ANALISIS SENTIMEN OPINI PENGGUNA APLIKASI VIDIO PADA ULASAN *PLAYSTORE* MENGGUNAKAN ALGORITMA *NAIVE BAYES*

Muhammad Yusuf Rismanda Gaja, Iqbal Maulana, Oman Komarudin
Program Studi Teknik Informatika S1, Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang, Karawang Jawa Barat
1910631170109@student.unsika.ac.id

ABSTRAK

Peningkatan popularitas layanan streaming online sejalan dengan perkembangan teknologi informasi saat ini. Salah satu aplikasi streaming video yang terkenal di Indonesia adalah Vidio. Meskipun Vidio menawarkan beragam film, serial, dan siaran langsung, namun penilaian aplikasi ini melalui komentar dan ulasan pengguna di Google Playstore tidak begitu positif. Oleh karena itu, diperlukan analisis terhadap ulasan tersebut guna memahami pandangan pengguna terhadap aplikasi ini. Penelitian ini menggunakan 1259 data yang dianalisis melalui pelabelan menggunakan kamus Inset dan paket valder dalam RStudio. Klasifikasi data dilakukan dengan algoritma Naïve Bayes serta seleksi fitur TF-IDF. Hasil klasifikasi menunjukkan bahwa akurasi tertinggi diperoleh ketika data dilabeli dengan metode Inset pada persentase 90% data pelatihan dan 10% data pengujian, mencapai 76%. Sementara itu, akurasi tertinggi dengan pelabelan menggunakan *Vader* hanya mencapai 73% pada skema pembagian data 70:30. Nilai presisi tertinggi ditemukan pada pengujian dengan pembagian data 90:10 menggunakan metode Inset, mencapai 73%. Namun, akurasi tertinggi dengan pelabelan *Vader* hanya mencapai 72,6% pada skema pembagian data 90:10. Selain itu, pengujian menunjukkan bahwa nilai recall tertinggi diperoleh pada skema pembagian data 80:20 dengan metode Inset, mencapai 81%. Akan tetapi, akurasi tertinggi dengan pelabelan *Vader* hanya mencapai 79% pada pembagian data 70:30. Kesimpulan dari hasil analisis ini adalah bahwa pelabelan menggunakan metode *Inset* lebih optimal dalam menganalisis sentimen dalam teks berbahasa Indonesia, dibandingkan dengan penggunaan metode *Vader*.

Kata kunci: *Analisis Sentimen, Vidio, Naïve bayes*

1. PENDAHULUAN

Perkembangan teknologi informasi terkini telah memicu lonjakan penonton yang beralih dari tayangan televisi ke layanan streaming online. Lebih lanjut, popularitas layanan ini didorong oleh kehadiran aplikasi streaming berbasis mobile yang memungkinkan masyarakat menikmati tayangan visual secara fleksibel. Fenomena ini semakin mencuat di Indonesia setelah pandemi Covid-19 mendorong masyarakat untuk berlangganan aplikasi streaming video. Dalam konteks ini, Vidio, sebuah aplikasi streaming video untuk ponsel android, telah menjadi perhatian. Diluncurkan pada 19 April 2015 di Google Playstore, Vidio adalah aplikasi streaming video pertama buatan Indonesia yang menawarkan 21 saluran gratis, termasuk film lokal dan internasional serta acara olahraga nasional dan internasional. Dengan basis pelanggan yang terus bertambah, mencapai 62 juta pengguna, dan lebih dari 2,3 juta pelanggan berbayar, Vidio telah membuktikan popularitasnya [1]. Analisis sentimen, sebagai disiplin ilmu, muncul untuk menganalisis pendapat, opini, penilaian, evaluasi, sikap, dan emosi individu terhadap berbagai entitas seperti produk, layanan, organisasi, individu, isu, acara, topik, dan atribut [2]. Dalam kerangka analisis sentimen, komentar dan ulasan digolongkan sebagai sentimen positif dan negatif, memberikan pandangan yang jelas terhadap kelebihan dan kekurangan layanan aplikasi. Hasil dari analisis ini bermanfaat untuk meningkatkan mutu produk dan layanan. Lebih lanjut, analisis sentimen mengubah informasi yang awalnya tidak terstruktur menjadi data terstruktur. Di antara berbagai metode dalam text mining, *Naïve Bayes* menjadi pilihan untuk mengklasifikasikan sentimen. Metode ini mengadaptasi *Teorema Bayes* yang diperkenalkan oleh ilmuwan Inggris, *Thomas Bayes*. Studi sebelumnya telah membuktikan keefektifan *Naïve Bayes* dalam analisis dokumen [3]. *Naïve Bayes* digunakan karena memiliki beberapa kelebihan yaitu cepat, sederhana, dan menghasilkan hasil prediksi yang cukup baik. Metode ini menawarkan kecepatan, kesederhanaan, dan hasil prediksi yang akurat. Oleh karena itu, penelitian ini mengusung tujuan mengklasifikasikan ulasan pengguna aplikasi Vidio di Google Play Store menjadi sentimen positif dan negatif menggunakan metode *Naïve Baye*s. Harapannya, hasil dari penelitian ini akan memberikan klasifikasi ulasan yang akurat dan rekomendasi evaluasi yang berguna bagi pengguna dan pengembang aplikasi dalam meningkatkan kualitas layanan.

