JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989

DOI: 10.26418/justin.v12i4.82758
Vol. 12, No. 4, Oktober 2024

Evaluasi Teknik Preprocessing terhadap Kinerja Multinomial Naïve Bayes dalam Klasifikasi Pertanyaan *Insincere*

Khadijah Fahmi Hayati Holle[1], Rizha Alfianita[2], Hikmatul Maulidia Putri[3]

Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim
Jl. Gajayana No. 50 Malang, Jawa Timur 65144 Indonesia

[1]khadijah.holle@uin-malang.ac.id,
[2] rizhaalfianita1412@gmail.com,
[3] hikmaulidia17@gmail.com

Abstrak

Platform komunitas tanya-jawab atau *Community Question Answering* (CQA) telah menjadi sumber informasi yang penting namun menghadapi tantangan, salah satunya adalah adanya pertanyaan *insincere*. Pertanyaan *insincere* ini mengacu pada pertanyaan yang tidak tulus dan sering didasarkan pada asumsi keliru, yang dapat mengganggu kenyamanan pengguna dan menyebabkan penyebaran informasi yang menyesatkan. Oleh karena itu, diperlukan deteksi pertanyaan *insincere*. Penelitian ini bertujuan untuk mengevaluasi pengaruh teknik preprocessing teks terhadap kinerja algoritma *Multinomial Naïve Bayes* (MNB) dalam mengklasifikasikan pertanyaan *insincere*. Data yang digunakan terdiri dari 4000 pertanyaan dari Quora, dengan masing-masing 2000 pertanyaan berlabel *insincere* dan 2000 berlabel *sincere*. Pembobotan kata dilakukan menggunakan TF-IDF. Terdapat 4 skenario pengujian yang berfokus pada variasi tahap *preprocessing* untuk mengetahui pengaruh *preprocessing* terhadap akurasi sistem. Skenario tersebut adalah MNB dengan *stemming*, MNB dengan *lemmatization*, MNB tanpa *stemming*, dan MNB dengan *stemming* tanpa *stopword removal*. Pengujian dilakukan menggunakan teknik k-Fold Cross Validation. Hasil uji coba menunjukkan bahwa skenario MNB dengan *stemming* tanpa *stopword removal* memberikan hasil terbaik dengan akurasi 83%, presisi 78%, recall 94%, dan F1-score 85%. Sehingga dapat disimpulkan bahwa pemilihan teknik pemrosesan teks yang tepat sangat penting untuk meningkatkan kinerja teks, khususnya dalam mendeteksi pertanyaan insincere pada platform CQA.

Kata kunci: Klasifikasi pertanyaan, Pertanyaan *Insincere,* Quora, *Multinomial Naïve Bayes, Preprocessing Text*

Evaluation of Preprocessing Techniques on the Performance of Multinomial Naïve Bayes in Classifying Insincere Questions

Abstract

Community Question Answering (CQA) platforms have become vital sources of information, yet they face challenges such as insincere questions. Insincere questions are not genuine and often based on false assumptions, disrupting user experience and spreading misleading information. Therefore, detecting insincere questions is essential. This study aims to evaluate the impact of text preprocessing techniques on the performance of the Multinomial Naïve Bayes (MNB) algorithm in classifying insincere questions. The dataset comprises 4000 questions from Quora, equally divided into 2000 insincere and 2000 sincere questions. Word weighting is performed using the TF-IDF technique. Four testing scenarios focusing on preprocessing variations are examined to determine their effect on system accuracy. The scenarios include MNB with stemming, MNB with lemmatization, MNB without stemming, and MNB with stemming without stopword removal. Testing is conducted using k-Fold Cross Validation. The results show that the MNB scenario with stemming without stopword removal yields the best performance, with an accuracy of 83%, precision of 78%, recall of 94%, and F1-score of 85%. It can be concluded that selecting the appropriate text preprocessing techniques is crucial to enhancing text classification performance, particularly in detecting insincere questions on CQA platforms.

