Terakreditasi SINTA Peringkat 3
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi Nomor 225/E/KPT/2022
masa berlaku mulai Vol.7 No. 1 tahun 2022 s.d Vol. 11 No. 2 tahun 2026

Terbit online pada laman web jurnal:
http://publishing-widyagama.ac.id/ejournal-v2/index.php/jointecs

JOINTECS

JOINTECS
(Journal of Information Technology and Computer Science)

Vol. 9 No. 1 (2024) 21 - 30 e-ISSN:2541-6448 p-ISSN:2541-3619

Perbandingan Performa SVM dan *Naïve Bayes* Pada Analisis Sentimen Aplikasi Game Online

Galang Paksi Permana[1], Danang Aditya Nugraha[2], Heri Santoso[3]
Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas PGRI Kanjuruhan
[1]190403010004@mhs.unikama.ac.id, [2]danang.adty@unikama.ac.id, [3]heri@unikama.ac.id

Abstract

Nowadays, online games developed rapidly, especially Clash of Clans. This game, which is already more than 10th years old, still can compete with others. The reason is the developer who always develops and renews it. However, the renewal and development cannot be accepted by all gamers of Clash of Clans. Many gamers give reviews in Google Play Store. It found the desynchronization between the rating and reviews that are given. Therefore, sentiment analysis is needed to know the type of sentiment given. This research optimized the extraction of TF-IDF (Term Frequency - Inverse Document Frequency) and N-Gram feature, then the selection of Chi-Square and SelectKBest feature. On sentiment analysis, it used the algorithm of Support Vector Machine and Naïve Bayes on purpose to apply and find the best classification model to get a more accurate result. The results show that the classification model with TF-IDF and N-Gram feature extraction, as well as Chi-Square and SelectKBest feature selection, is the most optimal classification model with the highest accuracy results of 93% on the Support Vector Machine at percentage ratio of 70:30. In comparison, the highest accuracy obtained by Naïve Bayes is 91.6% with the same percentage ratio.

Keywords: sentiment analysis; feature optimization; support vector machine; *naïve bayes*.

Abstrak

Game online saat ini berkembang sangat pesat terutama *game online* berbasis *mobile*, salah satunya adalah *Clash of Clans*. *Game* yang sudah berusia lebih dari 10 tahun ini masih mampu bersaing dengan *game mobile* terbaru lainnya. Hal tersebut tidak lepas dari pengembangan dan pembaruan yang dilakukan oleh *developer*. Namun tentu pengembangan dan pembaruan tersebut tidak bisa sepenuhnya diterima oleh seluruh pemain *game Clash of Clans*. Banyak ulasan yang diberikan oleh pengguna di *Google Play Store*. Ditemukan ketidak sinkronan antara rating yang diberikan dengan ulasan yang diberikan, oleh karena itu dibutuhkan analisis sentimen untuk mengetahui jenis sentimen yang diberikan. Pada penelitian ini dilakukan penerapan optimasi ekstraksi fitur TF-IDF (*Term Frequency - Inverse Document Frequency*) dan *N-Gram*, serta seleksi fitur *Chi-Square* dan *SelectKBest* pada analisis sentimen pengguna aplikasi *game online Clash of Clans* menggunakan algoritma *Support Vector Machine* dan *Naïve Bayes* dengan tujuan menerapkan dan mencari model klasifikasi terbaik untuk mendapatkan hasil akurasi yang lebih akurat. Hasil pengujian menunjukkan bahwa model klasifikasi dengan ekstraksi fitur TF-IDF (*Term Frequency - Inverse Document Frequency*) dan *N-Gram*, serta seleksi fitur *Chi-Square* dan *SelectKBest* adalah model klasifikasi paling optimal dengan hasil akurasi tertinggi 93% pada *Support Vector Machine* pada presentase rasio 70:30, sementara akurasi tertinggi yang diperoleh *Naïve Bayes* sebesar 91,6% dengan presentase rasio yang sama.

Kata kunci: analisis sentimen; optimasi fitur; *support vector machine*; *naïve bayes*.

Diterima Redaksi : 29-07-2023 | Selesai Revisi : 28-05-2024 | Diterbitkan Online : 04-06-2024

1. Pendahuluan

Game merupakan salah satu sarana hiburan yang dapat digunakan untuk menghilangkan kepenatan setelah beraktivitas [1]. Persaingan *game online* saat ini sangat ketat, terlebih *game online mobile*. Anak muda zaman sekarang sebagian besar lebih memilih memainkan game di *smartphone* mereka ketimbang di *PC.* Menurut survei yang dilakukan terhadap pemain *game online* di Indonesia pada tahun 2021, hasil survey tersebut mengungkapkan bahwa jumlah pemain *game mobile* berada di urutan pertama dengan jumlah pemain 121,7 juta, kemudian disusul oleh pemain *game PC* sebanyak 53,4 juta pemain [2]. Hal tersebut terjadi karena kondisi ekonomi yang ada khususnya di Indonesia yang menjadikan *game mobile* lebih banyak penggunanya ketimbang *game PC*. Faktor lainnya adalah *game* itu sendiri yang menggunakan *mode multiplayer* dalam gamenya sehingga membuat pengguna cenderung lebih memilih bermain bersama teman atau *player* lain dalam memainkan *game* tersebut [3].

