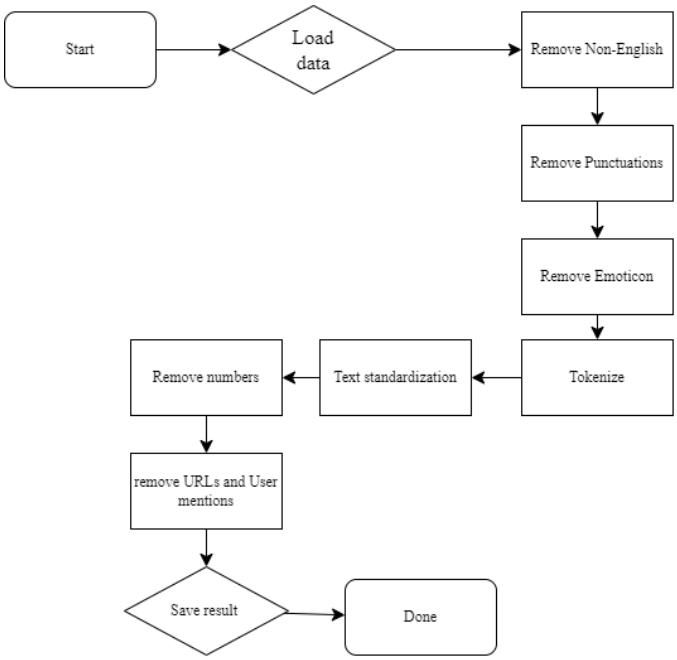
2.1 Tahap Pengumpulan Data

Pada proses ini dilakukan pengambilan data ulasan Instagram yang tersedia di Google Play Store dengan cara *Web Scraping* atau *spidering*. *Web Scraping* atau *spidering* adalah metode yang memungkinkan ekstraksi data secara otomatis dari Internet. Google Play Store merupakan salah satu platform distribusi yang menyediakan aplikasi dan konten dgital pada perangkat berbasis android. Data yang diperoleh dari pengambilan data disimpan dalam format .csv.

2.2 Tahap *Preprocessing*

Tahap *Preprocessing* adalah komponen penting dalam persiapan data. Hal ini karena struktur data yang dikumpulkan pada tahap pengumpulan sering kali tidak teratur, yang dapat menghambat kelancaran proses-proses berikutnya. Dalam studi ini, beberapa tahap Preprocessing yang akan dibahas meliputi penghapusan kata non-Inggris, penghapusan tanda baca, penghapusan emotikon, tokenisasi, dan standarisasi teks.

Tahap-tahap *Preprocessing* ini bekerja secara bersama-sama untuk menyempurnakan dan menstandarisasi data tekstual, sehingga menjadi lebih sesuai untuk analisis. Melalui penerapan prosedur-prosedur ini, kami memastikan bahwa data tetap konsisten, fokus pada informasi tekstual yang relevan, dan dipersiapkan dengan baik untuk pemrosesan yang akurat dan efisien dalam fase analitis berikutnya.



Gambar 2. Flowchart Pre-Processing

Yang dimana berdasarkan Gambar 2. Maka langkah-langkah preprocessing dapat dideskripsikan sebagai berikut:
a. *Remove Non-English* : menghapus elemen teks non-Inggris, memastikan dataset yang homogen dan hanya berisikan data berbahasa Inggris.
b. *Remove Punctuation* : menghapus karakter non-alfanumerik seperti titik, koma , tanda seru dan tanda tanya.
c. *Remove emoticon* : menghapus emotikon dengan mengganti subtrings yang terdeteksi.
d. *Tokenize* : memecah teks menjadi unit-unit kecil terutama kata-kata individu.
e. *Text standardization* : mengubah kata-non standar menjadi bentuk standar melalui penghapusan stopwords, konversi huruf kecil, penghapusan karakter duplikat dan penerapan stemming
f. *Remove numbers* : menghaspus elemen numerik yang tidak relevab
g. *Remove URLs and User Mentions* : menhapus tautan web (IRL) dan referensi pengguna