Keywords: *Question Classification, Insincere Question, Quora, Multinomial Naïve Bayes, Preprocessing Text*

I. PENDAHULUAN

Platform komunitas tanya-jawab atau Community Question Answering (CQA) menyimpan berbagai macam pertanyaan dan jawaban serta menyediakan informasi yang tidak bisa dengan mudah diakses melalui mesin pencarian umum [1]. CQA memberikan banyak manfaat bagi penggunanyanya, seperti memudahkan mereka dalam mencari jawaban atas pertanyaan yang ada sehingga dapat

menghemat waktu dan usaha. Namun, meskipun memiliki banyak dampak positif, platform CQA ini juga menghadapi beberapa tantangan, seperti menyaring konten toxic, menghapus pertanyaan duplikat, dan mengidentifikasi pengguna yang ahli [2].

Salah satu masalah utama pada platform CQA adalah rendahnya partisipasi pengguna dalam menjawab pertanyaan. Kebanyakan pengguna cenderung hanya mengajukan pertanyaan tanpa memberikan jawaban kepada pengguna lain. Hal ini menyebabkan perbedaan besar antara jumlah pertanyaan yang diajukan dengan jumlah jawaban yang diberikan [3]. Selain itu, banyaknya pertanyaan yang muncul membuat pengguna kesulitan menemukan pertanyaan yang relevan dengan keahlian mereka, terutama jika pertanyaan tersebut tertutup oleh konten toxic [4].

Konten toxic adalah ancaman serius yang dihadapi oleh 48% pengguna internet. Media sosial seperti Facebook, Twitter, dan YouTube sering kali menghadapi masalah dari pengguna yang tidak bertanggung jawab yang mencoba merusak integritas platform tersebut [5]. Platform CQA seperti Quora, Stack Overflow, Reddit, dan forum diskusi lainnya juga tidak luput dari masalah ini, di mana pengguna sering kali membuat postingan yang bersifat toxic dan menyerang pihak tertentu. Salah satu jenis konten toxic adalah pertanyaan *insincere*.

Quora telah mengidentifikasi karakteristik dari pertanyaan *insincere*. Pertanyaan *insincere* adalah pertanyaan yang didasarkan pada asumsi yang salah atau yang dimaksudkan untuk membuat pernyataan daripada mencari jawaban yang membantu. Beberapa karakteristik dari pertanyaan insincere antara lain: nada yang tidak netral, bersifat meremehkan atau menghasut, tidak berdasarkan kenyataan, dan menggunakan konten seksual. Nada yang tidak netral berarti pertanyaan yang bernada berlebihan dengan tujuan menjatuhkan sekelompok orang. Pertanyaan yang bersifat meremehkan atau menghasut mendorong tindakan diskriminatif terhadap suatu kelompok. Pertanyaan yang tidak berdasarkan kenyataan mengandung informasi yang salah atau asumsi yang tidak masuk akal, sementara pertanyaan dengan konten seksual mengandung konten yang tidak pantas seperti pedofilia, inses, dan bestialitas. [6]

Deteksi pertanyaan *insincere* penting bagi platform CQA. Pertanyaan *insincere* dapat mengganggu diskusi yang konstruktif dan mengurangi kualitas informasi yang tersedia di platform tersebut. Selain itu, pertanyaan-pertanyaan semacam ini dapat mempengaruhi kepercayaan pengguna terhadap platform dan menurunkan tingkat partisipasi aktif pengguna. Oleh karena itu, deteksi dan penghapusan pertanyaan insincere menjadi krusial untuk menjaga integritas dan kualitas dari platform-platform tersebut.

Terdapat beberapa metode klasifikasi yang biasa digunakan untuk klasifikasi teks, seperti Support Vector Machine (SVM)[7], [8], [9], Random Forest, algoritma Deep Learning[10], maupun Naïve Bayes[7], [11], [12], [13]. Meskipun relatif sederhana, banyak penelitian menunjukkan bahwa klasifikasi teks menggunakan Naïve Bayes cukup memberikan hasil yang baik. Variasi algoritma naïve bayes yang secara khusus dirancang untuk menangani data dengan representasi berbasis hitungan adalah algoritma Multinomial Naïve Bayes (MNB)[9], [11], [14], [15], [16].

Multinomial Naïve Bayes (MNB) telah terbukti efektif dalam memahami pola-pola teks untuk mengerjakan tugas klasifikasi teks. Namun, penggunaan algoritma ini masih memiliki tantangan, terutama terkait dengan kualitas data yang digunakan. Salah satu upaya untuk meningkatkan kinerja MNB adalah dengan melakukan teknik preprocessing yang tepat pada data teks yang akan diklasifikasikan. Teknik preprocessing ini meliputi berbagai tahapan seperti case folding, tokenizing, stopword removal, stemming, dan lemmatization. Setiap teknik memiliki peran penting dalam menyederhanakan dan membersihkan teks sehingga lebih mudah dianalisis oleh algoritma klasifikasi.