Ada berbagai macam game dengan jenis yang berbeda-beda seperti *simulation, racing, battle royal, strategy* dan jenis game lainya [4]. Saat ini, orang biasanya memeriksa ulasan dan peringkat game sebelum bermain di perangkat mereka [5]. *Clash of Clans* berada pada urutan keempat *game mobile* paling digemari di Indonesia. Game *multiplayer* yang bergenre strategi ini rilis pada 2 Agustus 2012 di *platform iOS* dan 7 Agustus 2012 di *platform Android*. *Game* ini memiliki latar belakang fantasi di mana pemain berperan sebagai kepala desa. *Clash of Clans* mengajak pemain untuk membangun desa mereka sendiri dengan menggunakan sumber daya yang diperoleh melalui serangan ke desa pemain lain atau melalui produksi di desa mereka sendiri. Banyak perubahan yang terjadi dalam *game online Clash of Clans* yang membuatnya masih bisa terus eksis dan bersaing dengan *game mobile* lainnya. *Google Play Store* memiliki beragam ulasan dari pengguna yang telah memainkan *Clash of Clans*. Ulasan pengguna mengandung informasi penting dan berguna dalam perbaikan dan pengembangan game [6]. Dari ulasan yang ada, ditemukan ketidak sinkronan antara ulasan dan rating yang diberikan. Maka dari itu perlu dilakukan analisis sentimen ulasan pengguna *game online Clash of Clans* untuk dapat mengetahui jenis sentimen yang diberikan oleh pengguna.

Analisis sentimen ialah bidang keilmuan yang mempelajari opini, evaluasi, sentimen, sikap, penilaian, dan emosi seseorang terhadap suatu objek seperti barang, orang, peristiwa, organisasi, dan masalah konkrit. Tujuan utama dari analisis sentimen adalah untuk memahami dan menganalisis opini dan sentimen yang diungkapkan oleh orang-orang dalam teks atau media sosial. Analisis sentimen banyak digunakan dalam berbagai aplikasi seperti pemantauan merek, penelitian pasar, politik, dan manajemen reputasi online [7]. Analisis sentimen dapat menggunakan metode data *mining*. Data mining merupakan salah satu teknik untuk membuat data yang berukuran besar menjadi informasi yang sangat penting. Data mining bertujuan menghasilkan berbagai pola yang sebelumnya tidak diketahui. Banyak persoalan di bidang ekonomi, bisnis, pertanian, kehutanan, pemerintahan, ekologi, dan lain sebagainya dapat diselesaikan dengan menggunakan data mining [8].

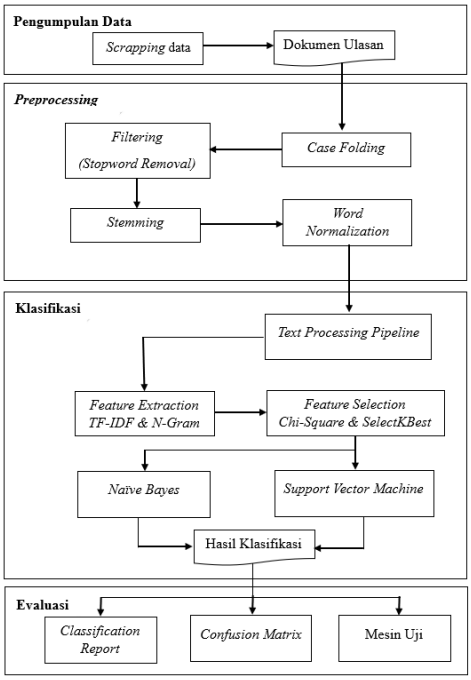
Berdasarkan penelitian analisis sentimen yang telah dilakukan memiliki perbedaan mengenai metode atau algoritma serta penambahan fitur pembobotan kata yang digunakan. Mardiana [9] melakukan perbandingan antara lima algoritma yaitu *Neural Network, K-Nearest Neighbor, Naïve Bayes, Support Vector Machine*, dan *Decision Tree* untuk mengekstraksi atribut dari dokumen atau teks berisi komentar. Berdasarkan hasil akurasi dari setiap metode, diketahui bahwa *Neural Network, Support Vector Machine* dan *Naïve Bayes* memberikan hasil klasifikasi yang baik dan dapat dikategorikan sebagai *Good Classification*. Pada penelitian Arifin [10] dilakukan *text classification* menggunakan algoritma *Support Vector Machine* dengan ekstraksi fitur *N-Gram*. Hasil penelitian menunjukkan bahwa penggunaan ekstraksi fitur *N-Gram* mampu meningkatkan akurasi *Support Vector Machine* sebesar 30% menjadi 60%. Dalam penelitian Zuliana [11] program migrasi TV digital dianalisis sentimennya menggunakan algoritma *Naïve Bayes* dengan seleksi fitur *Chi-Square* yang berfungsi untuk menghapus atribut atau fitur yang kurang relevan dan mempercepat pemrosesan. Hasil pengujian pada penelitian ini menunjukkan bahwa algoritma *Naïve Bayes* mendapatkan nilai akurasi sebesar 96%, *precision* 93%, dan *recall* 100%. Sementara itu penelitian Fitri [12] melakukan analisis klasifikasi penyederhanaan kriteria dalam pemilihan pakan ternak ikan menggunakan algortima C4.5 dengan seleksi fitur SelectKBest dan tanpa seleksi fitur. Hasilnya menunjukkan bahwa metode C4.5 dengan menerapkan seleksi fitur SelectKBest menghasilkan akurasi yang lebih baik dari pada metode C4.5 tanpa menggunakan seleksi fitur dengan hasil akurasi 92%, sementara tanpa seleksi fitur dengan hasil akurasi 86,8%.

Tujuan penelitian ini adalah untuk mengetahui performa model klasifikasi antara *Support Vector Machine* dan *Naïve Bayes* dalam konteks analisis sentimen. SVM mempunyai kemampuan dalam menggeneralisasi ke tingkat akurasi yang cukup tinggi dalam mengklasifikasikan sebuah pola. Sedangkan *Naïve Bayes* memiliki keunggulan pemrosesan cepat pada dataset yang besar dan hasil akurasi tinggi pada data *training* kecil. Sentimen analisis adalah teknik yang digunakan untuk mengidentifikasi, memahami, dan mengklasifikasikan sentimen atau opini dalam teks. Penelitian ini mengimplementasikan ekstraksi fitur TF-IDF dan *N-Gram*, serta seleksi fitur *Chi-Square* dan *SelectKBest* untuk meningkatkan performa akurasi dari SVM dan *Naïve Bayes* dalam pengklasifikasian teks

berupa sentimen pada aplikasi game online yang didapat dari *Google Play Store*. Hasil dari penelitian ini dapat membandingkan hasil analisis sentimen antara *Support Vektor Machine* dan *Naïve Bayes* berdasarkan dengan tingkat akurasi klasifikasi yang dihasilkan.