2.3 Pelabelan Data

Dalam penelitian ini, pelabelan data menggunakan metode berbasis lexicon dengan integrasi pustaka sentimen *VADER*. *Valence Aware Dictionary and Sentiment Reasoner (VADER)* adalah metode analisis berbasis lexicon. *VADER* menganalisis teks berdasarkan lexicon untuk menghasilkan kelas sentimen bersama dengan skor gabungan. *VADER* Sentiment Lexicon adalah salah satu dari leksikon ini, mencakup token yang terdiri dari istilah bahasa Inggris, emotikon, dan sentimen yang terkait dengan akronim dan inisial. Data yang diambil akan diberi label "Positif" atau "Negatif" [12].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

Keuntungan menggunakan *VADER* terletak pada leksikonnya yang komprehensif, yang berisi nilai-nilai sentimen untuk setiap kata. Berbagai perintah pemrograman Python yang memanfaatkan *VADER* akan dijalankan untuk menganalisis teks. *VADER* akan memanggil data leksikon dari server *NLTK* (*Natural Language Toolkit*) untuk menghitung kelas polaritas sentimen [13]. Skor gabungan, metrik yang merangkum sentimen keseluruhan dari teks, akan digunakan untuk menentukan apakah ulasan tersebut positif atau negatif.

2.4 N-gram

N-gram adalah model yang digunakan untuk memprediksi kata berikutnya yang mungkin mengikuti urutan dari N-1 kata sebelumnya [14]. Setelah tahap preprocessing, data akan melalui proses tokenisasi menggunakan jenis token N-gram. Dalam studi ini, tiga jenis tokenisasi yang digunakan adalah Unigram, Bigram, dan Trigram.

2.5 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) dirancang untuk memberikan signifikansi pada kata-kata (istilah) yang ditemukan dalam dokumen yang telah melalui tahap N-gram sebelumnya. Metode *TF-ID*F menawarkan beberapa keuntungan, seperti representasi istilah yang sederhana, penilaian signifikansi istilah, pengurangan bobot untuk istilah yang sering muncul, dan skalabilitas [15]. Pendekatan ini menggabungkan dua konsep untuk menghitung kepentingan kata: frekuensi kemunculan kata dalam dokumen tertentu dan frekuensi invers dari dokumen yang mengandung kata tersebut. *TF-IDF menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF)* untuk setiap dokumen.
Rumus TF-IDF adalah sebagai berikut:

$$TF - IDF = TD(d,t) \times \left[ln\ ln\ \left(\frac{Nd+1}{df(t)+1}\right) + 1\right] \quad (1)$$

Dalam library *TF-IDF* di *scikit-learn*, perhitungan logaritma dilakukan menggunakan logaritma alami, yang juga dikenal sebagai logaritma dengan basis e, di mana nilai e kira-kira adalah 2.718281828459.

2.6 Multinomial Naive Bayes

Klasifikator *Naive Bayes* adalah salah satu algoritma yang menggunakan konsep probabilitas dan sering diterapkan dalam klasifikasi analisis sentimen [16]. Klasifikator *Naive Bayes*, yang didasarkan pada Teorema Bayes, memiliki kemampuan klasifikasi yang mirip dengan pohon keputusan dan jaringan saraf. Klasifikator *Naive Bayes* terbukti memiliki akurasi dan efisiensi yang tinggi, terutama pada basis data berskala besar, serta dikenal karena proses *preprocessing* yang cepat [17].

Model *Naive Bayes* multinomial adalah model yang umum digunakan dalam klasifikasi teks. Model ini termasuk dalam kategori metode pembelajaran terawasi (*supervised learning*), yang memerlukan pelabelan setiap titik data sebelum proses pelatihan [18]. Saat ini, model multinomial dianggap sebagai pendekatan pemodelan utama, mengungguli model Bernoulli multivariat dalam hal efisiensi ketika mengintegrasikan pemodelan bahasa ke dalam pencarian informasi [19].

$$C_{map} = arg_{c\epsilon\{cl,cs\}}\ max\ P(c|d) \quad (2)$$

$$= arg_{\ c\epsilon\{cl,cs\}} max\ P(c)\prod_{k+1}^{m} \quad P(tk|c) \quad (3)$$