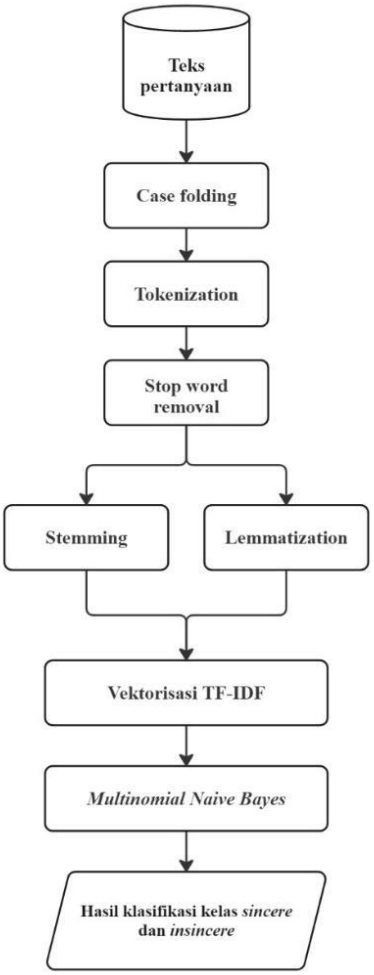
Penelitian ini bertujuan untuk mengevaluasi pengaruh berbagai teknik preprocessing terhadap kinerja algoritma MNB dalam mengklasifikasikan pertanyaan *insincere*. Melalui penelitian ini, diharapkan dapat diketahui teknik preprocessing mana yang paling efektif dalam meningkatkan akurasi dan kinerja keseluruhan dari algoritma MNB. Penelitian ini juga bertujuan untuk memberikan panduan praktis bagi peneliti dan praktisi dalam memilih dan menerapkan teknik preprocessing yang sesuai untuk keperluan klasifikasi teks, khususnya dalam konteks deteksi pertanyaan insincere pada platform CQA.

Penelitian ini akan mengevaluasi empat skenario preprocessing yang berbeda, yang masing-masing mewakili kombinasi yang berbeda dari tahapan preprocessing termasuk penghapusan stopwords, stemming, dan lemmatization. Kami akan meneliti pengaruh masing-masing teknik preprocessing terhadap kinerja MNB dalam mengidentifikasi pertanyaan-pertanyaan *insincere*. Sehingga dapat diketahui bagaimana pengaruh preprocessing terhadap akurasi sistem klasifikasi ini.

Dengan demikian, kontribusi penelitian ini tidak hanya terletak pada peningkatan akurasi klasifikasi tetapi juga pada pemahaman yang lebih baik tentang proses pra-pemrosesan data dalam konteks klasifikasi teks. Hasil penelitian ini diharapkan dapat memberikan wawasan bagi para peneliti dan praktisi dalam mengoptimalkan sistem klasifikasi teks mereka.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan beberapa teknik preprocessing untuk meningkatkan kinerja model Multinomial Naïve Bayes dalam klasifikasi pertanyaan insincere. Teknik preprocessing yang dipilih meliputi tokenisasi, stopword removal, stemming, dan representasi Term Frequency-Inverse Document Frequency (TF-IDF). Desain sistem yang diilustrasikan pada Gambar 1 menggambarkan alur keseluruhan dari sistem klasifikasi, dimulai dari masukan data hingga menghasilkan teks yang telah dikelompokkan ke dalam kategori yang sesuai.



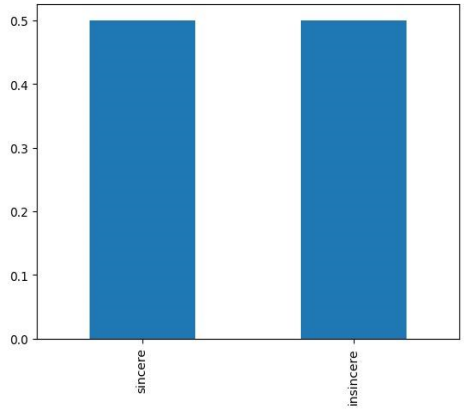
Gambar 1. Desain Sistem

Proses klasifikasi pada penelitian ini dimulai dengan memasukkan data teks pertanyaan sebagai input. Data tersebut kemudian melewati tahap preprocessing untuk membersihkan dan mempersiapkan data sebelum proses analisis lebih lanjut. Pada tahap preprocessing terdapat lima proses, yaitu *case folding*, *tokenizing*, *stopword removal*, *stemming*, dan *lemmatization*.