2. Metode Penelitian

Penelitian ini menggunakan data hasil dari *scraping* ulasan pengguna aplikasi *game online Clash of Clans* pada *Google Play Store* dengan bantuan *Google Colab* menggunakan bahasa pemrograman *python*. *Scraping* adalah teknik untuk mendapatkan informasi secara otomatis tanpa harus menyalinnya secara manual [13]. Ulasan yang digunakan terbatas hanya ulasan yang berasal dari negara Indonesia dan berjumlah 1000 ulasan. Berikut adalah diagram alur dari metode penelitian seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

Gambar 1 diawali dengan pengambilan data dari *Google Play Store* (*scraping*), data yang didapat dilabeli pada setiap ulasan yang didapat berdasarkan polaritas dan subjektifitas, kemudian akan diproses pada tahap *preprocessing*. *Preprocessing* teks adalah proses mengubah format data tidak terstruktur ke format terstruktur sesuai kebutuhan untuk mengekstrak informasi dalam proses penambangan [14]. Kata dari seluruh dokumen akan diberi bobot (*term weighting*) menggunakan ekstraksi fitur TF-IDF dan *N-Gram* serta seleksi fitur *Chi-Square* dan *SelectKBest* yang bertujuan untuk memudahkan dalam evaluasi model dan mengetahui model paling optimal. Pembobotan kata adalah proses perhitungan bobot tiap kata yang akan dicari pada keseluruhan dokumen untuk mengetahui ketersediaan dan kemiripan kata pada dokumen [15]. Dataset yang sudah bersih kemudian akan dibagi menjadi dua bagian yaitu data latih dan data uji. Metode yang digunakan dalam penelitian ini adalah *Support Vector Machine* dan *Naive Bayes* sebagai perbandingan untuk mendapatkan hasil akurasi terbaik dalam pengklasifikasian data teks sentimen ulasan.

2.1. Pengumpulan Data

Dataset yang peneliti gunakan pada penelitian ini merupakan dataset ulasan dari pengguna aplikasi *Clash of Clans* dari hasil *scraping* pada *Google Play Store*. Data ini memiliki 10 atribut dan 1000 data ulasan. Langkah pertama yang harus dilakukan adalah menginstall *google play scraper,* kemudian *install library pandas* dan *numpy.* Arahkan *google play scraper* ke "*com.supercell.clashofclans*" untuk melakukan scrapping data ulasan pada aplikasi *Clash of Clans*. Data yang diambil berjumlah 1000 data ulasan *newest* atau terbaru dengan bahasa dan negara Indonesia pada 5 Januari 2023 sampai 8 Januari 2023.

2.2. *Preprocessing*

Data yang didapat dari hasil *scraping* ulasan pengguna aplikasi *game online Clash of Clans* inilah yang akan dilakukan tahapan *preprocessing*. Tahap *Preprocessing* yang terdapat dalam penelitian ini adalah *case folding*, *filtering* (*stopword removal*), *stemming* dan *word normalization*. Semua fungsi yang telah dibuat pada tahapan *preprocessing* kemudian dijalankan pada satu fungsi yaitu *Text Processing Pipeline*.

2.3. Ekstraksi dan Seleksi Fitur

Setelah proses *preprocessing* selesai, dilakukan ekstraksi fitur dan seleksi fitur. Ekstraksi fitur menjadi bagian yang sangat penting dalam pengolahan dokumen pada mesin pencari karena sangat menentukan keberhasilan proses *text mining.* Jika nilai fitur yang dihasilkan tidak tepat, maka informasi yang digali dalam *text mining* tidak bisa memenuhi kriteria yang diinginkan. Akibatnya, informasi yang ditampilkan oleh mesin pencari tidak akan memenuhi keinginan pengguna [16]. Metode ekstraksi fitur yang digunakan untuk mendapatkan fitur terpilih pada dataset adalah TF-IDF dan *N-Gram*. Sedangkan Seleksi Fitur adalah proses dimana seleksi dilakukan secara otomatis atau manual memilih fitur-fitur yang berkontribusi paling besar pada variabel prediksi atau output. Memilih fitur yang tidak relevan dalam dataset dapat mengurangi keakuratan model dan membuat model belajar berdasarkan fitur yang tidak relevan. Manfaat melakukan seleksi fitur sebelum memodelkan data yaitu mengurangi *overfitting*, meningkatkan akurasi dan mengurangi waktu pelatihan [17]. *Chi-Square* dan *SelectKBest* adalah metode seleksi fitur yang akan diterapkan pada penelitian ini. Dilakukan optimasi pada ekstraksi fitur dan seleksi fitur dengan cara melakukan 6 kali pengujian kombinasi fitur dengan tujuan untuk mendapatkan model klasifikasi yang paling optimal dan menghasilkan hasil akurasi yang lebih tinggi dan akurat.