Parameter P(tk|c) (probabilitas likelihood) diestimasi dengan menghitung frekuensi tk di semua dokumen pelatihan dalam c, menggunakan prior Laplace:

$$p(c) = \frac{1+N_k}{|v|+N'} \quad (4)$$

Di mana Nk mewakili jumlah kejadian tk dalam dokumen pelatihan di c, dan N mewakili total kejadian kata dalam c.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, pendekatan digunakan untuk menguji suatu asumsi. Dalam strategi eksperimen pada sentimen, teks diklasifikasikan ke dalam kategori "Positif" dan "Negatif" sebelum dianalisis untuk menentukan tingkat sentimen yang terkandung di dalamnya. Proses ini menggunakan metode Naive Bayes untuk mengukur akurasi, presisi, dan recall.

3.1 Tahap pengumpulan data

3.1.1 Python Library

Sebelum menilai dan mengambil data sentimen ulasan Instagram, penting untuk menginstal beberapa pustaka untuk memastikan proses analisis berjalan lancar. Beberapa pustaka penting yang perlu diinstal antara lain *google-play-*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

scraper yang digunakan untuk mengambil data ulasan Instagram dari Google Play Store, *panda* untuk pemrosesan data, dan *numpy* untuk membuat array dari struktur data. Pustaka *NLTK* bertanggung jawab untuk memproses data bahasa alami, sedangkan *vaderSentiment* digunakan untuk menganalisis sentimen teks, khususnya dalam bahasa Inggris, dengan fokus pada analisis sentimen media sosial. Selain itu, pustaka *Langdetect* digunakan untuk mendeteksi bahasa dalam teks, dan pustaka *Sklearn* membantu dalam pemrosesan dan pelatihan data untuk tujuan pembelajaran mesin atau ilmu data.

3.1.2 Crawling Data

Langkah awal dalam proses ini melibatkan pengimporan pustaka Python yang diperlukan. Pustaka *google_play_scraper* menyediakan fungsi-fungsi yang diperlukan untuk mengambil ulasan aplikasi dari Google Play Store, sedangkan pustaka pandas digunakan untuk mengorganisasi dan menganalisis data yang dikumpulkan.

Proses dimulai dengan mengimpor *fungsi reviews_all* dan enum sort dari pustaka *google_play_scraper*, serta pustaka pandas untuk manipulasi data. Fungsi *reviews_all* memungkinkan ekstraksi semua ulasan yang tersedia untuk aplikasi tertentu, dan enum sort menawarkan opsi untuk menentukan urutan ulasan tersebut.

Dalam konteks penelitian ini, pengumpulan data untuk ulasan aplikasi Instagram dilakukan dengan menggunakan fungsi reviews_all untuk mengekstraksi umpan balik pengguna dari *Google Play Store*, dengan menargetkan ID paket *com.instagram.android*. Proses pengambilan data dioptimalkan melalui konfigurasi parameter spesifik, termasuk penundaan 0 milidetik antara permintaan API, penyaringan bahasa Inggris, pengurutan berdasarkan relevansi, dan tanpa penyaringan berdasarkan skor.

Penelitian ini menggunakan tiga dataset berbeda: Dataset A, yang terdiri dari hingga 20.000 ulasan; Dataset B, yang berisi 199 ulasan; dan Dataset C, yang terdiri dari 398 ulasan. Dataset ini, yang bervariasi dalam skala, memberikan wawasan multifaset tentang pengalaman pengguna Instagram. Dataset A menawarkan perspektif berskala besar yang komprehensif, sedangkan Dataset B dan C menyajikan sampel yang lebih terkonsentrasi. Pendekatan bertingkat ini memfasilitasi analisis yang mendalam tentang berbagai aspek aplikasi, dengan memanfaatkan umpan balik pengguna dari berbagai ukuran sampel. Penggunaan beberapa dataset memungkinkan penelitian untuk mengeksplorasi potensi korelasi antara ukuran sampel dan hasil analisis dalam konteks ulasan berbahasa Inggris.