Setelah tahap preprocessing selesai, setiap kata dalam data diberi bobot menggunakan pembobotan TF-IDF. Fitur-fitur yang dihasilkan dari TF-IDF kemudian diproses dengan menggunakan metode *Multinomial Naïve Bayes* untuk melakukan klasifikasi. Dalam klasifikasi ini, terdapat dua kelas yang dihasilkan, yaitu *sincere* dan *insincere*.

A. Pengumpulan Data

Dataset yang digunakan pada penelitian ini berupa data sekunder yang dikumpulkan dari situs Kaggle.com dan dikenal dengan nama "*Quora Insincere Questions Classification*". Dataset ini terdiri dari pertanyaan-pertanyaan dalam bahasa Inggris yang dikumpulkan dari platform tanya jawab Quora. Selain teks pertanyaan, dataset ini juga mencangkup label untuk setiap pertanyaan, yaitu *sincere* dan *insincere*. Setiap kelas memiliki 2000 pertanyaan yang sudah ditandai. Gambar 2 menunjukkan pembagian data yang digunakan.



Gambar 2. Pembagian Data

Pemilihan dataset ini didasarkan pada relevansinya dengan fokus penelitian kami, yang bertujuan untuk mengklasifikasikan pertanyaan-pertanyaan *insincere*. Melalui dataset ini, kami dapat mengidentifikasi pola-pola dalam pertanyaan yang tidak tulus, yang diharapkan dapat membantu dalam pengembangan sistem deteksi dan penanganan konten yang tidak pantas di platform tanya jawab tersebut.

Dalam konteks dataset ini, pertanyaan *sincere* merujuk pada pertanyaan yang diajukan dengan niat yang tulus untuk memperoleh jawaban yang bermanfaat, sementara pertanyaan *insincere* merujuk pada pertanyaan yang didasarkan pada asumsi yang salah, berisi provokasi, atau tidak bermaksud mencari informasi yang berguna.

TABEL 1. CONTOH PERTANYAAN

Kalimat Pertanyaan	Label
What options are available for me with 79.6% in class 12th?	*Sincere*
How can I distinguish between vanity and purpose?	*Sincere*
Why do vegans always try to bully me are they brain damaged?	*Insincere*
Why Indian people are not so humble?	*Insincere*

B. Preprocessing Teks

Tahap preprocessing merupakan langkah penting dalam mempersiapkan *corpus* untuk pemodelan dan memiliki dampak signifikan terhadap hasil dari system yang dibangun [17]. Preprocessing bertujuan untuk menyederhanakan teks menjadi bentuk yang lebih terstruktur dan memungkinkan analisis yang lebih efektif. Dalam penelitian ini, berbagai metode preprocessing digunakan, termasuk case folding, tokenization, stop word removal, stemming, dan lemmatization [18].

Tahapan preprocessing yang dilakukan dalam penelitian ini terdiri dari:

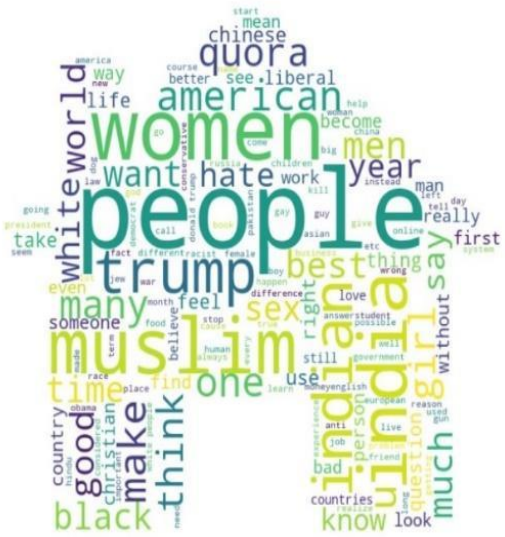
1) *Case Folding:* tahapan ini melibatkan pengubahan semua huruf dalam teks menjadi huruf kecil (lowercase). Hal ini dilakukan untuk memastikan konsistensi dalam analisis teks.
2) *Tokenization:* proses tokenization memecahkan teks menjadi unit-unit yang lebih kecil, yaitu kata-

kata atau token. Setiap kata dalam teks dianggap sebagai token yang independent dan dapat dianalisis secara terpisah.

