JOURNAL OF
SOFTWARE ENGINEERING AND
COMPUTATIONAL INTELLIGENCE (JSECI)
INFORMATICS ENGINEERING, FACULTY OF COMPUTER SCIENCE
UNIVERSITAS INDO GLOBAL MANDIRI

Optimasi Akurasi Sentimen Komentar Xiaomi SU7 di YouTube Menggunakan *Naive Bayes* dan *Chi-Square*

Dicky Ryanto Fernandes[1], Nicolas Jacky Pratama Hasan[2], Novan Wijaya[3*]
[1,2]Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
[3]Manajemen Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang

[1]dicky.ryanto2303@mhs.mdp.ac.id , [2]nicolasjacky2004@mhs.mdp.ac.id , [3*]novan.wijaya@mdp.ac.id

ABSTRACT

The purpose of this study is to use the Naive Bayes and Chi-square methodologies to analyze the sentiment of comments made about the Xiaomi SU7 product on the YouTube platform. Preprocessing, dataset labeling, dataset mining, and using the SMOTE approach to address class imbalance are the primary steps of this study methodology. The process of data mining involves gathering user comment data from Xiaomi SU7-related YouTube videos. Following data collection, a labeling process is performed to categorize comments into positive, negative, and neutral sentiment categories. The preparation step involves removing extraneous components from the data, such as special characters, numerals, and punctuation. Next, we utilize the Synthetic Minority Oversampling Technique (SMOTE) to address the issue of class imbalance, which arises when the proportion of comments expressing a particular attitude is lower than the total number of comments. According to the study's findings, the Naive Bayes model's accuracy was 76.1%; however, the accuracy was raised to 78.5% by using chi-square feature selection. This demonstrates how applying chi-square can raise sentiment prediction accuracy by 2.4%. The study's findings indicate that the Naive Bayes model with chi-square feature selection performs better than the Naive Bayes model without this feature selection in predicting the sentiment of user comments. Businesses may create more efficient marketing strategies and react to customer feedback faster by having a better knowledge of user emotions. Additionally, the study's findings lay the groundwork for future advancements in sentiment analysis, particularly as they relate to digital marketing and product reputation management. As a result, this research advances both academic knowledge and the corporate community's ability to comprehend and address the dynamics of consumer psychology in the contemporary digital era.

Keywords: *Chi-Square, Naive Bayes, Sentiment Analysis, SMOTE*

ABSTRAK

Tujuan dari penelitian ini adalah untuk menganalisis sentimen komentar terkait produk Xiaomi SU7 pada platform YouTube dengan menggunakan metode Naive Bayes dan Chi-square. Metode penelitian ini meliputi beberapa tahapan utama: mining dataset, pelabelan dataset, preprocessing, dan penerapan metode SMOTE untuk mengatasi ketidakseimbangan kelas. Penambangan data dilakukan dengan cara mengumpulkan data komentar pengguna dari video YouTube terkait Xiaomi SU7. Setelah data terkumpul, kemudian dilakukan langkah pelabelan untuk mengklasifikasikan komentar menjadi sentimen positif, negatif, atau netral. Tahap preprocessing meliputi pembersihan data dari unsur-unsur yang tidak diperlukan seperti tanda baca, angka, dan karakter khusus. Kami kemudian menerapkan Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi masalah ketidakseimbangan kelas, dimana jumlah komentar dengan sentimen tertentu lebih sedikit dibandingkan jumlah komentar lainnya. Hasil penelitian ini menunjukkan bahwa akurasi model Naive Bayes mencapai 76,1%, sedangkan penggunaan seleksi fitur chi-square meningkatkan akurasi menjadi 78,5%. Hal ini menunjukkan bahwa penggunaan chi-square dapat meningkatkan akurasi prediksi sentimen sebesar 2,4%. Kesimpulan dari penelitian ini adalah model Naive Bayes dengan pemilihan fitur chi-square lebih efektif dalam memprediksi sentimen komentar pengguna dibandingkan model Naive Bayes tanpa pemilihan fitur tersebut. Dengan memahami emosi pengguna, bisnis dapat mengembangkan strategi pemasaran yang lebih efektif dan merespons opini konsumen dengan lebih cepat. Hasil penelitian ini juga memberikan landasan untuk pengembangan lebih lanjut di bidang analisis sentimen, khususnya dalam konteks pemasaran digital dan manajemen reputasi produk. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi pada pemahaman akademis tetapi juga memiliki implikasi praktis yang penting bagi dunia usaha dalam memahami dan merespons dinamika psikologi konsumen di era digital modern.