Setelah pengumpulan ulasan, hasilnya dikonversi menjadi DataFrame menggunakan pustaka pandas. Format tabular ini memudahkan analisis dan manipulasi data. Untuk memverifikasi keberhasilan proses pengumpulan data, lima baris pertama DataFrame ditampilkan untuk inspeksi.Terakhir, jumlah total ulasan yang dikumpulkan dicetak untuk mengonfirmasi bahwa tujuan pengumpulan data telah tercapai. DataFrame kemudian disimpan ke file *CSV* untuk mendukung upaya analisis dan dokumentasi. Tabel 1 berikut menunjukkan ulasan-ulsan yang belum diproses atau dikelola.

Tabel 1. Contoh beberapa data ulasan

ID	Content
117	Very best app to share your daily life fellings and also share news to everyone around this world ðŸ˜Š
207	I love the app, however. Out if curiosity, I changed my year of birth from 2006 to 1999 cuz to see what would be different on a new account. And both my main instagram and Facebook account are stuck at the year 1999 for my year of birth and it says I have to wait a few days. Can I contact instagram and or Facebook to get this mistake changed quickly please?

3.2 Pre-Processing

Pada tahap ini, penghapusan kata-kata non-Inggris, penghapusan tanda baca, penghapusan emotikon, tokenisasi, standarisasi teks, penghapusan angka, dan penghapusan URL serta penyebutan pengguna dilakukan. Penghapusan kata-kata non-Inggris adalah langkah pra pengolahan penting yang memastikan keseragaman linguistik dalam korpus. Prosedur ini diimplementasikan melalui penggunaan pustaka *langdetect*, yang memfasilitasi identifikasi dan penghapusan elemen teks yang tidak sesuai dengan bahasa Inggris. Dengan pendekatan metodologis ini, kita memastikan bahwa dataset yang dihasilkan hanya terdiri dari konten berbahasa Inggris. Proses ini meningkatkan akurasi dan keandalan analisis dan interpretasi selanjutnya dengan menciptakan dasar linguistik yang lebih seragam dan konsisten. Penghapusan tanda baca melibatkan penghapusan karakter non-alfanumerik seperti titik, koma, tanda seru, tanda tanya, dan lainnya. Proses ini penting dalam analisis teks karena tanda baca dapat mengganggu algoritma yang digunakan untuk mengevaluasi sentimen. Biasanya, tanda baca dianggap berlebihan dalam konteks analisis sentimen karena tidak berkontribusi pada makna emosional teks. Dengan menghapus tanda baca, kita memastikan fokus tetap pada kata-kata yang merupakan pembawa utama sentimen.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

Penghapusan emotikon dilakukan dengan mengganti substrings emotikon yang terdeteksi dengan spasi menggunakan fungsi *replace_emoji* dari pustaka emoji. Meskipun emotikon dapat menyampaikan sentimen, mereka menunjukkan inkonsistensi dan mungkin tidak diakui secara universal oleh alat analisis, sehingga perlu dihapus. Tokenisasi, yang melibatkan pemecahan teks menjadi kata-kata, adalah proses dekomposisi data tekstual menjadi unit-unit diskrit, terutama kata-kata individu. Pendekatan metodologis ini memfasilitasi analisis yang lebih rinci. Setelah tokenisasi, proses ini mengubah kata-kata non-standar menjadi bentuk standar mereka dan mencakup beberapa langkah penting: menghapus kata-kata umum yang dianggap tidak bermakna (*stopwords*), mengonversi semua kata ke huruf kecil, menghapus karakter duplikat dalam kata, dan menerapkan stemming untuk mengurangi kata ke bentuk dasar atau akar mereka. Pendekatan metodologis ini memastikan keseragaman semua elemen tekstual, mengubahnya ke format yang konsisten (misalnya, karakter huruf kecil). Keseragaman ini penting untuk menjaga konsistensi dan keterbandingan analitis di seluruh dataset.