3) *Stopword Removal:* tahap ini bertujuan untuk menghapus kata-kata yang umumnya tidak memberikan kontribusi signifikan pada makna teks, seperti kata-kata penghubung dan kata-kata umum lainnya. Contoh kata-kata seperti "*and*", "*or*", dan "*in*" biasanya dianggap sebagai stopwords dan dihapus dari teks.
4) *Stemming:* proses stemming dilakukan pengubahan kata-kata ke bentuk dasarnya atau akar kata, dengan menghapus imbuhan atau afiksasi. Ini dilakukan untuk mengurangi variasi dalam representasi kata, sehingga kata-kata dengan bentuk yang sama dapat dianggap sama. Hal ini dapat mengurangi kompleksitas model dengan mengurangi jumlah fitur unik.
5) *Lemmatization:* Lemmatization mirip dengan stemming namun lebih kompleks karena memperhatikan stuktur dan makna kata. Proses ini mengubah kata-kata ke bentuk dasarnya dalam kamus, sehingga memastikan konsistensi dan akurasi dalam analisis teks [19].

Tahapan preprocessing tersebut penting untuk menghasilkan data yang bersih dan terstruktur, yang dapat dianalisis dengan efektif menggunakan algoritma klasifikasi seperti Multinomial Naïve Bayes. Setiap tahapan memiliki peran khusus dalam menyederhanakan dan mengoptimalkan representasi teks untuk analisis lebih lanjut.

Terms output tahap preprocessing dapat ditampilkan dalam bentuk *word cloud* seperti yang ditunjukkan pada Gambar 3. Dari *word cloud* tersebut dapat diketahui kata yang sering muncul pada dataset. Tiga kata yang paling sering muncul diantaranya "*People*", "*Woman*", dan "Muslim".



Gambar 3. Word Cloud

C. Ekstraksi Fitur TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) adalah algoritma yang digunakan untuk mengekstrak atribut atau fitur yang dapat mendeskripsikan dokumen yang tidak dikategorikan dengan baik. Metode TF-IDF bekerja dengan memberikan bobot pada setiap *term* di dokumen yang direpresentasikan dalam bentuk matriks [21]. TF-IDF berguna dalam menilai pentingnya kata dalam dokumen tertentu relatif terhadap koleksi dokumen lainnya.

Pada TF-IDF, jika suatu *term* sering muncul dalam satu atau beberapa dokumen tetapi tidak umum disemua dokumen, *term* tersebut dianggap penting dan diberi bobot tinggi. Namun, jika suatu *term* muncul dihampir semua dokumen, maka *term* tersebut dianggap umum dan diberi bobot rendah [22].

TF-IDF terdiri dari dua komponen utama, yaitu Term Frequency (TF) dan Invers Document Frequency (IDF). TF mengukur mengukur frekuensi kemunculan suatu kata dalam dokumen tertentu. IDF mengukur seberapa penting kata tersebut dalam seluruh koleksi dokumen. Kata yang jarang muncul dibanyak dokumen akan memiliki nilai IDF yang lebih tinggi, menunjukkan bahwa kata tersebut memberikan informasi yang lebih spesifik [23].

Rumus untuk TF dan IDF diberikan pada Persamaan (1) dan (2):

$$TF\ (w) = \frac{frekuensi\ muncul\ kata\ w\ di\ dokumen\ d}{Total\ kata\ dokumen\ d} \quad (1)$$

$$IDF\ (w) = log_e \frac{Jumlah\ total\ dokumen}{Total\ dokumen\ pada\ kata\ w} \quad (2)$$

Komponen TF mengukur seberapa sering kata *w* muncul dalam dokumen *d*, sedangkan komponen IDF mengukur seberapa unik atau jarang kata tersebut muncul dalam koleksi dokumen. Nilai TF-IDF diperoleh dengan mengalikan nilai TF dan IDF sebagaimana ditunjukkan pada Persamaan (3)

$$TF - IDF\ (w, d) = TF(w)\ \times\ IDF(w) \quad (3)$$

Contohnya, jika kata "machine" muncul 3 kali dalam sebuah dokumen dengan total 100 kata, nilai TF dari "machine" adalah $\frac{3}{100} = 0.03$. Jika "machine" muncul dalam 10 dari 1000 dokumen dalam koleksi, nilai IDF dari "machine" adalah $\left(\frac{1000}{10}\right) = 2.3$. Maka, nilai TF-IDF dari "machine" dalam dokumen tersebut adalah 0.03 × 2.3 = 0.069.