Vol. 2 No. 1
Article info: *Submitted June 5, 2024*
Revised June 10, 2024, Accepted June 20,2024

Kata kunci: Analisis Sentimen, *Chi-Square*, *Naive Bayes*, SMOTE

1. PENDAHULUAN

Kemunculan platform media sosial secara signifikan mempengaruhi perilaku konsumen dalam berbagai aktivitas pemasaran. Saat ini, perusahaan dapat dengan cepat menjalankan dan melaksanakan strategi mereka tanpa batasan ruang dan waktu [1][2].Berkat fitur-fitur yang disediakan platform media sosial seperti komentar dan opsi live chat, konsumen Indonesia berkesempatan berinteraksi langsung dengan brand favorit, selebriti, dan sesama pengguna di ruang digital [2][3].Analisis sentimen adalah analisis komputasi atas opini, peringkat, dan sentimen pihak lain dalam entitas, peristiwa, dan atribut [4]. Mempelajari analisis sentimen dapat memberikan informasi berharga. Analisis sentimen di jejaring sosial seperti Twitter, YouTube, dan Facebook telah menjadi alat yang ampuh untuk mempelajari lebih lanjut opini pengguna [5]. Analisis sentimen adalah metode menganalisis data untuk mengetahui perasaan orang. Analisis sentimen dapat dibagi menjadi tiga tugas: pengenalan teks informasi, ekstraksi informasi, dan klasifikasi emosi (pengenalan emosi, polaritas) [6].

Teknik analisis sentimen juga dapat digunakan untuk menganalisis pendapat orang dalam sebuah teks dan menentukan apakah sentimen tersebut positif, negatif, atau netral [7][8]. Analisis sentimen juga mencakup pemrosesan bahasa alami (NLP) untuk mendeteksi informasi subjektif dalam dokumen [9][10].
Penelitian ini menggunakan informasi dari YouTube yang awalnya diolah menggunakan *text mining*. Penambangan teks adalah siklus mengekstraksi informasi berguna dari kumpulan data tidak terstruktur. Metode ini dirancang untuk memecahkan masalah dalam proses ekstraksi informasi dan dapat digunakan untuk mencari informasi yang hilang atau tidak diketahui. Umumnya, penambangan teks melibatkan proses pembersihan data atau pra-pemrosesan data [11][12][13] Sedangkan analisis sentimen merupakan metode yang secara otomatis memahami, mengekstraksi, dan mengolah data sastra untuk memperoleh data emosional yang terkandung dalam teks opini. Metode ini berasal dari penambangan teks dan cocok untuk mengklasifikasikan ulasan ke dalam kelas positif dan negatif [14][15].

Pada tahap *preprocessing* data, dari dataset yang digunakan dalam penelitian ini ditemukan masalah yaitu kelas tidak seimbang yang sangat besar dimana data sentimen negatif jauh lebih banyak dibandingkan data sentimen positif. Oleh karena itu, diperlukan metode *preprocessing* untuk menyelesaikan masalah ketidakseimbangan kelas tersebut. Salah satu teknik *oversampling* yang bisa digunakan adalah teknik *Synthetic Minority Oversampling* (SMOTE). SMOTE dapat melakukan replikasi data sintetik, sehingga dapat memecah masalah distribusi data yang berbeda [16].

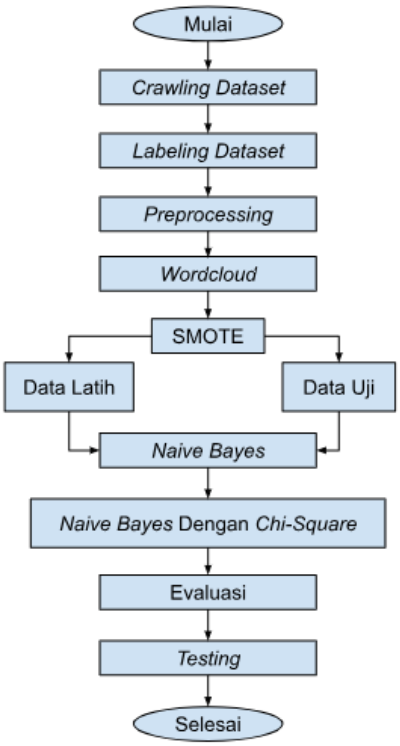
Naive Bayes Classifier (NBC) adalah salah satu algoritma klasifikasi di bidang pembelajaran mesin (ML) [17]. *Naive Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut tidak bergantung satu sama lain secara kondisional jika diberi nilai keluaran. Keuntungan menggunakan *Naive Bayes* adalah metode ini hanya memerlukan sedikit data latih untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi. Naive Bayes sering kali berkinerja jauh lebih baik dari yang diharapkan dalam situasi dunia nyata yang paling kompleks [18].
Penelitian ini menggunakan *Chi-Square* sebagai reduksi dimensi untuk mengatasi permasalahan *curse of dimensionality* [19][20][21]. Reduksi dimensi data dapat meningkatkan akurasi algoritma *Naive Bayes*. Salah satu metode reduksi dimensi adalah metode seleksi fitur. Seleksi fitur adalah proses pemilihan *subset* variabel input yang relevan dari dataset besar yang akan digunakan untuk membangun model [21]. Pemilihan fitur bertujuan untuk mengurangi fitur yang tidak relevan dan berlebihan, membutuhkan lebih sedikit waktu untuk melatih model, dan dapat membantu meningkatkan kinerja pengklasifikasi yang dihasilkan [22].