Elemen numerik mungkin tidak relevan untuk jenis analisis tekstual tertentu, terutama dalam analisis sentimen atau pemodelan topik. Penghapusan mereka dapat membantu memfokuskan analisis pada konten tekstual yang relevan. *URL* dan referensi pengguna (misalnya, @username dalam posting media sosial) sering kali tidak relevan dengan analisis konten inti dan dapat menambah kebisingan. Penghapusan elemen-elemen ini membantu memfokuskan analisis pada konten tekstual utama. Tabel 2 berikut menunjukkan proses preprocessing dilakukan.

Tabel 2. Penerapan Preprocessing

ID	Sebelum penerapan Preprocessing	Setelah Penerapan Preprocessing
117	Very best app to share your daily life fellings and also share news to everyone around this world ðŸ˜Š	best share daily life also share news everyone around world
207	I love the app, however. Out if curiosity, I changed my year of birth from 2006 to 1999 cuz to see what would be different on a new account. And both my main instagram and Facebook account are stuck at the year 1999 for my year of birth and it says I have to wait a few days. Can I contact instagram and or Facebook to get this mistake changed quickly please?	love however curiosity year birth see would different new account main account stuck year year birth wait days contact get mistake quickly please

3.3 Data Labelling

Setelah melalui tahap preprocessing, proses berikutnya adalah tahap labelling yang dimana data akan dibagi menjadi dua sentimen yaitu sentimen “positif” dan sentimen ‘Negatif’ yang dimana menggunakan integrasi pustaka sentimen Vader. Tabel 3 menunjukan proses data labelling.dilakukan.

Tabel 3. Penerapan Tokenisasi Unigram

ID	Setelah Penerapan Preprocessing	Sentiment
117	best share daily life also share news everyone around world	Positive
207	love however curiosity year birth see would different new account main account stuck year year birth wait days contact get mistake quickly please	Positive

3.4 N-Gram

Penelitian ini memanfaatkan tiga tingkat tokenisasi N-gram, masing-masing menawarkan wawasan unik mengenai struktur linguistik teks. Tokenisasi Unigram, Bigram, dan Trigram digunakan untuk menganalisis teks dari berbagai perspektif, dengan setiap metode memberikan pendekatan yang berbeda dalam pemahaman dan analisis teks.

3.4.1 Tokenisasi Unigram

Tokenisasi Unigram adalah metode dasar yang membagi teks menjadi token individu, biasanya berupa kata atau karakter. Unigram, yang mewakili unit analisis yang paling atomik, memungkinkan peneliti untuk memeriksa item leksikal secara terpisah. Pendekatan ini sangat berguna untuk tugas-tugas seperti analisis kosakata, perhitungan frekuensi istilah, dan klasifikasi teks dasar. Dengan memfokuskan pada token tunggal, Unigram membantu dalam menentukan frekuensi kemunculan kata serta istilah-istilah penting dalam teks. Tabel 4 menunjukan proses tokenisasi unigram dilakukan.

Tabel 4. Penerapan Tokenisasi Unigram

ID	Setelah Penerapan Preprocessing	Penerapan Tokenisasi Unigram

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

117	best share daily life also share news everyone around world	[‘best’, ‘share’, ‘daily’, ‘life’, ‘also’, ‘share’, ‘news’, ‘everyone’, ‘around’, ‘world’]
207	love however curiosity year birth see would different new account main account stuck year year birth wait days contact get mistake quickly please	['love', 'however', 'curiosity', 'year', 'birth', 'see', 'would', 'different', 'new', 'account', 'main', 'account', 'stuck', 'year', 'year', 'birth', 'wait', 'days', 'contact', 'get', 'mistake', 'quickly', 'please']

3.4.2 Tokenisasi Bigram

Tokenisasi Bigram, di sisi lain, menangkap ketergantungan jangka pendek dalam teks, memberikan wawasan berharga tentang asosiasi kata yang langsung dan kolokasi. Metode ini menganalisis pasangan kata yang berurutan, memungkinkan identifikasi pola hubungan kata yang lebih jelas dan pengertian tentang bagaimana kata-kata saling terkait dalam konteks yang lebih luas. Bigram berguna untuk memahami kolokasi kata dan hubungan kontekstual antara dua kata yang sering muncul bersamaan. Tabel 5 menunjukan penerapan tokenisasi Bigram dilakukan