Dengan menggunakan TF-IDF, fitur yang diekstraksi dari teks dapat diolah untuk melakukan klasifikasi lebih lanjut, memungkinkan algoritma seperti Multinomial Naïve Bayes untuk bekerja lebih efektif dalam mengidentifikasi pertanyaan *insincere*.

D. Algoritma Multinomial Naïve Bayes

Multinomial Naïve Bayes adalah varian dari algoritma Naïve Bayes yang khusus digunakan untuk klasifikasi teks. Naïve Bayes sendiri merupakan metode machine learning yang menerapkan konsep probabilistik berdasarkan Teorema Bayes. Metode ini mengasumsikan bahwa setiap fitur (atau atribut) bersifat independent satu sama lain [24].

Multinomial Naïve Bayes bekerja dengan mempertimbangkan frekuensi kemunculan kata dalam dokumen[25]. Dalam konteks ini, kata-kata dalam dokumen diperlakukan sebagai fitur yang digunakan untuk memprediksi kelas dari dokumen tersebut, seperti "sincere" atau "insincere". Karena metode ini termasuk supervised learning, maka data yang digunakan harus berlabel.

Menurut Teorema Bayes, *Multinomial Naïve Bayes* mengkombinasikan *priori* (probalititas awal) dengan kontribusi dari masing-masing fitur (term) atau *likelihood*, kemudian membagi hasilnya dengan *evidence* [26]. Konsep ini dikenal sebagai *maximum a posteriori* (MAP). Untuk menghitung MAP, perlu diketahui nilai *priori* yang dapat dilihat pada Persamaan (4):

$$P(class) = \frac{\sum class}{\sum sentence} \quad (4)$$

Pada Persamaan (4), $P(class)$ merepresentasikan *priori* dari kelas *sincere* dan *insincere*, menunjukkan probabilitas kalimat-kalimat yang termasuk ke dalam masing-masing kelas sebelum mempertimbangkan *evidence*. *Priori* dihitung dengan membagi jumlah kalimat yang termasuk ke dalam suatu kelas dengan total kalimat dari seluruh data yang ada pada data *training* [27].

Selanjutnya, terdapat perhitungan dari *likelihood*. *Likelihood* mengukur seberapa baik fitur kata menjelaskan kelas dari suatu kalimat. Perhitungan dari *likelihood* terdapat pada Persamaan (5):

$$\sum P(term_i|class) = \frac{count(term_i, class) + 1}{\sum count(term) + |V|} \quad (5)$$

Pertama, dilakukan perhitungan berapa kali suatu fitur (term) muncul di suatu kelas, dalam konteks ini kelas *sincere* atau *insincere*. Selanjutnya, dihitung juga total fitur (term) yang ada di dalam suatu kelas dan dijumlahkan dengan total vocabulary yang ada di kelas tersebut dimana data yang digunakan adalah data *training*. Hasil dari keduanya dibagi dan menjadi nilai dari *likelihood* [28]. Karena *Multinomial Naïve Bayes* mengalikan semua fitur *likelihood*, jika terdapat satu *likelihood* yang bernilai 0 maka probabilitas *posteriori* atau outputnya akan menjadi 0 tidak peduli berapa nilai *likelihood* yang lain [29]. Untuk mengatasi hal ini, dilakukan *laplacian smoothing* [30], yaitu dengan menambahkan angka 1 pada perhitungan *likelihood* seperti yang dapat dilihat pada Persamaan (5).