Metode seleksi fitur yang digunakan pada penelitian ini adalah *chi-square*. Penerapan model *chi-square* adalah untuk mengurangi tingkat kesalahan akibat terlalu banyak atribut dan mengurangi dimasukkannya atribut-atribut tersebut dalam analisis sentimen. Lebih lanjut, langkah reduksi merupakan langkah penting dalam pengenalan pola untuk mengklasifikasikan fitur diagnostik ke dalam atribut [23]. *Naive Bayes* dengan *chi-square* lebih akurat dibandingkan dengan *Naive Bayes* tanpa *chi-square* [24].
Dalam penelitian yang berjudul "Optimasi Akurasi Sentimen Komentar Xiaomi SU7 di Youtube Menggunakan *Naive Bayes* dan *Chi-Square*" akan dilakukan optimasi akurasi pada algoritma *Naive Bayes* menggunakan seleksi fitur *Chi-Square* berdasarkan komentar pada konten youtube mengenai produk Xiaomi SU7.

2. METODE PENELITIAN.

Adapun tahapan penelitian yang dilakukan pada penelitian ini adalah *crawling dataset* (pengambilan data), *labeling dataset* (melabel sentimen), *preprocessing* (meliputi *Cleaning, Case Folding, Tokenizing, Filtering & Stemming)*, *wordcloud* (visualisasi kata-kata yang paling sering muncul dalam dokumen dataset), SMOTE (*oversampling* data

latih), pemilihan proporsi data latih dan data uji, membuat model *Naïve Bayes*, membuat model *Naïve Bayes* dengan seleksi fitur *Chi-Square*, melakukan evaluasi pada model *Naïve Bayes* dan model *Naïve Bayes* dengan seleksi fitur *Chi-Square* serta membandingkan tingkat akurasinya (*include* data *precision, recall, f1-score, & support)*, dan melakukan *testing* dengan cara memprediksi hasil analisis sentimen untuk kata-kata yang baru diinput. Tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1. *Crawling Dataset.*

Perayapan web menggunakan *Uniform Resource Locators* (URL) untuk menemukan halaman web dan mengembalikan data yang relevan langsung ke pengguna. Pengguna tidak perlu menelusuri halaman web individual untuk mengakses informasi, sehingga menghemat waktu dan energi serta meningkatkan akurasi pengumpulan data. Data yang terkumpul dapat diolah lebih lanjut dan dianalisis sesuai kebutuhan pengguna [25].

2.2. *Labeling Dataset.*

Pada proses ini dilakukan pelabelan pada komentar-komentar youtube yang ada dalam file csv yang bertujuan untuk memberikan 3 kategori pada dataset tersebut (Positif, Negatif dan Netral). Proses *labeling* melibatkan analisis sentimen, yaitu menentukan apakah opini yang diungkapkan dalam sebuah kalimat atau dokumen bersifat positif, negatif, atau netral dengan mengklasifikasikan polaritas teks [26]. Label positif untuk komentar yang mendukung produk Xiaomi SU7, label negatif untuk komentar yang menentang produk Xiaomi SU7, dan label netral untuk komentar yang tidak berkaitan dengan produk Xiaomi SU7.

2.3. *Preprocessing.*

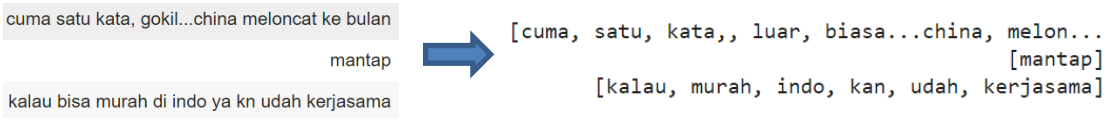
Tahap *preprocessing* berfungsi untuk menyeragamkan bentuk dan format teks sehingga dapat dipersiapkan untuk diolah menjadi data pada tahap berikutnya. *Cleaning*, *case folding*, *tokenizing*, *filtering*, dan *stemming* adalah semua bagian dari proses *preprocessing* teks [27].