Tabel 5. Penerapan Tokenisasi Bigram

ID	Setelah Penerapan Preprocessing	Penerapan Tokenisasi Bigram
117	best share daily life also share news everyone around world	['best share', 'share daily', 'daily life', 'life also', 'also share', 'share news', 'news everyone', 'everyone around', 'around world']
207	love however curiosity year birth see would different new account main account stuck year year birth wait days contact get mistake quickly please	['love however', 'however curiosity', 'curiosity year', 'year birth', 'birth see', 'see would', 'would different', 'different new', 'new account', 'account main', 'main account', 'account stuck', 'stuck year', 'year year', 'year birth', 'birth wait', 'wait days', 'days contact', 'contact get', 'get mistake', 'mistake quickly', 'quickly please']

3.4.3 Trigram

Tokenisasi Trigram lebih lanjut memperluas konteks dengan mempertimbangkan urutan tiga kata atau karakter berturut-turut. Trigram memberikan representasi yang lebih mendalam dari struktur linguistik lokal dan dapat mengungkapkan pola penggunaan kata yang lebih kompleks. Dengan menganalisis triplet kata, Trigram memungkinkan peneliti untuk mengidentifikasi pola berulang dan struktur frasa yang lebih kompleks dalam teks. Pada tabel 6 menunjukan penerapan tokenisasi trigram dilakukan.

Tabel 6. Penerapan Tokenisasi Trigram

ID	Setelah Penerapan Preprocessing	Penerapan Tokenisasi Trigram
117	best share daily life also share news everyone around world	['best share daily', 'share daily life', 'daily life also', 'life also share', 'also share news', 'share news everyone', 'news everyone around', 'everyone around world']
207	love however curiosity year birth see would different new account main account stuck year year birth wait days contact get mistake quickly please	['love however curiosity', 'however curiosity year', 'curiosity year birth', 'year birth see', 'birth see would', 'see would different', 'would different new', 'different new account', 'new account main', 'account main account', 'main account stuck', 'account stuck year', 'stuck year year', 'year year birth', 'year birth wait', 'birth wait days', 'wait days contact', 'days contact get', 'contact get mistake', 'get mistake quickly', 'mistake quickly please']

Secara keseluruhan, penerapan Unigram, Bigram, dan Trigram dalam penelitian ini memberikan pendekatan yang komprehensif untuk menganalisis teks. Unigram menawarkan analisis dasar dari item leksikal, Bigram mengeksplorasi hubungan kata yang lebih dekat, dan Trigram memberikan gambaran lebih dalam tentang pola linguistik yang lebih kompleks. Metode-tokenisasi ini memungkinkan peneliti untuk mengeksplorasi dan memahami berbagai aspek dari teks untuk analisis sentimen yang lebih efektif.

3.5 Evaluasi kinerja model

Dalam penelitian ini, 20% data akan digunakan sebagai data pengujian sementara 80% data akan digunakan sebagai data pelatihan dengan rasio 8:2. Dengan menerapkan N-grams, Penelitian ini menghasilkan hasil-hasil sebagai dibawah ini:

a. Dataset A (20.000 data):
   Berdasarkan Tabel 7, metode Unigram menunjukan kinerja terbaik dengan akurasi 75%, presisi 77 %, recall 74% dan F1-Score sebesar 73%.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

Tabel 7. Hasil dengan Menggunakan Dataset A (20.000 Data)

Testing	Accuracy	Precision	Recall	F1-Score
Unigram	0.75	0.77	0.74	0.73
Bigram	0.73	0.76	0.72	0.71
Trigram	0.69	0.71	0.68	0.67

b. Dataset B (199 data):
Berdasarkan Tabel 8, metode Unigram juga menunjukan kinerja terbaik dengan akurasi 70% , presisi 72%, recall 70% dan F1-Socre sebesar 70%.