2. TINJAUAN PUSTAKA

2.1. Vidio

Vidio adalah platform streaming video yang didirikan pada tahun 2014. Situs ini memberi penggunanya keuntungan untuk menonton dan menikmati banyak video dan layanan lain melalui internet dan streaming. Vidio dapat digunakan melalui platform seluler dan tablet (iOS, Android), smart TV Vidio (Vidio, 2021)

2.2. *Google Play Store*

Google Playstore adalah platform *Google* yang menyediakan berbagai konten digital kepada pengguna. Semula bernama Android Market, dirilis pada 22 Oktober 2008, kemudian berganti nama menjadi Google Playstore pada Maret 2012. Konten digital yang disediakan *Google Playstore* meliputi game, film, musik, dan buku [4] . Selain itu, *Google Play Store* menawarkan kesempatan kepada pengguna untuk menilai konten mereka dengan menyediakan fungsi ulasan. Fitur ini memungkinkan pengguna yang telah mencoba dan menggunakan layanan yang disediakan Google Playstore untuk memberikan ulasan dan komentar tentang produk atau layanan tersebut.

2.3. Ulasan

Definisi ulasan menurut Kamus Besar Bahasa Indonesia (KBBI) adalah tafsiran, komentar, dan kupasan. Menurut [5] kecenderungan seorang pengguna suatu produk atau layanan aplikasi ingin mengetahui pendapat dan pengalaman terhadap produk dan layanan aplikasi tersebut.

2.4. Analisis Sentimen

Analisis sentimen adalah sebuah studi yang menganalisis pendapat, sentimen, penilaian, sikap, evaluasi, dan emosi seseorang terhadap entitas seperti sebuah produk, layanan, organisasi, individual, masalah, topik atau peristiwa, dan semacamnya. Analisis Sentimen digunakan untuk mendeteksi suatu opini kedalam opini positif atau negatif dalam suatu teks, yang dapat berupa dokumen, paragraf, kalimat ataupun klausa [6]. Analisis sentimen dapat juga disebut dengan Opinion Mining karena dalam prosesnya kita harus melihat dan menganalisis emosi yang terdapat pada suatu tulisan, seperti tulisan komentar pelanggan pada suatu produk dan pelayanan ataupun tulisan pengguna sosial media seperti facebook dan twitter.

2.5. *Knowledge Discovery In Database*

Knowledge Discovery in Database (KDD) adalah Istilah yang sering digunakan untuk menggambarkan proses penemuan pengetahuan dalam basis data besar. Tahapan-tahapan dalam proses KDD mencakup pemilihan data, pembersihan data, transformasi data, eksplorasi data, dan evaluasi hasil. Tahap pemilihan data melibatkan seleksi data dari basis data operasional dan menyimpan data yang dipilih dalam sebuah berkas terpisah. Tahap pembersihan data melibatkan proses membuang duplikasi, memeriksa inkonsistensi dan kesalahan pada data [7].

2.6. *Text Mining*

Text mining, juga dikenal sebagai *Text Data Mining* (TDM) atau *Text Knowledge Mining* (KDT), adalah bidang khusus dari data mining yang didefinisikan sebagai data mining dalam bentuk teks. Sumber data biasanya diambil dari dokumen dengan tujuan untuk menemukan frase yang dapat mewakili isi dokumen sehingga dapat dilakukan analisis keterkaitan antar dokumen. Biasanya dokumen tersebut tidak dalam bentuk yang terstruktur. Hal ini disebabkan karakteristik data tekstual yang lebih kompleks. Oleh karena itu, diperlukan mekanisme yang tepat untuk menggali teks-teks tersebut untuk mendapatkan informasi yang lebih terstruktur dan berharga. Langkah-langkah preprocessing text mining meliputi pembersihan, pelipatan kasus, enkripsi, penyaringan, penghapusan [8].

2.7. *Web Scraping*

Web Scraping adalah suatu teknik pengambilan suatu informasi atau data yang sangat banyak untuk digunakan sebagai keperluan riset, analisis dan lain-lain. Mengambil data dari website lalu disimpan kedalam database untuk dianalisis. *Web Miner* merupakan alat untuk mengambil data dari situs web. Salah satu contoh alat *Web Scraping* yang disediakan oleh Google adalah *Data Miner* [9].

2.8. *Data Miner*

Data Miner merupakan sebuah ekstensi pada browser Google Chrome yang dikembangkan oleh Data Miner untuk mengikis data yang terdapat dalam halaman web lalu mengekspor halaman web tersebut menjadi file XLS, CSV, XLSX, atau TSV.

2.9. Algoritma *Naïve Bayes*

Algoritma *Naive Bayes* adalah metode klasifikasi probabilistik sederhana berdasarkan teorema Bayes yang diajukan oleh ilmuwan Inggris Thomas Bayes, yang mengasumsikan tingkat kemandirian yang tinggi dari semua kondisi dan peristiwa [10]. Meskipun metode operasinya sederhana, namun kinerja metode ini sebanding dengan algoritma lain seperti pohon keputusan dan jaringan saraf C. Salah satu keunggulan dari algoritma Naive Bayes adalah kemampuannya untuk menghasilkan parameter klasifikasi dengan menggunakan sejumlah kecil data pelatihan. Hal ini membuatnya menjadi pilihan yang lebih akurat saat diterapkan pada kumpulan data yang besar.