1. *Cleaning* & *Case Folding* : *Cleaning* merupakan proses mengolah data yang tidak lengkap, penghapusan data duplikat, pemeriksaan data yang tidak konsisten dan perbaikan kesalahan data [28]. Pada tahap *case folding*, semua huruf akan diubah menjadi *lowercase*, atau huruf kecil [29].

Cuma satu kata, gokil...China meloncat ke bulan	cuma satu kata, gokil...china meloncat ke bulan
Mantap	mantap
Kalau bisa murah di indo ya kn udah kerjasama	kalau bisa murah di indo ya kn udah kerjasama

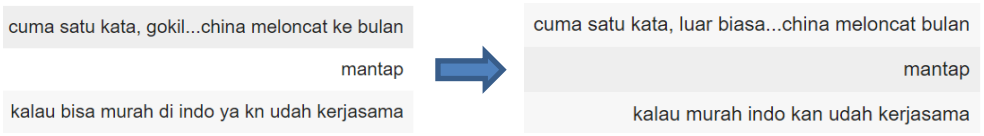
Gambar 2. Tahapan Cleaning & Case Folding

2. *Tokenizing* : Data yang dibersihkan kemudian dipecah menjadi unit-unit yang lebih kecil seperti kata dan frasa menggunakan *tokenizer*. Setiap token menerima indeks dan kata yang sesuai, membentuk kamus (kosa kata) [30]. Gambar 3. menjelaskan proses sebelum (kiri) dan proses sesudah (kanan) *Tokenizing*.



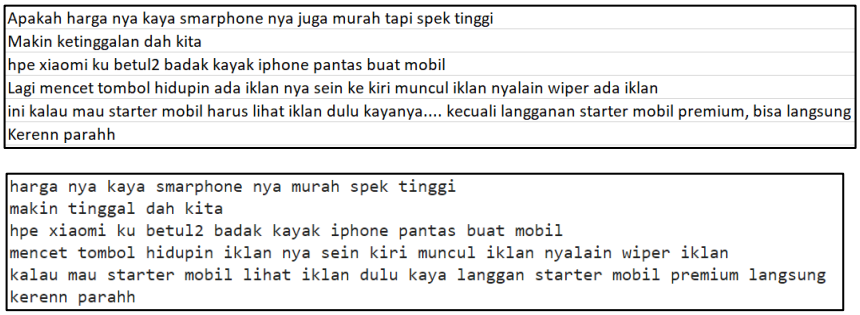
Gambar 3. Tahapan Tokenizing

3. *Filtering* : Menghapus kata-kata yang umum dan tidak berhubungan akan mengubah akhiran setiap kata yang difilter menjadi kata dasar yang berisi teks yang dihapus [27]. Proses ini disebut juga dengan *stopword removing*, yaitu menghapus kata-kata yang tidak bermakna dan tidak memiliki pengaruh terhadap analisis sentimen. Gambar 4. menjelaskan proses sebelum (kiri) dan proses sesudah (kanan) *Filtering*.



Gambar 4. Tahapan Tokenizing

4. *Stemming* : tahap dimana dilakukan pencarian kata dasar dari setiap hasil *filtering* kata [27]. Gambar 5. menjelaskan proses sebelum (atas) dan proses sesudah (bawah) *Stemming*.



Gambar 5. Tahapan Stemming

2.4. *Wordcloud.*

Word Cloud adalah metode penambangan teks yang menampilkan grafik frekuensi kata yang menyorot kata-kata yang lebih sering muncul dalam teks sumber. Semakin besar kata dalam gambar, semakin sering kata tersebut muncul dalam dokumen [31].

2.5. SMOTE.

Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE) adalah metode yang meningkatkan keakuratan algoritme yang digunakan dan memungkinkan algoritme menangani kelas yang tidak seimbang [32].

2.6. Proporsi data latih dan data uji.

Pemilihan persentase dalam membagi data dalam penelitian ini didasarkan pada prinsip Pareto atau dikenal dengan aturan 80:20. Ini pada dasarnya adalah teori yang menyatakan bahwa 80% keluaran atau hasil dihasilkan dari 20% koordinasi masukan dan keluaran yang tidak seimbang [33].