Output dari model klasifikasi *Multinomial Naïve Bayes* adalah probabilitas posterior (*posteriori*). Kelas dari suatu kalimat ditentukan berdasarkan hasil probabilitas posterior terbesar atau *maximum a posteriori* (MAP). Dengan demikian, persamaan *posteriori* untuk model klasifikasi dapat ditulis seperti pada Persamaan (6):

$$y = argmax_{kelas}\, P(term_1, \dots, term_n|kelas)\, P(kelas) \quad (6)$$

dimana y menunjukkan kelas yang akan diprediksi berdasarkan fitur-fitur dari suatu kalimat. Untuk menghitungnya, nilai *priori* suatu kelas dikalikan dengan *likelihood* dari fitur ke-1 hingga ke-n. Kelas yang memiliki probabilitas posterior terbesar akan dipilih sebagai kelas yang diprediksi.

E. Cross Validation

Cross Validation adalah teknik yang digunakan untuk membagi data menjadi set pelatihan dan pengujian, yang bertujuan untuk mengevaluasi efektivitas suatu model [20]. Data pelatihan digunakan untuk mengajarkan algoritma dalam proses komputasi, sedangkan data pengujian digunakan untuk memvalidasi perhitungan tersebut. Salah satu teknik cross validation yang umum digunakan adalah k-Fold Cross Validation.

Pendekatan k-fold cross validation melibatkan pembagian dataset menjadi k partisi atau "folds". Pada setiap iterasi, satu fold digunakan sebagai data pengujian, sementara k-1 fold lainnya digunakan sebagai data pelatihan. Proses ini diulang sebanyak k kali, sehingga setiap fold digunakan sekali sebagai data pengujian. Hasil dari iterasi ini kemudian dirata-rata untuk memberikan estimasi kinerja model yang lebih akurat dan mengurangi variabilitas hasil.

Dalam penelitian ini, kami menggunakan 5-fold dan 10-fold cross validation untuk mengevaluasi model. Pembagian ini dilakukan secara acak oleh sistem untuk memastikan bahwa setiap fold representative dari keseluruhan dataset.

F. Evaluasi Model

Evaluasi model pada penelitian ini dilakukan terhadap 4 variasi berdasarkan tahap preprocessing yang digunakan. Setiap scenario dirancang untuk menguji pengaruh Teknik preprocessing tertentu terhadap kinerja algoritma Multinomial Naïve Bayes. Keempat scenario tersebut adalah:

1. Skenario pertama, model diuji menggunakan kombinasi tahapan stopword removal dan stemming. Pada tahap preprocessing disini, kata-kata umum yang tidak memiliki makna penting (stopwords) dihapus, dan kata-kata dikonversi ke bentuk dasarnya menggunakan stemming.
2. Skenario kedua, model diuji menggunakan kombinasi tahapan stopword removal dan lemmatization. Disini stopwords dihapus, dan kata-kata dikonversi ke bentuk dasarnya menggunakan lemmatization yang memperhatikan konteks dan makna kata.
3. Skenario ketiga, model diuji menggunakan kombinasi tahapan stopword removal tanpa stemming dan lemmatization. Dalam skenarion ini, hanya stopwords yang dihapus, dan kata-kata tetap dalam bentuk aslinya.
4. Skenario keempat, model diuji menggunakan kombinasi tahapan stemming tanpa stopword removal. Kata-kata dikonversi ke bentuk dasarnya menggunakan stemming, tetapi stopwords tidak dihapus.

Pengujian empat skenario tersebut dilakukan menggunakan evaluasi model dengan *k-fold cross validation*. Proses ini memastikan bahwa hasil evaluasi memperhitungkan variasi dalam dataset dan memberikan gamabran yang lebih akurat tentang konerja model di berbagai kondisi preprocessing.

III. HASIL DAN PEMBAHASAN

Penelitian ini mengevaluasi empat skenario preprocessing berbeda untuk menilai kinerja Multinomial Naïve Bayes (MNB) dalam klasifikasi pertanyaan insincere, menggunakan data dari *Quora Insincere Questions Classification*. Data tersebut terdiri dari pertanyaan berbahasa Inggris, yang dibagi menjadi pertanyaan *sincere* dan *insincere*.