2.4. Klasifikasi *Support Vector Machine*

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya kedalam kelas tertentu dari jumlah kelas yang tersedia [18]. Metode SVM adalah metode terbaik yang digunakan dalam melakukan prediksi berdasarkan kerangka kerja statistik yang dikembangkan oleh seorang peneliti yang bernama Chervonenkis dan Vapnik. SVM di dalam *machine larbibf* merupakan salah satu algoritme pembelajaran yang terawasi sehingga dapat digunakan dalam melakukan analisis untuk kebutuhan klasifikasi dan regresi [19]. Konsep SVM bermula dari masalah klasifikasi dua kelas sehingga membutuhkan *training set* positif dan negatif. SVM berusaha menemukan *hyperplane* (pemisah) terbaik untuk memisahkan ke dalam dua kelas dan memaksimalkan margin antara dua kelas tersebut. Klasifikasi SVM dinotasikan sebagai rumus 1.

$$f(x) = W^T X + b \quad (1)$$

Dalam klasifikasi SVM**,** fungsi klasifikasi dilambangkan dengan $f(x)$. Fungsi $f(x)$ memprediksi kelas target *y* berdasarkan fitur input *x*. *W* adalah vektor bobot, *X* adalah vektor fitur masukan**,** dan *b* adalah bias. Sehingga diperoleh rumus 2 dan rumus 3.

$$[(W^T.x_i) + b] \geq 1 \; untuk \; y_i = +1 \quad (2)$$

$$[(W^T.x_i) + b] \leq 1 \; untuk \; y_i = -1 \quad (3)$$

Klasifikasi dilakukan dengan dataset latih yang dinotasikan dengan $f(x)$ = himpunan data *training*, ke - i = 1, 2, ..., n dan, yi = label kelas dari x_i. Pengelompokkan menggunakan *hyperplane linier* seperti terlihat pada rumus. Parameter w menotasikan *vector* terhadap *hyperplane*, b adalah *offset.*

2.5. Klasifikasi *Naïve Bayes*

Algoritma *Naïve Bayes* merupakan suatu bentuk klasifikasi data dengan menggunakan metode probabilitas dan statistik. Metode ini pertama kali dikenalkan oleh ilmuwan Inggris Thomas Bayes, yaitu digunakan untuk memprediksi peluang yang terjadi di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *teorema Bayes* [20]. Model *naïve bayes* yang digunakan pada penelitian ini adalah *gaussian naïve bayes* yang merupakan sebuah teknik klasifikasi yang digunakan dalam *machine learning* dengan menggunakan metode *probability* dan *distribusi gaussian*. Model ini mengasumsikan bahwa setiap *feature* pada data memiliki pengaruh yang independen dalam memprediksi target. Kombinasi prediksi dari seluruh parameter adalah prediksi akhir dengan *probability* dari target *variable* yang diklasifikasikan ke dalam dua kelas. Klasifikasi akhirnya adalah hasil *probability* yang lebih tinggi dari grup target.

$$P(X|Y) = \frac{P(Y|X).P(X)}{P(Y)} \quad (4)$$

Pada rumus 4 digunakan untuk mendapatkan nilai token dari setiap kata menggunakan model *naïve bayes*. Dimana *X,Y* adalah kejadian, *P*(*X*|*Y*) merupakan probability untuk *X* ketika *Y* benar. *P*(*Y*|*X*) adalah probability untuk *Y* ketika *X* benar. *P*(*X*)*, P*(*Y*) adalah probability independent untuk *X* dan *Y*.

2.6. Evaluasi

Tahapan evaluasi dilakukan diakhir proses penelitian. Pada dataset ulasan pengguna aplikasi *game online Clash of Clans* di *Google Play Store* yang telah melalui tahap ekstraksi fitur dan seleksi fitur. Uji akurasi dilakukan untuk menentukan model klasifikasi yang paling optimal, dengan menggunakan 6 kombinasi fitur dan tiga presentase rasio pengujian, yaitu 70:30, 80:20, dan 90:10 untuk data *training* dan *data testing*, dengan total hasil pengujian akurasi sebanyak 36 hasil akurasi. Tahapan evaluasi pada penelitian ini dilakukan dengan metode *Confusion Matrix, Classification Report* dan mesin uji. *Confusion Matrix* dan *Classification Report* merupakan metode yang digunakan dalam melakukan perhitungan akurasi dan evalusi model klasifikasi. Sementara mesin uji berfungsi untuk melakukan pengujian model yang berjalan pada algoritma dengan menghasilkan label positif dan negatif, yaitu 0 dan 1.

3. Hasil dan Pembahasan

Hasil dan pembahasan merupakan penjelasan hasil dari penelitian yang telah dilakukan mengenai perbandingan performa SVM dengan *Naïve Bayes* dan analisis sentimen terhadap data *Clash of Clans*. Pada tahap ini, hasil akan dijelaskan dari Pengambilan Data, *Preprocessing* Data, Ekstraksi dan Seleksi Fitur, *Machine Learning Model*, Hasil dari Klasifikasi Sentimen Analisis menggunakan SVM dan *Naïve Bayes* dan Evaluasi terhadap hasil yang telah diperoleh dalam penelitian. Berikut penjelasan mengenai hasil dan pembahasan dari setiap langkah penelitian.

3.1 Pengambilan Data

Dataset yang peneliti gunakan pada penelitian ini merupakan dataset ulasan dari pengguna aplikasi Clash of Clans dari hasil scrapping pada Google Play Store. Data ini memiliki 10 atribut dan 1000 data ulasan. Langkah pertama yang harus dilakukan adalah menginstall google play scraper, kemudian install library pandas dan numpy. Arahkan google play scraper ke “com.supercell.clashofclans” untuk melakukan scrapping data ulasan pada aplikasi Clash of Clans. Data yang diambil berjumlah 1000 data ulasan newest atau terbaru dengan bahasa dan negara Indonesia pada 5 Januari 2023 sampai 8 Januari 2023. Tabel 1 menunjukan dataset ulasan yang berisi *score*, at dan content.