Tabel 8. Hasil dengan Menggunakan Dataset B (199 Data)

Testing	Accuracy	Precision	Recall	F1-Score
Unigram	0.70	0.72	0.70	0.70
Bigram	0.65	0.42	0.65	0.51
Trigram	0.65	0.42	0.65	0.51

c. Dataset B (398 data):
Berdasarkan Tabel 9, metode Unigram, Bigram dan Trigram menunjukan kinerja terbaik dam konsisten dengan akurasi 92% , presisi 84%, recall 91% dan F1-Score sebesar 87%.

Tabel 9. Hasil dengan Menggunakan Dataset C (398 Data)

Testing	Accuracy	Precision	Recall	F1-Score
Unigram	0.92	0.84	0.91	0.87
Bigram	0.92	0.84	0.91	0.87
Trigram	0.92	0.84	0.91	0.87

Dataset C memberikan hasil yang paing konsisten dan akurat, meskipun ukurannya lebih kecil dari Dataset A. Dataset A menunjukan hasil yang sangat baik dengan Unigram, Sementara Dataset B memiliki kinerja yang kurang optimal terutama pada metode Bigram dan Trigram

4. KESIMPULAN

Dalam Penelitian ini dilakukan studi perbandingan analisis sentimen pada aplikasi Instagram menggunakan tiga dataset berbed, yaitu dataset A yang berskala besar serta Dataset B dan C yang lebih terfokus. Hasilnya menunjukan bahwa terdapat perbedaan akurasi yang signifikan antara ketida dataset karena pengaruh faktor-faktor seperti ukuran sampel, keragaman komposisi data, pertimbangan temporal, dan potensi bias pengambilan sampel.Dataset A yang lebih besar memberikan gambaran yang lebih representatif dan stabil, sementara Dataset B dan C memberikan wawasan yang lebih terokus. Perbedaan periode pengumpulan data juga mempengaruhi hasil, mencerminkan fase pengembangan aplikasi atau fluktuasi kepuasan pengguna yang berbeda.

Secara keseluruhan, studi ini menggarisbawahi bahwa ukuran dataset, komposisi, dan faktor temporal memiliki peran yang signifikan dalam menghasilkan hasil analisis sentimen yang valid dan bermanfaat. Hasil penelitian menunjukkan bahwa penggunaan Dataset C dengan metode Unigram menghasilkan akurasi sebesar 92%, presisi sebesar 84%, recall sebesar 91%, dan F1-Score sebesar 87%. Pendekatan yang cermat dalam pemilihan dataset dan interpretasi hasil adalah kunci untuk mengembangkan model analisis sentimen yang efektif dan dapat diandalkan dalam konteks aplikasi media sosial.

REFERENCES

[1] K. J. Syahrina, N. Siregar and N. Harahap, "PENELITIAN TENTANG INSTAGRAM," *Maktabun: Jurnal Perpustakaan dan Informasi*, pp. 20 - 26, 2022.

[2] C. M. Annur, "Indonesia Jadi Negara dengan Pengguna Instagram Terbanyak ke-4 di Dunia," Databoks, 28 11 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/11/28/indonesia-jadi-negara-dengan-pengguna-instagram-terbanyak-ke-4-di-dunia. [Accessed 21 april 2024].

[3] N. Julius, "Data Jumlah Pengguna Instagram di Indonesia 2024," Upgraded, 14 06 2024. [Online]. Available: https://upgraded.id/data-jumlah-pengguna-instagram-di-indonesia. [Accessed 18 06 2024].

[4] J. Callaham, "From Android Market to Google Play: a history of the Play Store," Android Authority, 2017 03 2017. [Online]. Available: https://www.androidauthority.com/android-market-google-play-history-754989/. [Accessed 12 05 2024].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 921−929
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5615

[5] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara and M. Nappi, "Discrepancy Detection between Actual User Reviews and Numeric Ratings of Google App Store using Deep Learning," *Expert Systems with Applications*, vol. 181, pp. 1-11, 2021.

[6] Y. Asri, W. N. Suliyanti, D. Kuswardani and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," *PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika*, vol. 15, no. 2, pp. 264-275, 2022.

[7] D. S. Sayogo, B. Irawan and A. Bahtiar, "ANALISIS SENTIMEN ULASAN INSTAGRAM DI GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAÏVE BAYES," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3314-3319, 2023.