2.10. Metode *Lexicon*

Metode leksikon adalah pendekatan yang menggunakan pendekatan berbasis kamus atau kamus. *Lexicon* adalah daftar leksikal, atau yang disebut daftar kosa kata, yang berisi informasi tentang bagian-bagian pembicaraan. Kata-kata ini sering mengandung kata sifat, kata kerja, atau kata keterangan yang digunakan sebagai indikator kalimat pendapat, seperti baik, buruk, indah, mudah, atau sulit. Untuk menggunakan metode ini, kita memerlukan daftar kata yang diberi label dengan nilai polaritasnya. Nilai dari -1 (untuk kelas negatif) hingga +1 (untuk kelas positif). Aplikasi itu dapat menggunakan paket apa pun yang tersedia untuk pelabelan. *syuzhet*, *Vader*, *NLC*, dll [11].

2.11. *Term Frequency Inverse Document Frequency* (TF-IDF)

Term Frequency Inverse Document Frequency (TF-IDF) adalah sebuah algoritma yang digunakan untuk mengubah data dari data tekstual ke dalam data numerik untuk dilakukan pembobotan pada tiap kata atau fitur. TF *(Term Frequency)* adalah seberapa banyak kemunculan setiap kata pada dokumen sedangkan IDF *(Inverse Document Frequency)* merupakan jumlah dokumen terkait yang memiliki kata tertentu. Jadi TF – IDF adalah gabungan dari TF dan IDF dimana keduanya dapat menutupi kekurangan untuk menjadikan prediksi menjadi relevan [12].

2.12. *Confusion Matrix*

Confusion matrix adalah sebuah metode yang berbentuk matrix berfungsi untuk mencatat dan menghitung serta mengevaluasi suatu kinerja algoritma yang berbasis klasifikasi atau *classifier*. *Confusion matrix* berisi informasi mengenai perbandingan antara hasil klasifikasi oleh suatu model dengan hasil aktualnya [13].

2.13. *Rstudio*

Rstudio adalah platform open source dengan bahasa pemrograman R yang digunakan untuk mengolah data untuk Windows, Linux, Macintosh, UNIX. R menyediakan lebih dari 6000 paket (package), sehingga terbilang lengkap dan banyak. saat ini R sudah dikenal luas sebagai salah perangkat lunak yang kuat untuk analisis data dan Data Science. *RStudio* merupakan *Integrated Development Environment* (IDE) untuk R yang banyak digunakan saat ini.

2.14. *Wordcloud*

Wordcloud adalah Suatu hasil yang diperoleh dari metode penambangan teks adalah kemampuan untuk memvisualisasikan istilah-istilah populer yang terkait dengan kata kunci seperti "internet" dan "data teks". Word cloud digunakan sebagai alat untuk memeriksa istilah-istilah yang paling umum atau tren berdasarkan seberapa sering mereka muncul dalam teks. Ukuran kata dalam word cloud mencerminkan frekuensinya; semakin besar kata tersebut, semakin sering muncul dalam teks, dan sebaliknya.

3. METODE PENELITIAN

Metodologi yang digunakan dalam penelitian ini yaitu *Knowledge Discovery in Database* (KDD) dengan tahapannya yaitu *data selection, data preprocessing, data transformation, data mining* dan *interpretation/evaluation.* Pada tahap *data preprocessing* akan dilakukan beberapa tahapan yaitu diantaranya *Cleaning, Case folding, Tokenizing, Filtering, Stemming.* Adapun tahapan KDD sebagai berikut:

3.1. *Data Selection*

Pada tahap awal ini, data-data pada database akan diseleksi kemudian disimpan didalam berkas yang terpisah.

3.2. *Preprocessing*

Tahapan ini merupakan tahapan yang penting dalam penelitian ini sebelum berlanjut ke tahap sebelumnya. Pada tahap ini akan dilakukan proses pembersihan data untuk mengurangi atribut yang kurang berpengaruh supaya memudahkan pengolahan data pada saat proses klasifikasi.

3.3. *Transformation*

Pada tahap transformation dilakukan transformasi data menjadi numerik menggunakan TF-IDF (Term Frequency Inverse Document Frequency) yaitu dengan memberikan bobot terhadap kemunculan kata pada sebuah dokumen. Tahapan ini bertujuan untuk mencari representasi nilai dari tiap-tiap dokumen dari suatu kumpulan data training.

3.4. *Data Mining*

Pada tahapan ini data sudah siap dan akan dilakukan proses klasifikasi menggunakan RStudio. Algoritma Naïve Bayes akan diimplementasikan untuk mengklasifikasi sebuah data. Pada tahap ini akan dilakukan perhitungan mengenai persentase ulasan sentimen positif dan sentimen negatif dan melakukan proses klasifikasi menggunakan algoritma Naïve Bayes berdasarkan data yang telah diberikan label oleh ahli bahasa. Dalam proses pengklasifikasian terdapat dari 2 proses yang dilakukan yaitu proses training dan proses testing.