2.7. *Naive Bayes.*

Pengklasifikasi *Naive Bayes* adalah metode klasifikasi berdasarkan teorema *Bayes*. Metode klasifikasi ini menggunakan probabilitas dan statistik, atau memprediksi peluang masa depan berdasarkan pengalaman masa lalu, yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Oleh karena itu, teorema ini dikenal dengan teorema *Bayes*. Penerapan metode *Naïve Bayes* mencakup klasifikasi dokumen teks, metode pembelajaran mesin untuk menghasilkan probabilitas, diagnosis medis otomatis, serta deteksi dan pemfilteran spam [34].

2.8. *Naive Bayes* dengan *Chi-Square.*

Chi-square adalah metode yang sering digunakan di bidang statistika [35]. Uji *chi-square* merupakan jenis uji perbandingan nonparametrik yang paling umum dilakukan terhadap dua variabel, dimana variabel kedua mempunyai skala data nominal. Jika hanya ada satu variabel dengan skala nominal bivariat, maka dilakukan analisis *chi-square*, yang menunjukkan bahwa metode ini sebaiknya digunakan jika ordenya tidak sama [36].

2.9. Evaluasi

Tahap evaluasi merupakan tahap yang dilakukan untuk melihat seberapa baik kinerja algoritma klasifikasi yang digunakan dalam penelitian. Tolak ukur yang digunakan untuk mengukur kinerja adalah akurasi, presisi, *recall*, dan *F-measure* [37].

2.10. *Testing*

Tahapan testing merupakan percobaan yang dilakukan menggunakan data uji yang bertujuan untuk menentukan kata yang kita masukkan merupakan sentimen positif atau negatif.

3. HASIL DAN PEMBAHASAN

3.1. *Crawling Dataset*

Pada penelitian ini menggunakan dataset yang diambil dari video youtube berjudul “XIAOMI BIKIN MOBIL: XIAOMI SU7” kemudian dilakukan *crawling dataset* menggunakan *website* Netlytic dan menghasilkan file berformat CSV. File dataset dinamai “Dataset_xiaomi_su7”. Gambar 6. merupakan proses *Crawling Dataset* yang telah dilakukan.

Gambar 6. Crawling Dataset

3.2. *Labeling Dataset*

File “Dataset_xiaomi_su7” diberi label dengan cara manual, yaitu dengan menambahkan kolom “Sentimen” yang berisi Positif, Netral atau Negatif. Tabel 1. merupakan *sample* proses labeling yang telah dilakukan.

Tabel 1. Labeling Dataset

author	description	sentimen
@aweyjr4758	hahaa edasss	Positif
@unangsaputra1331	Jepang ketawa melihat teknologi cina.	Positif

@bayusubekti4058	Cuma satu kata, gokil...China meloncat ke bulan	Positif
@muhdazmirahim8478	Mantap	Positif
@sahabatkhazanah	Kalau bisa murah di indo ya kn udah kerjasama	Netral

3.3. *Preprocessing*

Preprocessing dilakukan dimulai dari *cleaning, case folding*, menghapus kolom yang tidak diperlukan, normalisasi teks, *tokenizing, filtering, stemming*, *cleaning stop-words*, *tokenizing*, *stemming*, penggabungan data, *filtering*, *label encoding*,

Tabel 2. Hasil Akhir Preprocessing

author	text_cleaning	sentimen
@aweyjr4758	hahaa hebat	Positif
@unangsaputra1331	jepang ketawa melihat teknologi cina	Positif
@bayusubekti4058	cuma satu kata, luar biasa...china meloncat bulan	Positif
@muhdazmirahim8478	mantap	Positif
@sahabatkhazanah	kalau murah indo kan udah kerjasama	Netral

3.4. *Wordcloud*

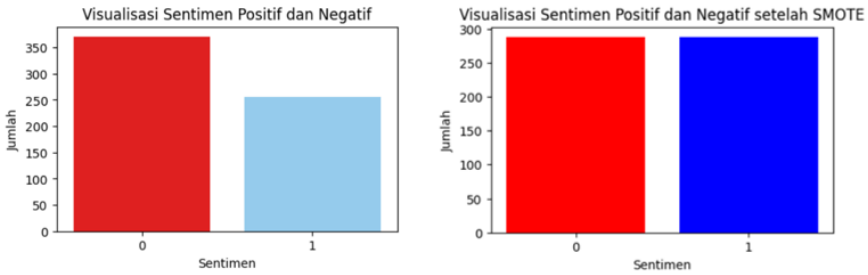
Wordcloud merupakan visualisasi kata-kata yang paling sering muncul dalam dokumen dataset yang memiliki sentimen negatif maupun sentimen positif dengan hasil seperti Gambar 7. :

Gambar 7. Visualisasi Sentimen Negatif dan Sentimen Positif

Gambar 7. Menjelaskan bahwa pada sentimen negatif terdapat kata-kata yang mendominasi seperti “china”, “harga”, “xiaomi”, “iklan”, “mahal”, “teknologi”, “spam”, “*bloatware*”, “Indonesia” dan lainnya. Sedangkan pada sentimen positif, kata-kata yang mendominasi adalah “china”, “keren”, “Indonesia”, “luar biasa”, “lebih murah”, “mobil listrik”, “xiaomi”, “mantap”, dan lainnya. Hal ini ditandai dengan ukuran kata yang semakin besar sehingga menunjukkan bahwa kata tersebut semakin mendominasi.