[8] T. Widiyaningtyas, I. A. Zaeni and R. A. Farisi, "Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Semarang, 2019.

[9] R. B. Trianto, A. Triyono and D. M. P. Arum, "Klasifikasi Rating Otomatis pada Dokumen Teks Ulasan Produk Elektronik Menggunakan Metode N-gram dan Naïve Bayes," *Jurnal Informatika Universitas Pamulang* , vol. 5, no. 3, pp. 295-301, 2020.

[10] L. A. Sari, N. F. Ramadhita and F. N. Hasan, "ANALYSIS OF PUBLIC SENTIMENT ON GOOGLE PLAY STORE TIJE APPLICATION USERS USING NAÏVE BAYES CLASSIFIER METHOD," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 1, pp. 243-251, 2024.

[11] A. Sunjaya, N. Wijaya, N. P. Wong, S.KOM., M.T.I and S. Winardi, S.KOM., M.T.I, "Implementation of Sentiment Analysis of Shopee E-Commerce Reviews using Naïve Bayes, N-Gram, and Information Gain," in *2023 Eighth International Conference on Informatics and Computing (ICIC) (2023)*, Malang, 2023.

[12] A. Bayhaqy, K. Nainggolan, S. Sfrenrianto and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," 2012.

[13] D. Abimanyu, E. Budianita, E. P. Cynthia, F. Yanto and Y. , "Analisis Sentimen Akun Twitter Apex Legends Menggunakan VADER," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 5, no. 3, pp. 423-431, 2022.

[14] D. D. Suhartono,S.Kom., M.T.I, "N-Gram," BINA NUSANTARA : sCHOOL OF COMPUTER SCIENCE, 19 12 2019. [Online]. Available: https://socs.binus.ac.id/2019/12/31/n-gram/. [Accessed 01 03 2024].

[15] D. Septiani and I. Isabela, "ANALISIS TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DALAM TEMU KEMBALI INFORMASI PADA DOKUMEN TEKS," *SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia*, vol. 1, no. 2, pp. 81-88, 2022.

[16] E. Indarbensyah and N. Rochamawati, "Penerapan N-gram Menggunakan Algoritma Random Forest dan Naives Bayes Classifier pada Analisis Sentimen Kebijakan PPKM 2021," *JINACS: Journal of Informatics and Computer Science*, vol. 2, no. 4, pp. 235-244, 2021.

[17] I. E. Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review," *JOSCEK : Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 1-7, 2020.

[18] D. H. Kalokasari, D. M. Shofi and A. H. Setyaning, "IMPLEMENTASI ALGORITMA MULTINOMIAL NAIVE BAYES CLASSIFIER PADA SISTEM KLASIFIKASI SURAT KELUAR (Studi Kasus : DISKOMINFO Kabupaten Tangerang)," *JURNAL TEKNIK INFORMATIKA*, vol. 10, no. 2, pp. 109-118, 2017.

[19] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon and A. Ahmed, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS International Journal of Computer Science and Network Security*, vol. 19, no. 3, pp. 62-67, 2019.

[20] E. H. Muktafin, K. and E. T. Luthfi, "Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing," *Jurnal Eksplora Informatika*, vol. 10, no. 1, pp. 32-42, 2020.

[21] U. Rahardja, T. Hariguna and W. M. Baihaqi, "Opinion mining on e-commerce data using sentiment analysis and k-medoid clustering," in *Proceedings - 2019 12th International Conference on Ubi-Media Computing, Ubi-Media 2019*, 2019.

[22] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - Journal of Business and Audit Information Systems*, vol. 4, no. 2, 2021.

[23] L. Ardiani, H. Sujaini and T. , "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *Jurnal Sistem dan Teknologi Informasi (Justin)*, vol. 8, no. 2, p. 183, 2020.

[24] P. A. Sumitro, R. D. I. Mulyana and W. Saputro, "Analisis Sentimen Terhadapat Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based," *Jurnal Informatika dan Teknologi Komputer*, pp. 50-57, 2021.