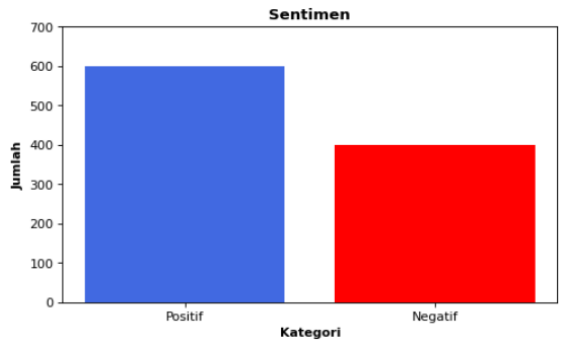
Tabel 1. Dataset Ulasan

No	score	at	content
1	5	1/8/2023 8:17	Ppp
2	5	1/8/2023 8:13	Saya suka gem ini
3	4	1/8/2023 8:12	Yang lain bisa download aku ngak bisa dwonlad ðŸ˜¢ðŸ˜¢ðŸ˜¢ gam ini selalu menemaniku dari umur ku 6 tahun ke 16 tahun lalu aku ngak bisa dwonload gam ini kembali semoga supercel bisa membantu ya plis ðŸ˜¯ðŸ˜¯ðŸ˜¯
…	…	…	…
1000	5	1/5/2023 7:23	Bagus

Data hasil scrapping tersebut kemudian diberi label "Sentiment" secara manual, label tersebut merupakan label positif dan label negatif. Pelabelan dibantu oleh seorang guru bahasa Indonesia dari salah satu sekolah swasta di Kota Malang. Data label Sentiment yang telah dibuat dilakukan proses *replace value*, yaitu Positif menjadi 0 dan Negatif menjadi 1 yang ditunjukan Tabel 2. Dari 1000 data ulasan yang didapatkan, dibuat sebuah model untuk mengetahui berapa banyak ulasan yang bernilai positif dan negative seperti Gambar 2.

Tabel 2. Dataset Hasil Pelabelan Manual dan *Replace Value*

No	score	at	content	Sentiment
1	5	1/8/2023 8:17	Ppp	0
2	5	1/8/2023 8:13	Saya suka gem ini	0
3	4	1/8/2023 8:12	Yang lain bisa download aku ngak bisa dwonlad ðŸ˜¢ðŸ˜¢ðŸ˜¢ ðŸ˜¢ gam ini selalu menemaniku dari umur ku 6 tahun ke 16 tahun lalu aku ngak bisa dwonload gam ini kembali semoga supercel bisa membantu ya plis ðŸ˜¯ðŸ˜¯ðŸ˜¯	1
…	…	…	…	…
1000	5	1/5/2023 7:23	Bagus	0



Gambar 2. Jumlah Sentimen Positif dan Negatif

3.2. *Preprocessing* Data

*T*ahap *preprocessing data* adalah tahap dimana data yang diambil dari google play store menggunakan teknik scrapping akan dibersihkan sebelum dianalisa lebih lanjut. Data mentah tersebut akan melalui beberapa proses, yang pertama adalah *Case Folding* yaitu mengubah seluruh kata menjadi lower case, menghapus *URL*, angka dan karakter tanda baca termasuk emoji. *Filtering (Stopword Removal)* yaitu menghapus kata atau karakter tanda baca dengan informasi rendah atau ganda. *Stemming* yaitu menghapus imbuhan. Proses terakhir adalah *Word Normalization* yaitu mengubah kata slang menjadi kata baku atau kata dasar. Tabel 3 merupakan hasil *preprocessing* data pada salah satu ulasan pengguna yaitu *raw* data pada baris ke 771. Keempat proses tersebut dilakukan untuk mendapatkan data yang bersih sehingga memudahkan proses pembobotan kata dan evaluasi model.

Tabel 3. Hasil *Preprocessing Data*

Metode	Data
Raw Data	Masuk tahun 2023 game nya malah sering putus sambungan. Ga keren min.
Case Folding	masuk tahun game nya malah sering putus sambungan ga keren min
Filtering (Stopword Removal)	masuk tahun game nya malah sering putus sambungan ga keren min
Stemming	masuk tahun game nya malah sering putus sambung ga keren min
Word Normalization	masuk tahun game nya malah sering putus sambung tidak keren min

3.3. Ekstraksi Fitur dan Seleksi Fitur

Ekstraksi fitur adalah tahapan di mana fitur-fitur baru diekstraksi dari fitur-fitur asli melalui suatu pemetaan fungsional, dan nantinya nilai-nilai yang dihasilkan akan dianalisis untuk tahapan berikutnya. Seleksi fitur digunakan untuk mengurangi atribut atau fitur yang tidak relevan adalam dataset dan mempercepat pemrosesan data. Dalam penelitian ini, dilakukan ekstraksi fitur menggunakan metode TF-IDF dan *N-Gram*, serta seleksi fitur menggunakan metode *Chi-Square* dan *SelectKBest*.

Penerapan optimasi ekstraksi fitur dan seleksi fitur dengan cara melakukan 6 kali uji kombinasi pada fitur. Tujuan pengujian ini untuk mendapatkan model klasifikasi yang paling optimal. Selain itu, perbandingan pengujian untuk mendapatkan hasil akurasi yang lebih akurat. Enam kombinasi uji fitur yang dilakukan yaitu TF-IDF, TF-IDF dan *N-Gram*, TF-IDF dan *Chi-Square*, TF-IDF *N-Gram* dan *Chi-Square*, TF-IDF *Chi-Square* dan *SelectKBest*, TF-IDF *N-Gram Chi-Square* dan *SelectKBest.*

TF IDF dan *N-Gram* merupakan kombinasi dari metode ekstraksi fitur. Kombinasi tersebut digunakan untuk mengatasi kelemahan yang ada pada TF-IDF dimana

fitur yang dihasilkan tidak memperhatikan urutan dari kata dalam sebuah kalimat. Hal ini tentunya bisa menghilangkan makna sesungguhnya yang terdapat dalam sebuah kalimat. Maka dari itu diperlukan *N-Gram* yang berfungsi untuk mempertahankan susunan dan urutan kata pada sebuah kalimat. Sementara kombinasi dua seleksi fitur yang digunakan, yaitu *Chi-Square* dan *SelectKBest*. *Chi-Square* berfungsi untuk mengukur hubungan antara variabel nominal dan variabel nominal lainnya, sedangkan *SelectKBest* digunakan untuk meningkatkan akurasi prediksi atau kinerja dataset dengan memilih fitur-fitur yang paling relevan.