3.5. *Interpretation/Ealuation*

Pada tahapan ini akan dilakukan pengujian metode Naïve Bayes terhadap hasil klasifikasi dengan menggunakan *Confusion Matrix* yang hasilnya akan diperoleh berupa nilai *accuracy, precision, recall,* dan AUC. Berdasarkan nilai-nilai tersebut akan terlihat keakuratan dari proses klasifikasi yang dilakukan.

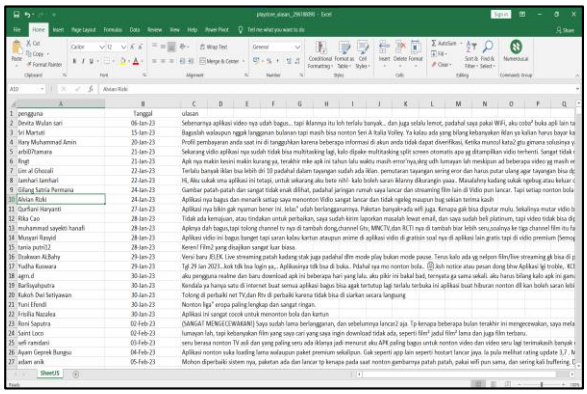
4. HASIL DAN PEMBAHASAN

Hasil penelitian yang telah dilakukan yaitu mengetahui sentimen dari opini pengguna aplikasi Vidio dengan cara mengklasifikasikan opini pengguna aplikasi Vidio pada ulasan Google Play menggunakan algoritma Naive Bayes. lalu berdasarkan hasil analisis yang didapatkan akan diberikan rekomendasi yang tepat yang nantinya akan dijadikan rujukan dalam perbaikan dan peningkatan kualitas aplikasi kedepannya.

4.1. *Data Selection*

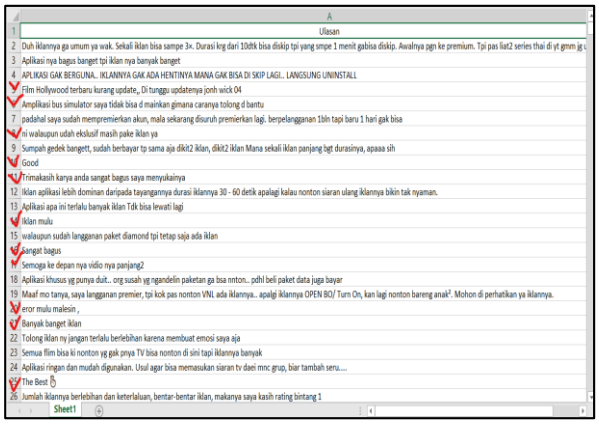
Tahap ini dimulai dengan melakukan pengambilan data ulasan menggunakan teknik *web scraping* dengan cara menambahkan ekstensi Google Chrome yang bernama Data Miner. Setelah itu membuka halaman web Google Play dan menuju halaman aplikasi Vidio lalu membuka pada bagian rating dan ulasan untuk mengambil data ulasan dari aplikasi tersebut. Data hasil *scrapping* menggunakan Data Miner terdapat menu pilihan format penyimpanan dengan 2 format yaitu *.csv* dan *.xlsx.*

Jumlah data ulasan yang berhasil diambil yaitu sebanyak 2000 data.

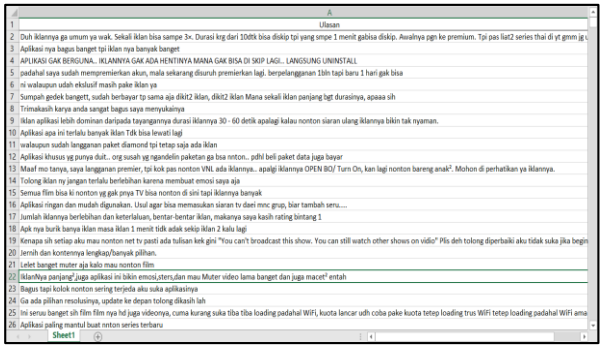


Gambar 1. Data hasil *Scrapping*

Setelah mendapatkan dataset, dataset di seleksi untuk memilih data yang sesuai dengan kriteria yang telah ditentukan lalu data ulasan berbahasa Indonesia, tidak berisi kata-kata kasar, dan merupakan kalimat yang berisi ulasan suatu aplikasi yang benar. Data ulasan yang didapatkan setelah melakukan seleksi data sebesar 1259 data ulasan. Data ulasan sebelum dan sesudah melalui tahap selection dapat dilihat pada Gambar 2 dan Gambar 3.



Gambar 2. Data ulasan sebelum melalui tahap *Selection*



Gambar 3. Data ulasan setelah melalui tahap *Selection*

4.2. *Preprocessing*

Pada tahapan ini melakukan pembersihan data dan mengurangi atribut data. Karena data yang dihasilkan dari Scrapping data mengandung banyak gangguan (*noise*) sehingga perlu dilakukan pembersihan dan pengurangan tersebut untuk menghindari adanya error dalam proses klasifikasi. Berikut ini adalah hasil dari *preprocessing* pada tabel 1.