3.5. SMOTE

Pada penelitian ini teridentifikasi bahwa data sentimen negatif jauh lebih banyak dibandingkan dengan data sentimen positif. Maka perlu melakukan SMOTE supaya data pelatihan memiliki distribusi yang seimbang antara kelas-kelasnya, yang dapat meningkatkan kinerja model dalam memprediksi kelas minoritas. Warna merah merupakan sentimen negatif sedangkan warna biru merupakan sentimen positif pada Gambar 8.



Gambar 8. Hasil SMOTE Dalam Bentuk visualisasi Sentimen Negatif dan Positif

Gambar 8. Menjelaskan perbandingan antara sentimen yang tidak menggunakan SMOTE (kiri) dan sentimen yang menggunakan SMOTE (kanan) (data sentimen positif dan negatif dengan konsidi seimbang).

3.6. Pembagian Data Latih dan Data Uji

Penelitian ini membagi data dengan proporsi data latih sebesar 80% dan data uji sebesar 20%

3.7. Evaluasi

Evaluasi yang dilakukan pada penelitian ini menghasilkan akurasi, presisi, *f1-score*, serta *support* pada *naive bayes* dan *naive bayes* dengan menggunakan seleksi fitur *chi-square* yang dapat dilihat pada gambar di bawah. Percobaan menggunakan *Naive Bayes* menghasilkan akurasi sebesar 76.19% sedangkan *Naive Bayes* dengan *Chi-Square* menghasilkan akurasi sebesar 78,57%. Hal ini membuktikan bahwa adanya peningkatan akurasi sekitar 2,38%.

Tabel 3. Evaluasi Dari Model Naive Bayes dan Naive Bayes Menggunakan Chi-Square

	precision	***recall***	***f1-score***	***support***
negatif	0.90	0.72	0.80	83
positif	0.61	0.84	0.71	43
accuracy		0.7619047619047619		126

	precision	***recall***	***f1-score***	***support***
negatif	0.88	0.78	0.83	83
positif	0.65	0.79	0.72	43
accuracy		0.7857142857142857		126

3.8. Testing

```
Akurasi model naive bayes dengan seleksi fitur chi-square 0.7857142857142857

Masukkan teks baru: anjir
Hasil Analisis Sentimen untuk Teks Baru: Negatif
```

Gambar 9. Hasil Testing

Gambar 9. merupakan hasil pengujian prediksi analisis sentimen untuk teks baru dengan kata “anjir” menggunakan model *Naive Bayes* dengan seleksi fitur *Chi-Square* yang akurasinya 78,5% mendapatkan hasil sebagai sentimen negatif.

4. KESIMPULAN

Kesimpulan dari penelitian ini adalah optimasi akurasi *Naive Bayes* menggunakan *Chi-Square* berhasil dilakukan. Hal ini dibuktikan dengan akurasi *Naive bayes* sebelum menggunakan *Chi-Square* berada di angka 76.1% mengalami kenaikan setelah menggunakan *chi-square* menjadi 78.5%. Jadi dengan menggunakan tambahan seleksi fitur *Chi-Square* pada *Naive Bayes*, optimasi akurasinya meningkat sebesar 2,4%. Dalam penelitian ini, kelebihannya termasuk penggunaan metode *Naive Bayes* yang telah terbukti efektif dalam analisis sentimen. Selain itu, penggunaan *Chi-Square* sebagai metode seleksi fitur membantu dalam mengatasi masalah *curse of dimensionality* dan meningkatkan akurasi prediksi sentimen. Namun, kelemahan penelitian ini mungkin terletak

pada keterbatasan dalam generalisasi hasil karena fokus pada satu produk dan platform tertentu.

Disarankan untuk memperluas cakupan analisis sentimen ke platform media sosial lainnya selain YouTube, seperti Twitter, Instagram, atau Facebook, untuk mendapatkan pemahaman yang lebih komprehensif tentang persepsi konsumen terhadap merek. Selain itu, mempertimbangkan penggunaan algoritma *machine learning* yang lebih canggih dan teknik pemrosesan bahasa alami yang lebih kompleks dapat meningkatkan akurasi klasifikasi sentimen.