Tabel 4. Hasil Ekstraksi Fitur dan Seleksi Fitur

Fitur	Hasil Ekstraksi Fitur dan Seleksi Fitur
TF-IDF	{'acara': 0.382, 'ada': 0.294, 'adakan': 0.431, 'adil': 0.324, 'admin': 0.177, 'adain': 0.523, 'agar': 0.532, 'adalah': 0.365}
TF-IDF & *N-Gram*	{'abdet': 0.184, 'acara': 0.160, 'acarachrismast': 0.224, 'google': 0.390, 'giveaway': 0.088, 'account': 0.205}
TF-IDF & *Chi-Square*	{'acount': 0.612, 'admin': 1.807, 'adil': 1.122, 'android': 0.026, 'mobile': 0.065, 'multipemain': 0.911, 'tropi': 0.659}
TF-IDF, *N-Gram* & *Chi-Square*	{'acount': 0.307, 'baguss': 0.108, 'base': 0.276, 'attack': 0.615, 'bug': 2.455, 'error': 1.520, 'legend': 0.278, 'mahal': 0.728}
TF-IDF, *Chi-Square* & *SelectKBest*	{'admin': 1.807, 'asik': 1.310, 'baik': 0.030, 'disconnected': 0.502, 'gems': 0.007, 'iklan': 0.280, 'wow': 0.341}
TF-IDF, *N-Gram*, *Chi-Square* & *SelectKBest*	{'adil': 0.562, 'royal': 0.074, 'supercell': 0.276, 'tropi': 0.348, 'update': 6.536, 'warliga': 0.135', 'upgrade': 2.088}

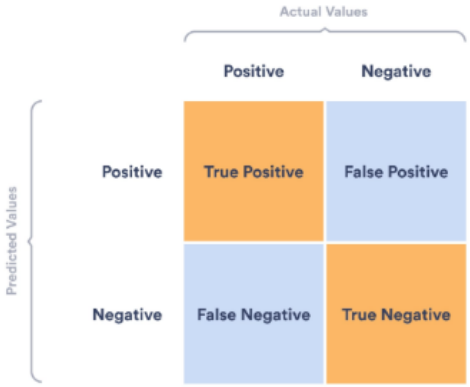
Tabel 4 berisi enam hasil pengujian ekstraksi fitur dan seleksi fitur. Pertama, hasil pembobotan *TF-IDF*. Kedua, hasil pembobotan *TF-IDF* dan *N-Gram*. Ketiga, hasil pembobotan *TF-IDF* dan *Chi-Square*. Keempat, hasil pembobotan *TF-IDF, N-Gram* dan *Chi-Square*. Kelima, hasil pembobotan *TF-IDF, Chi-Square* dan *SelectKBest*. Terakhir, hasil pembobotan *TF-IDF, N-Gram, Chi-Square* dan *SelectKBest*.

3.4 *Machine Learning Model*

Machine learning model adalah hasil dari fase pelatihan (*training phase*) di mana sistem belajar untuk menemukan pola di dalam data. *Train-test split* adalah salah satu metode evaluasi model *machine learning*. Dalam metode ini, dataset dibagi menjadi dua bagian yaitu *training* data dan *testing* data, masing-masing bagian digunakan untuk tujuan yang berbeda. *Training* data digunakan untuk melatih model *machine learning*, sementara *testing* data digunakan untuk mengevaluasi performa model tersebut. Proporsi pembagian antara *training* data dan *testing* data pada penelitian ini adalah 70:30, 80:20, dan 90:10.

3.5 Hasil dan Evaluasi

Setelah mendapatkan data uji dan data latih model akan dievaluasi menggunakan algoritma klasifikasi *Support Vector Machine* dan *Naïve Bayes*. Pada penelitian ini evaluasi model dilakukan dengan bantuan pustaka *sk-learn. Naïve Bayes* pada penelitian ini menggunakan variasi yang sering digunakan untuk klasifikasi teks yaitu *Gaussian Naïve Bayes*, sementara *Support Vector Machine* menggunakan variasi SVC. Setelah model diujikan akan dilakukan perhitungan akurasi untuk mengetahui performa pada tiap algoritma menggunakan tabel confusion matrix. Gambar 3 merupakan tabel *confusion matrix* yang memiliki dua kelas untuk memudahkan dalam proses perhitungan performa dari suatu model algoritma klasifikasi.



4Gambar 3. *Confusion Matrix*

Tabel 5. Hasil Akurasi *Support Vector Machine* (%)

Fitur	Presentase Rasio		
	70 : 30	80 : 20	90 : 10
TF-IDF	88	86,5	86
TF-IDF & *N-Gram*	89	88	85
TF-IDF & *Chi-Square*	90,3	89	85
TF-IDF, *N-Gram* & *Chi-Square*	89,3	88	86
TF-IDF, *Chi-Square* & *SelectKBest*	92,3	92	92
TF-IDF, *N-Gram*, *Chi-Square* & *SelectKBest*	93	90,5	87

Tabel 5 merupakan hasil pengujian yang didapatkan *Support Vector Machine*. Langkah pertama yang dilakukan adalah import *svm* menggunakan *sklern* agar proses klasifikasi yang dilakukan dapat berjalan dalam algoritma *Support Vector Machine*. Selanjutnya *import accuracy score* dari *sklearn metrics* untuk menampilkan akurasi dari hasil pengujian. Dilakukan tiga kali pengujian dengan perbandingan presentase rasio 90:10, 80:20, dan 70:30 dengan *random state* = 0.