Tabel 1. Hasil *Preprocessing*

Preprocessing	***Content***
Data Selection	Resolusi gambar utk kejernihan masih nunggu beberapa detik, pakai paket pula jadi malas buka ☺ Pindah STB jadinya ☺
Cleaning	Resolusi gambar utk kejernihan masih nunggu beberapa detik pakai paket pula jadi malas buka pindah STB jadinya
Case Folding	Resolusi gambar utk kejernihan masih nunggu beberapa detik jadi malas buka pindah stb jadinya
Tokenizing	‘resolusi’ ‘gambar’ ‘utk’ ‘kejernihan’ ‘masih’ ‘menunggu’ ‘beberapa’ ‘detik’ ‘jadi’ ‘malas’ ‘buka’ ‘pindah’ ‘stb’ ‘jadinya’
Filtering	“resolusi gambar kejernihan masih menunggu beberapa detik jadi malas buka pindah”
Stemming	“resolusi gambar jernih masih tunggu berapa detik jadi malas buka pindah”

4.3. Pelabelan Sentimen

Pada tahap ini, data ulasan yang sudah diolah sebelumnya akan diberi label secara otomatis menggunakan kamus kata dengan skor sentimen. Proses ini akan menentukan apakah sentimen dalam ulasan adalah positif atau negatif, dengan menggunakan dua label sentimen. Berdasarkan kumpulan kata bahasa Indonesia, RStudio akan secara otomatis melabelinya dengan menghitung skor kata positif dikurangi jumlah kata negatif dalam kalimat ulasan. Jika sebuah kalimat memiliki skor < 0 akan diklasifikasikan ke dalam kelas negatif, sebaliknya jika kalimat memiliki skor ≥ 0, maka akan diklasifikasikan ke dalam kelas positif. Pada tahap ini dilakukan pelabelan menggunakan dua kamus lexicon yang berbeda yaitu menggunakan package *Vader Sentiment* dan menggunakan kamus *Inset* (*Indonesia sentiment lexicon*). Berikut merupakan hasil perbandingan jumlah data dari pelabelan sentimen dengan menggunakan 2 kamus berbeda dapat dilihat pada tabel 2.

Tabel 2. Perbandingan hasil pelabelan menggunakan *Inset dan Vader*

Kelas Sentimen	**Kamus *Inset***	***Vader***
Positif	683	635
Negatif	576	624
Total	1259	

4.4. *Transformation*

Dalam proses pada data mining diperlukan dataset yang numerik karena pada dataset saat ini masih berbentuk string maka diperlukan perubahan menjadi bentuk kata. Tahapan ini melalui beberapa proses yaitu mengubah tipe data menjadi numerik dan melakukan pembobotan kata dengan menggunakan algoritma TF – IDF. Perhitungan bobot kata dilakukan dengan terlebih dahulu menentukan nilai *Term Frequency* (TF). Nilai *Term Frequency* (TF) dari data ulasan dengan dua hasil pelabelan berbeda dapat dilihat ada gambar 4 dan gambar 5.

Weighting : term frequency (tf)
Sample :

Terms / Docs	bayar	bintang	iklan	kuota	langgan	muncul	nonton	trus	video	vidio
101	0	1	0	2	0	0	1	0	0	0
106	0	0	4	0	0	1	1	0	0	0
112	4	0	0	0	0	0	2	0	0	0
113	0	1	1	0	0	0	2	0	0	1
121	0	1	0	0	0	0	1	0	3	1
135	0	0	2	0	0	4	0	2	0	0
136	2	0	1	2	0	0	4	0	0	0
138	0	0	0	0	0	0	0	0	1	0
140	0	0	4	0	0	1	0	0	0	0
147	0	0	0	0	2	0	0	2	0	1

Gambar 4. Nilai *Term Frequency* pelabelan *Inset*

Weighting : term frequency (tf)
Sample :

Terms / Docs	bayar	bintang	iklan	kuota	langgan	muncul	nonton	trus	video	vidio
101	0	1	0	2	0	0	1	0	0	0
106	0	0	4	0	0	1	1	0	0	0
112	4	0	0	0	0	0	2	0	0	0
113	0	1	1	0	0	0	2	0	0	1
121	0	1	0	0	0	0	1	0	3	1
135	0	0	2	0	0	4	0	2	0	0
136	2	0	1	2	0	0	4	0	0	0
138	0	0	0	0	0	0	0	0	1	0
140	0	0	4	0	0	1	0	0	0	0
147	0	0	0	0	2	0	0	2	0	1

Gambar 5. Nilai *Term Frequency* Pelabelan *Vader*

Hasil pembobotan kata menggunakan Term Frequency Inverse Document Frequency (TF-IDF) pada Rstudio dengan pelabelan Inset dapat dilihat pada gambar 6 dan gambar 7.