DAFTAR RUJUKAN

[1] M. Kim and T. H. Baek, "I'll follow the fun: The extended investment model of social media influencers," *Telematics and Informatics*, vol. 74, p. 101881, 2022, doi: https://doi.org/10.1016/j.tele.2022.101881.

[2] D. Vrontis, A. Makrides, M. Christofi, and A. Thrassou, "Social media influencer marketing: A systematic review, integrative framework and future research agenda," *Int J Consum Stud*, vol. 45, Jan. 2021, doi: 10.1111/ijcs.12647.

[3] S. V. Jin, A. Muqaddam, and E. Ryu, "Instafamous and social media influencer marketing," *Marketing Intelligence & Planning*, vol. 37, no. 5, pp. 567–579, Jan. 2019, doi: 10.1108/MIP-09-2018-0375.

[4] A. Fatihah and H. Munandar, "3 rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) 30 Agustus 2023-Jakarta," 2023.

[5] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics (Switzerland)*, vol. 9, no. 3, Mar. 2020, doi: 10.3390/electronics9030483.

[6] A. Mutoi Siregar, T. Astiyah Hasan, U. Buana Perjuangan karawang Jl HSRonggo Waluyo, and J. Barat, "Techno Xplore Jurnal Ilmu Komputer dan Teknologi Informasi."

[7] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," *Procedia Comput Sci*, vol. 197, pp. 143–150, 2022, doi: https://doi.org/10.1016/j.procs.2021.12.128.

[8] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp. 46–50, Oct. 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[9] G.-D. Pilar, S.-B. Isabel, P.-M. Diego, and G.-Á. José Luis, "A novel flexible feature extraction algorithm for Spanish tweet sentiment analysis based on the context of words," *Expert Syst Appl*, vol. 212, p. 118817, 2023, doi: https://doi.org/10.1016/j.eswa.2022.118817.

[10] J. T. Pintas, L. A. F. Fernandes, and A. C. B. Garcia, "Feature selection methods for text classification: a systematic literature review," *Artif Intell Rev*, vol. 54, no. 8, pp. 6149–6200, 2021, doi: 10.1007/s10462-021-09970-6.

[11] W. A. Prabowo and C. Wiguna, "Sistem Informasi UMKM Bengkel Berbasis Web Menggunakan Metode SCRUM," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, p. 149, Jan. 2021, doi: 10.30865/mib.v5i1.2604.

[12] S. Muhammad Habib, E. Haerani, S. Kurnia Gusti, S. Ramadhani, and T. H. Informatika UIN Sultan Syarif Kasim Riau Jl Soebrantas, "Klasifikasi Berita Menggunakan Metode Naïve Bayes Classifier," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 5, no. 2, 2022.

[13] R. Rasenda, H. Lubis, and R. Ridwan, "Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 369, Apr. 2020, doi: 10.30865/mib.v4i2.2051.

[14] J. Homepage, A. Harun, and D. P. Ananda, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Analysis of Public Opinion Sentiment About Covid-19 Vaccination in Indonesia Using Naïve Bayes and Decission Tree Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes dan Decission Tree," vol. 1, pp. 58–63, 2021.

[15] A. Dewandaru, J. Sasongko Wibowo, and A. History, "Jurnal Teknologi dan Manajemen Informatika Analisis Sentimen dan Klasifikasi Tweet Terkait Mutasi COVID-19 Menggunakan Metode Naïve Bayes Classifier Article Info ABSTRACT," vol. 8, pp. 32–38, 2022, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi

[16] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[17] A. Rozaq, Y. Yunitasari, K. Sussolaikah, E. R. N. Sari, and R. I. Syahputra, "Analisis Sentimen Terhadap Implementasi Program Merdeka Belajar Kampus Merdeka Menggunakan Naïve Bayes, K-Nearest Neighboars Dan Decision Tree," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 2, p. 746, Apr. 2022, doi: 10.30865/mib.v6i2.3554.

[18] I. B. Naive Bayes Untuk Menentukan Wadah Limbah and S. Karakteristik Gigih Putra Kawani, "Journal of Informatics, Information System, Software Engineering and Applications," vol. 1, no. 2, pp. 73–081, 2019, doi: 10.20895/INISTA.V1I2.

[19] G. Chao, Y. Luo, and W. Ding, "Recent Advances in Supervised Dimension Reduction: A Survey," *Mach Learn Knowl Extr*, vol. 1, no. 1, pp. 341–358, Dec. 2019, doi: 10.3390/make1010020.