Tabel 6. Hasil Akurasi *Naïve Bayes* (%)

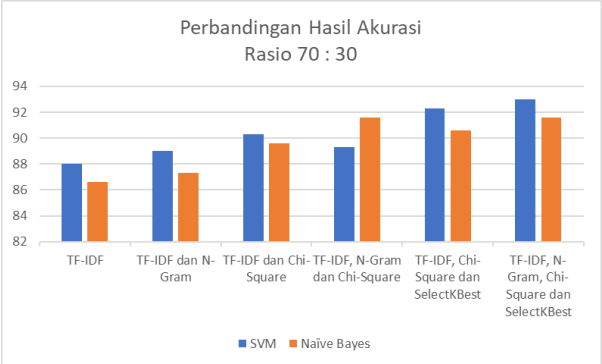
Fitur	Presentase Rasio		
	70 : 30	80 : 20	90 : 10
TF-IDF	86,6	83,5	81
TF-IDF & *N-Gram*	87,3	86,5	87
TF-IDF & *Chi-Square*	89,6	88,5	87
TF-IDF, *N-Gram* & *Chi-Square*	91,5	91,5	91
TF-IDF, *Chi-Square* & *SelectKBest*	90,6	90	88
TF-IDF, *N-Gram*, *Chi-Square* & *SelectKBest*	91,6	90,5	90

Tabel 6 merupakan hasil yang didapatkan pada *Naïve Bayes*. Langkah pertama yang dilakukan adalah import *GaussianNB* menggunakan *sklern naïve bayes* agar proses klasifikasi yang dilakukan dapat berjalan dalam algoritma *Naïve Bayes*. Selanjutnya *import accuracy score* dari *sklearn metrics* untuk menampilkan akurasi dari hasil pengujian. Dilakukan tiga kali pengujian dengan perbandingan presentase rasio 90:10, 80:20, dan 70:30 dengan *random state* = 0.

Tabel 7. Hasil Pengujian Akurasi (%)

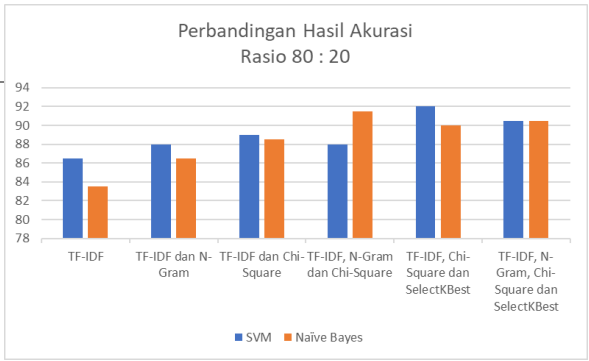
Fitur	Presentase Rasio					
	70 : 30		80 : 20		90 : 10	
	SVM	Naïve Bayes	SVM	Naïve Bayes	SVM	Naïve Bayes
TF-IDF	88	86,6	86,5	83,5	86	81
TF-IDF & *N-Gram*	89	87,3	88	86,5	85	87
TF-IDF & *Chi-Square*	90,3	89,6	89	88,5	87	87
TF-IDF, *N-Gram* & *Chi-Square*	89,3	91,5	88	91,5	86	91
TF-IDF, *Chi-Square* & *SelectKBest*	92,3	90,6	92	90	92	88
TF-IDF, *N-Gram*, *Chi-Square* & *SelectKBest*	93	91,6	90,5	90,5	87	90

Tabel 7 merupakan perbandingan hasil akurasi proses optimasi ekstraksi fitur dan seleksi fitur pada analisis sentimen dengan *Support Vector Machine* dan *Naïve* Bayes. Tabel 12 menerapkan dan mengetahui model terbaik serta mendapatkan hasil akurasi yang lebih akurat. Metode SVM lebih baik daripada Naïve Bayes yang diujikan dengan tiga variasi split rasio 70:30 sampai 90:10.



Gambar 4. *Bar Chart* Rasio 70:30

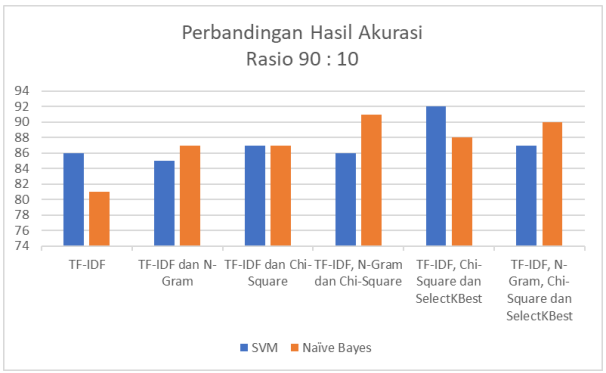
Gambar 4 adalah *bar chart* perbandingan hasil akurasi *Support Vector Machine* dan *Naïve Bayes* pada rasio perbandingan 70:30. Algoritma *TF-IDF, Chi Square* dan *SelectKBest* menghasilkan kinerja akurasi tertinggi 92.3%. Sedangkan algoritma *TF-IDF* mmenghasilkan akurasi terendah 86.6%. Rasio 70:30 menunjukan algoritma SVM lebih tinggi daripada Naïve Baye.



Gambar 5. *Bar Chart* Rasio 80:20

Gambar 5 adalah *bar chart* perbandingan hasil akurasi *Support Vector Machine* dan *Naïve Bayes* pada rasio perbandingan 80:20. Hasi tertinggi pada pengujian ini 92% dengan metode *TF-IDF, Chi Square* dan *SelectKBest*. Hasil terendah dengan metode *TF-IDF* 83.5%. Hasil rasio 80:20 terlihat metode SVM lebih tinggi daripada Naïve Bayes

Gambar 6 adalah *bar chart* perbandingan hasil akurasi *Support Vector Machine* dan *Naïve Bayes* pada rasio perbandingan 90:10. Metode *TF-IDF, Chi Square* dan *SelectKBest* menghasilkan kinerja akurasi tertinggi 92%. Algoritma *TF-IDF* mendapatkan kinerja terendah 81%. Kinerja algoritma SVM lebih baik daripada Naïve Bayes pada rasio 90.10



Gambar 6. *Bar Chart* Rasio 90:10

4. Kesimpulan

Dari hasil pengujian tersebut, dapat disimpulkan bahwa penerapan optimasi ekstraksi fitur TF-IDF dan *N-Gram*, serta seleksi fitur *Chi-Square* dan *SelectKBest* adalah model klasifikasi terbaik dengan menghasilkan akurasi tertinggi dalam analisis sentimen ulasan pengguna *game online Clash of Clans*. Akurasi tertinggi dicapai oleh *Support Vector Machine*, dengan hasil akurasi sebesar 93% pada *training* data dan *testing* data dengan presentase rasio 70:30. Sementara itu, *Naïve Bayes* mencapai akurasi sebesar 91,6% pada presentase rasio pengujian yang sama.