Terms / Docs	aplikasi	bagus	bayar	film	iklan	langgan	nonton	paket	video	vidio
154	0.000000	0	0.000000	0.000000	0.000000	0.000000	4.68666	0.000000	0.000000	2.574969
27	0.000000	0	6.663984	0.000000	0.000000	6.743783	0.00000	10.988117	0.000000	2.574969
326	0.000000	0	0.000000	8.794362	5.113155	0.000000	0.00000	0.000000	0.000000	0.000000
39	1.732674	0	0.000000	0.000000	0.000000	0.000000	3.12444	5.494059	0.000000	2.574969
459	0.000000	0	0.000000	0.000000	0.000000	0.000000	1.56222	0.000000	0.000000	5.149937
651	1.732674	0	3.331992	0.000000	0.000000	0.000000	3.12444	2.747029	0.000000	0.000000
700	0.000000	0	0.000000	11.725815	0.000000	0.000000	1.56222	2.747029	0.000000	2.574969
897	0.000000	0	0.000000	0.000000	3.408770	0.000000	0.00000	0.000000	0.000000	2.574969
911	0.000000	0	0.000000	0.000000	0.000000	2.247928	0.00000	0.000000	3.056985	2.574969
922	5.198022	0	0.000000	0.000000	0.000000	0.000000	1.56222	0.000000	0.000000	2.574969

Gambar 6. Bobot kata pada *Data Training* pelabelan *Inset*

Terms / Docs	apk	aplikasi	bagus	bayar	film	iklan	langgan	nonton	video	vidio
10	0.000000	3.865772	0.000000	0.000000	0	1.534336	0.00000	0.000000	2.729352	0.000000
134	3.119299	0.000000	0.000000	0.000000	0	3.068673	0.00000	5.020498	0.000000	2.889817
138	3.119299	0.000000	0.000000	6.444785	0	7.671682	0.00000	3.346998	0.000000	0.000000
199	0.000000	0.000000	0.000000	0.000000	0	0.000000	4.33985	0.000000	5.458705	0.000000
233	3.119299	0.000000	2.485427	0.000000	0	0.000000	0.00000	0.000000	0.000000	0.000000
249	9.357897	0.000000	2.485427	0.000000	0	1.534336	0.00000	3.346998	2.729352	0.000000
28	0.000000	3.865772	0.000000	6.444785	0	0.000000	4.33985	0.000000	10.917410	0.000000
37	0.000000	5.798657	2.485427	3.222392	0	0.000000	0.00000	0.000000	0.000000	0.000000
69	0.000000	1.932886	2.485427	0.000000	0	4.603009	0.00000	0.000000	0.000000	5.779634
8	6.238598	0.000000	0.000000	0.000000	0	1.534336	0.00000	0.000000	0.000000	0.000000

Gambar 7. Bobot kata pada *Data Testing* pelabelan *insert*

Hasil pembobotan kata menggunakan *Term Frequency Inverse Document Frequency* (TF-IDF) pada Rstudio dengan pelabelan *Vader* dapat dilihat pada gambar 8 dan gambar 9 berikut.

Terms / Docs	aplikasi	bagus	bayar	film	iklan	langgan	nonton	paket	video	vidio
154	0.000000	0	0.000000	0.000000	0.000000	0.000000	4.68666	0.000000	0.000000	2.566457
27	0.000000	0	6.663984	0.000000	0.000000	6.743783	0.00000	10.949773	0.000000	2.566457
326	0.000000	0	0.000000	8.794362	5.113155	0.000000	0.00000	0.000000	0.000000	0.000000
39	1.727920	0	0.000000	0.000000	0.000000	0.000000	3.12444	5.474886	0.000000	2.566457
459	0.000000	0	0.000000	0.000000	0.000000	0.000000	1.56222	0.000000	0.000000	5.132914
651	1.727920	0	3.331992	0.000000	0.000000	0.000000	3.12444	2.737443	0.000000	0.000000
700	0.000000	0	0.000000	11.725815	0.000000	0.000000	1.56222	2.737443	0.000000	2.566457
897	0.000000	0	0.000000	0.000000	3.408770	0.000000	0.00000	0.000000	0.000000	2.566457
911	0.000000	0	0.000000	0.000000	0.000000	2.247928	0.00000	0.000000	3.056985	2.566457
922	5.183761	0	0.000000	0.000000	0.000000	0.000000	1.56222	0.000000	0.000000	2.566457

Gambar 8. Bobot kata pada *data training* pelabelan *Vader*

Terms / Docs	apk	aplikasi	bagus	bayar	film	iklan	langgan	nonton	video	vidio
10	0.000000	3.865772	0.000000	0.000000	0	1.534336	0.00000	0.000000	2.729352	0.000000
134	3.119299	0.000000	0.000000	0.000000	0	3.068673	0.00000	5.020498	0.000000	2.889817
138	3.119299	0.000000	0.000000	6.444785	0	7.671682	0.00000	3.346998	0.000000	0.000000
199	0.000000	0.000000	0.000000	0.000000	0	0.000000	4.33985	0.000000	5.458705	0.000000
233	3.119299	0.000000	2.485427	0.000000	0	0.000000	0.00000	0.000000	0.000000	0.000000
249	9.357897	0.000000	2.485427	0.000000	0	1.534336	0.00000	3.346998	2.729352	0.000000
28	0.000000	3.865772	0.000000	6.444785	0	0.000000	4.33985	0.000000	10.917410	0.000000
37	0.000000	5.798657	2.485427	3.222392	0	0.000000	0.00000	0.000000	0.000000	0.000000
69	0.000000	1.932886	2.485427	0.000000	0	4.603009	0.00000	0.000000	0.000000	5.779634
8	6.238598	0.000000	0.000000	0.000000	0	1.534336	0.00000	0.000000	0.000000	0.000000