[20] O. Saini, "A Review on Dimension Reduction Techniques in Data Mining," 2018. [Online]. Available: www.iiste.org

[21] R. Aziz, C. K. Verma, and N. Srivastava, "Dimension reduction methods for microarray data: a review," *AIMS Bioeng*, vol. 4, no. 1, pp. 179–197, 2017, doi: 10.3934/bioeng.2017.1.179.

[22] G. Kicska and A. Kiss, "Comparing swarm intelligence algorithms for dimension reduction in machine learning," *Big Data and Cognitive Computing*, vol. 5, no. 3, Sep. 2021, doi: 10.3390/bdcc5030036.

[23] R. Rosdiana, V. Novalia, I. Saputra, M. Ula, and M. Danil, "Application of Artificial Intelligence Chi-Square Model and Classification Of KNN in Heart Disease Detection," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 6, no. 1, pp. 180–188, Jul. 2022, doi: 10.31289/jite.v6i1.7343.

[24] R. Febriasto, N. L. P. S. P. Paramita, and W. Wibawati, "Prediksi Kuat Tekan Semen untuk Produk Portland Composite Cement (PCC) di PT. Semen Indonesia (Persero) Tbk. Menggunakan Support Vector Regression (SVR) Dengan Feature Selection," *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, Feb. 2020, doi: 10.12962/j23373520.v8i2.43071.

[25] D. Peng, T. Li, Y. Wang, and C. L. P. Chen, "Research on Information Collection Method of Shipping Job Hunting Based on Web Crawler," in *2018 Eighth International Conference on Information Science and Technology (ICIST)*, 2018, pp. 57–62. doi: 10.1109/ICIST.2018.8426183.

[26] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[27] N. Khasanah and A. Salim, "Rachman Komarudin 4) , Yana Iqbal Maulana 5) 1) Teknik Informatika, Fakultas Teknologi Informasi, Universitas Nusa Mandiri 2,3) Sistem Informasi, Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika 4) Sistem Informasi, Fakultas Teknologi Informasi, Universitas Nusa Mandiri 5) Teknik Informatika," 2022.
[28] E. Rahm and H. Do, "Data Cleaning: Problems and Current Approaches," *IEEE Data Eng. Bull.*, vol. 23, pp. 3–13, Jan. 2000.
[29] E. Y. Hidayat, R. W. Hardiansyah, and A. Affandy, "Analisis Sentimen Twitter untuk Menilai Opini Terhadap Perusahaan Publik Menggunakan Algoritma Deep Neural Network," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 2, pp. 108–118, Sep. 2021, doi: 10.25077/teknosi.v7i2.2021.108-118.
[30] P. B. Wintoro, H. Hermawan, M. A. Muda, and Y. Mulyani, "Implementasi Long Short-Term Memory pada Chatbot Informasi Akademik Teknik Informatika Unila," *EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi*, vol. 12, no. 1, p. 68, Jun. 2022, doi: 10.36448/expert.v12i1.2593.
[31] A. Alamsyah and F. Nuruz Zuhri, "Measuring Public Sentiment Towards Services Level in Online Forum using Naive Bayes Classifier and Word Cloud."
[32] L. Amatullah, Y. Widiastiwi, and N. Chamidah, "Penerapan Klasifikasi Random Forest Terhadap Data Gangguan Spektrum Autisme (ASD) Pada Anak-Anak Menggunakan Seleksi Fitur Principal Component Analysis".
[33] I. Kurniawan *et al.*, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 1, 2023, [Online]. Available: http://jurnal.mdp.ac.id
[34] A. Felicia Watratan, A. B. Puspita, D. Moeis, S. Informasi, and S. Profesional Makassar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," 2020. [Online]. Available: http://journal.isas.or.id/index.php/JACOST
[35] J. Reynaldo, P. P. Adikara, and R. C. Wihandika, "Analisis Sentimen Mengenai Produk Toyota Avanza Menggunakan Metode Learning Vector Quantization Versi 3 (LVQ 3) dengan Seleksi Fitur Chi Square, Lexicon-Based Features serta Normalisasi Min-Max," 2020. [Online]. Available: http://j-ptiik.ub.ac.id
[36] H. Mardiansyah, R. Widia Sembiring, and S. Efendi, "Handling Problems of Credit Data for Imbalanced Classes using SMOTEXGBoost," *J Phys Conf Ser*, vol. 1830, no. 1, p. 012011, Apr. 2021, doi: 10.1088/1742-6596/1830/1/012011.
[37] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 2, pp. 785–795, Jun. 2022, doi: 10.35957/jatisi.v9i2.1835.