Penerapan optimasi ekstraksi fitur TF-IDF dan *N-Gram*, seleksi fitur *Chi-Square* dan *SelectKBest* pada analisis sentimen pengguna *game online Clash of Clans.* Algoritma *Support Vector Machine* dan *Naïve Bayes* tidak selalu meningkatkan hasil akurasi setelah proses klasifikasi. Metode tersebt hasil akurasi cenderung turun saat proses *training* data yang lebih besar.

Daftar Pustaka

[1] B. Hartanto, M. S. Ferdinand, and Mc. Albert Surya Wijaya, "Pembuatan Game Nonogram Multiplayer".

[2] L. Sihombing and D. Manurung, *Peta Ekosistem Industri Game Indonesia 2021*. 2022.

[3] E. R. Lidinillah, T. Rohana, and A. R. Juwita, "Analisis Sentimen Twitter Terhadap Steam Menggunakan Algoritma Logistic Regression dan Support Vector Machine," *TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika*, vol. 10, no. 2, pp. 154–164, Jul. 2023, doi: 10.37373/tekno.v10i2.440.

[4] Muhammad Alwi, Ninis Anggraini, and Rodia, "Analisis Data Mining Pada Pemilihan Jenis Game Terpopuler Menggunakan Algoritma Apriori," *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, vol. 11, no. 1, pp. 9–15, Apr. 2023, doi: 10.21063/jtif.2023.v11.1.9-15.

[5] R. Kusnadi, Y. Yusuf, A. Andriantony, R. Ardian Yaputra, and M. Caintan, "Analisis Sentimen Terhadap Game Genshin Impact Menggunakan Bert," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 6, no. 2, pp. 122–129, Jul. 2021, doi: 10.36341/rabit.v6i2.1765.

[6] M. S. Arifin, M. Ariyanti, and E. Nurhazizah, "Economics and Digital Business Review Analisis Kualitas Mobile Games Berdasarkan Ulasan Platform Google Play Di Indonesia Menggunakan Metode Text Mining," *Economics and Digital Business Review*, vol. 4, no. 1, pp. 357–368, 2023.

[7] R. Arief and K. Imanuel, "Analisis Sentimen Topik Viral Desa Penari Pada Media Sosial Twitter Dengan Metode Lexicon Based," *Jurnal Ilmiah MATRIK*, vol. 21, no. 3, 2019.

[8] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Jakarta: Penerbit Andi, 2020.

[9] T. Mardiana, H. Syahreva, and T. Tuslaela, "Komparasi Metode Klasifikasi Pasa Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 267–274, Sep. 2019, doi: 10.33480/pilar.v15i2.752.

[10] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) Dengan TF-IDF N-Gram Untuk Text Classification," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 6, pp. 129–136, 2021.

[11] V. Zuliana and I. Maulana, "Analisis Sentimen Program Migrasi TV Digital Menggunakan Algoritma Naive Bayes Dengan Chi Square," *Jurnal informasi dan Komputer*, vol. 10, no. 2, 2022.

[12] E. N. Fitri, S. Winarno, F. Budiman, A. Rohmani, J. Zeniarja, and E. Sugiarto, "Decision Tree Simplification Through Feature Selection Approach In Selecting Fish Feed Sellers," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 2, pp. 301–309, Mar. 2023, doi: 10.52436/1.jutif.2023.4.2.747.

[13] D. Deviacita, H. Sasty, and H. Muhardi, "Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace," *Jurnal Sistem dan Teknologi Informasi*, vol. 7, no. 4, 2019.

[14] F. Prasetya and F. Ferdiansyah, "Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 132, Sep. 2022, doi: 10.30865/json.v4i1.4852.

[15] A. T. Ni'mah and A. Z. Arifin, "Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis," *Rekayasa*, vol. 13, no. 2, pp. 172–180, Aug. 2020, doi: 10.21107/rekayasa.v13i2.6412.

[16] Y. Sergio, V. Putranta, B. Rahayudi, and W. Purnomo, "Analisis Sentimen Masyarakat terhadap Kebijakan Penghapusan Subsidi BBM pada Media Sosial Twitter menggunakan Algoritma Naive Bayes Classifier dengan Ekstraksi Fitur N-Gram TF-IDF," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[17] Y. Setiawan, "Data Mining Berbasis Nearest Neighbor dan Seleksi Fitur Untuk Deteksi Kanker Payudara," *Jurnal Informatika: Jurnal pengembangan IT (JPIT)*, vol. 8, no. 2, pp. 89–96, 2023.

[18] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 437, Apr. 2020, doi: 10.30865/mib.v4i2.2080.

[19] L. Hernando, V. A. Handayani, D. P. Caniago, and N. W. Nasution, "Penerapan Data Mining Dalam Analisa Profil Mahasiswa Menggunakan Metode Support Vector Machine (SVM)," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 2, pp. 477–483, Sep. 2023, doi: 10.37859/coscitech.v4i2.5107.

[20] R. Yendra *et al.*, "Klasifikasi Data Mining Untuk Seleksi Penerimaan Calon Pegawai Negeri Sipil Tahun 2017 Menggunakan Metode Naïve Bayes," *Jurnal Sains Matematika dan Statistika*, vol. 6, no. 1, 2020.

(JOINTECS) Journal of Information Technology and Computer Science Vol . 9 No. 1 (2024) 21 – 30

Halaman ini sengaja dikosongkan