Gambar 9. Bobot kata pada *Data Testing* pelabelan *Vader*

4.5. *Data Mining*

Di tahap ini, dilakukan pengelompokan data ulasan menggunakan algoritma Naïve Bayes. Sebelumnya, data dibagi menjadi dua bagian: data latih untuk membangun model, dan data uji untuk menguji kinerja klasifikasi yang telah dibuat. Data latih digunakan untuk membentuk pola, sedangkan data uji digunakan untuk menguji seberapa baik classifier dapat melakukan klasifikasi. Pembagian data latih dan data uji pada Rstudio dapat dilihat pada gambar 10 berikut.

```
df.train <- df[1:1007,]
df.test <- df[1008:1259,]
dtm.train <- dtm[1:1007,]
dtm.test <- dtm[1008:1259,]
corpus.clean.train <- corpus[1:1007]
corpus.clean.test <- corpus[1008:1259]
```

Data latih = data ke 1-1007
Data uji = data ke 1008-1259

Gambar 10. Partisi data latih dan data uji

Pada penelitian ini dilakukan 4 kali pengujian yaitu pertama menggunakan persentase 90% data latih dan 10% data uji, pengujian yang kedua menggunakan persentase 80% data latih dan 20% data uji, serta pengujian ketiga menggunakan persentase 70% data latih dan 30% data uji. pengujian yang keempat persentase 60% data latih dan 40% data uji Berikut merupakan jumlah data latih dan data uji pada masing-masing pengujian.

Tabel 3. Pembagian Data Latih dan Data Uji

Presentasi Latih : Uji (%)	Data Latih	Data Uji
90:10	1133	126
80:20	1007	252
70:30	881	378
60:40	755	504

4.6. *Interpretation/Evaluation*

Pada tahapan ini akan dilakukan pengujian/evaluasi untuk mengukur kinerja klasifikasi data menggunakan *confusion matrix*. Hasil dari *confusion matrix* akan mendapatkan nilai *accuracy*, *precision*, dan *recall*. Berikut hasil klasifikasi data pelabelan menggunakan *inset* dan *vader* dapat dilihat pada tabel 4 dan tabel 5.

Tabel 4. Hasil *Confusion Matrix Inset*

Model	*Accuracy*	*Precision*	*Recall*
90:10	76%	73%	78%
80:20	72%	67%	79%
70:30	74%	67%	80%
60:40	72%	64%	81%

Tabel 5. Hasil *Confusion Matrix Vader*

Model	*Accuracy*	*Precision*	*Recall*
90:10	68%	73%	68%
80:20	73%	68%	76%
70:30	73%	67%	77%
60:40	71%	67%	76%

Hasil klasifikasi menggunakan *Naïve Bayes Classifier* dengan data pelabelan menggunakan leksikon *Inset* (*Indonesian sentiment lexicon*) mencapai nilai akurasi tertinggi di perbandingan 90:10 dengan presentase nilai sebesar 76%. Lalu untuk pengujian klasifikasi menggunakan data ulasan dengan pelabelan leksikon *Vader* mencapai nilai akurasi tertinggi di perbandingan 70:30 dengan presentase nilai sebesar 73%. Untuk nilai *precision* tertinggi dengan klasifikasi menggunakan data ulasan yang pelabelannya menggunakan *Inset* (*Indonesian sentiment lexicon*) diperoleh dari pengujian dengan penggunaan data 90:10 yaitu sebesar 73%, sedangkan untuk dengan klasifikasi menggunakan data ulasan yang pelabelannya menggunakan *Vader* nilai *precision* tertinggi diperoleh dari pengujian dengan penggunaan data 90:10 yaitu sebesar 72,6%. Untuk nilai *recall* tertinggi dengan klasifikasi menggunakan data ulasan yang pelabelannya menggunakan *Inset* (*Indonesian sentiment lexicon*) diperoleh dari pengujian dengan penggunaan data 60:40 yaitu sebesar 81%, sedangkan untuk dengan klasifikasi menggunakan data ulasan yang pelabelannya menggunakan *Vader* nilai *precision* tertinggi diperoleh dari pengujian dengan penggunaan data 70:30 yaitu sebesar 79%.

Berikut ini adalah Hasil Klasifikasi analisis sentimen pengguna aplikasi Vidio dapat di visualisasikan dengan menggunakan *wordcloud* untuk mengetahui gambaran kata yang paling banyak muncul dalam data ulasan pengguna aplikasi Vidio. Berikut gambar visualisasi kata dari masing masing sentimen.

Gambar 11. *Wordcloud* ulasan positif dari data pelabelan *Inset*

Gambar 12. *Wordcloud* ulasan Positif dari data pelabelan *Vader*

Berdasarkan gambar 11 dan gambar 12 dapat dilihat bahwa kata yang sering muncul sedikit berbeda. Namun kata “bagus”, “transaksi”, “video”, “paket”, dan “film” tetap menjadi kata yang paling sering muncul yang digunakan untuk ulasan positif aplikasi Vidio pada penelitian ini. Semakin besar ukuran kata dalam *wordcloud* maka semakin tinggi pula frekuensi kata tersebut.

Gambar 13. *Wordcloud* ulasan Negatif dari data pelabelan *Inset*

Gambar 14. *Wordcloud* Negatif dari data pelabelan *Vader*

Berdasarkan gambar 13 dan gambar 14 dapat dilihat bahwa kata yang sering muncul sedikit berbeda. Namun kata "iklan", "bayar"," video", "paket", dan "film" tetap menjadi kata yang paling sering muncul yang digunakan untuk ulasan negatif aplikasi Vidio pada penelitian ini. Semakin besar ukuran kata dalam word cloud maka semakin tinggi pula frekuensi kata tersebut.

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang telah dilakukan terdapat beberapa kesimpulan yaitu dari hasil pembersihan, dan pemilihan data mendapatkan total 1259 data. Dimana setelah melakukan pelabelan menggunakan 2 leksikon mendapatkan hasil yang berbeda yaitu dengan *inset* mendapatkan 683 data dengan sentimen positif dan 576 data dengan sentimen negatif, sedangkan dengan *vader* mendapatkan 635 data positif dan 624 data negatif. Ini membuktikan bahwa dengan kamus pelabelan yang berbeda, pelabelan dengan metode leksikon akan menghasilkan perbedaan hasil sentimen yang berbeda juga. Hasil *accuracy* tertinggi klasifikasi menggunakan Algoritma *Naïve Bayes* yaitu dari pengujian data yang pelabelannya menggunakan *Inset* dengan persentase penggunaan 90% data *training*, dan 10% data *tesing* sebesar 76%,sedangkan accuracy ytertinggi dengan pelabelan vader hanya 73% di penggunaan data 70:30. Untuk nilai precision tertinggi juga diperoleh dari pengujian dengan penggunaan data 90:10 menggunakan pelabelan *Inset* yaitu sebesar 73%, sedangkan *accuracy* tertinggi dengan pelabelan *vader* hanya 72,6% di penggunaan data 90:10. Lalu nilai recall tertinggi diperoleh dari pengujian dengan penggunaan data 80:20 yaitu menggunakan pelabelan *Inset* sebesar 81%, sedangkan *accuracy* tertinggi dengan pelabelan *vader* hanya 79% di penggunaan data 70:30. Dari hasil informasi tersebut membuktikan bahwa pelabelan dalam analisis sentimen berbahasa Indonesia menggunakan pelabelan *Inset* menjadi pilihan terbaik dalam melabelkan data teks sentimen dengan metode leksikon dibanding dengan menggunakan pelabelan menggunakan package *Vader*. Saran untuk penelitian ini yaitu menggunakan metode klasifikasi lain seperti *Support Vector Machine* atau *K-Nearest Neighbor*, menggunakan dataset yang lebih banyak untuk memperluas pencarian data, metode pelabelan *lexicon* seperti *syuzhet lexicon, bing lexicon, nrc lexicon* dan lain lain.

DAFTAR PUSTAKA

[1] E. M. A. Ernamia and A. Herliana, "Analisis Sentimen Kuliah Daring Dengan Algoritma Naïve Bayes, K-Nn Dan Decision Tree," *J. Responsif Ris. Sains dan Inform.*, vol. 4, no. 1, pp. 70–80, 2022, doi: 10.51977/jti.v4i1.614.

[2] B. Liu, "Web data mining: exploring hyperlinks, contents, and usage data," vol. 1, 2011, [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-19460-3

[3] S. Wahyu Handani, D. Intan Surya Saputra, Hasirun, R. Mega Arino, and G. Fiza Asyrofi Ramadhan, "Sentiment analysis for go-jek on google play store," *J. Phys. Conf. Ser.*, vol. 1196, no. 1, 2019, doi: 10.1088/1742-6596/1196/1/012032.

[4] 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*, vol. 15, no. 2, pp. 264–275, 2022, doi: 10.33322/petir.v15i2.1733.

[5] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 785–795, 2022, doi: 10.35957/jatisi.v9i2.1835.

[6] R. Apriani *et al.*, "Analisis Sentimen dengan Naïve Bayes Terhadap Komentar Aplikasi Tokopedia," *J. Rekayasa Teknol. Nusa Putra*, vol. 6, no. 1, pp. 54–62, 2019, [Online]. Available: https://rekayasa.nusaputra.ac.id/article/view/86

[7] W. Gata and P. P. Purnomo, "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS," *Format*, vol. 6, no. 1, pp. 1–13, 2017.

[8] I. B. M. C. Education, "Text Mining." Retrieved from ibm. com: https://www. ibm. com/cloud/learn/text-mining, 2020.

[9] A. Josi, L. A. Abdillah, and Suryayusra, "Penerapan teknik web scraping pada mesin pencari artikel ilmiah," 2014, [Online]. Available: http://arxiv.org/abs/1410.5777

[10] R. Marcos de Moraes, E. A. de M. G. Soares, and L. dos S. Machado, "A double weighted fuzzy gamma naive bayes classifier," *J. Intell. Fuzzy Syst.*, vol. 38, no. 1, pp. 577–588, 2020.

[11] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *2017*

International Conference on Asian Language Processing (IALP), 2017, pp. 391–394.

[12] R. Melita, V. Amrizal, H. B. Suseno, T. Dirjam, T. Informatika, and F. Sains, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam)," *J. Tek. Inf.*, vol. 11, no. 2, pp. 149–164, 2018.

[13] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021, [Online]. Available: http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369