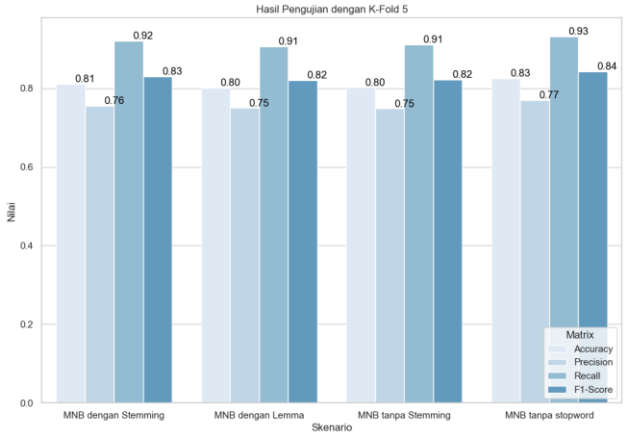
A. *Hasil Uji Coba*

Pengujian *k-fold* yang digunakan sebanyak 5 dan 10 partisi untuk masing-masing kondisi: dengan stemming, dengan lemmatization, tanpa stemming dan lemmatization, serta tanpa stopword removal. Perbandingan hasil waktu pemrosesan tiap skenario dapat dilihat dari Tabel 2 berikut.

TABEL 2. PERBANDINGAN WAKTU PEMROSESAN

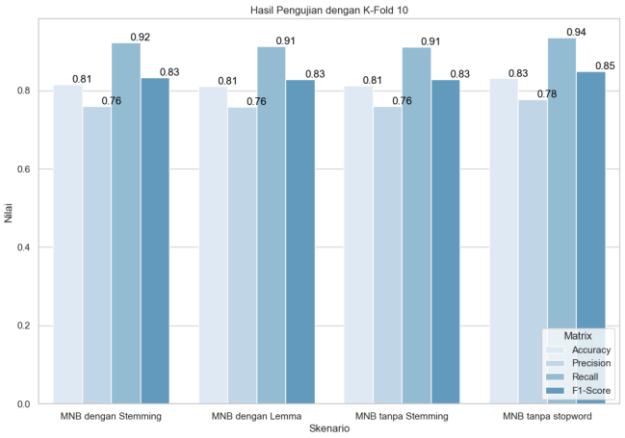
Metode	K-Fold		Waktu Preprocesing
MNB dengan *stemming*	5	10	4.6s
MNB dengan *lemmatization*	5	10	5.7s
MNB tanpa *stemming/lemma*	5	10	3.2s
MNB tanpa *stopword*	5	10	3.3s

Dari Table 2, diketahui bahwa metode *Multinomial Naïve Bayes* tanpa proses *stemming/lemmatization* dan metode *Multinomial Naïve Bayes* tanpa proses *stopword removal* memiliki waktu preprocessing yang relative sama dan lebih cepat dibandingkan metode MNB dengan *stemming* atau dengan *lemmatization.* Sedangkan perbandingan hasil performa masing-masing skenario ditunjukkan pada Gambar 4. Gambar grafik tersebut menunjukkan perbandingan nilai akurasi, recall, presisi, dan f-measure dari pengujian dengan 5-folds.



Gambar 4. Grafik hasil pengujian 4 skenario dengan 5-Folds

Berdasarkan hasil pengujian performa pada tiap skenario, tampak bahwa selisih nilai akurasi tiap skenario dengan 5-folds tidak menunjukkan perbedaan yang signifikan. Nilai akurasi tertinggi yang diperoleh sebesar 83% yaitu pada penggunaan metode *Multinomial Naïve Bayes* tanpa proses *stopword removal*. Selanjutnya metode MNB dengan tanpa stemming memiliki nilai akurasi yang sama dengan metode MNB dengan lemmatization yaitu sebesar 80%. Sedangkan metode MNB dengan stemming menghasilkan akurasi sebesar 81%.



Gambar 5. Grafik perbandingan pengujian K-Fold 10

Secara keseluruhan, pada pengujian 5-folds cross validation, penggunaan algoritma MNB tanpa stopword removal menghasilkan akurasi, recall, presisi, dan f1-score tertinggi yaitu akurasi 83%, presisi 77%, recall 93%, dan f1-score 84%. Sedangkan perbandingan tiap skenario dengan pengujian menggunakan k-fold 10 ditunjukkan pada Gambar 5. Grafik perbandingan pengujian performa pada tiap skenario pada gambar tersebut, menunjukkan bahwa hasil pengujian tiap skenario dengan k-fold 10 hampir sama dengan hasil pengujian k-fold 5. Penjabaran perolehan hasil skenario terbaik dengan perbandingan k-fold seperti pada Tabel 3.

TABEL 3. PERBANDINGAN HASIL UJI COBA SETIAP SKENARIO

No	Nilai K	Skenario	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1	5	MNB dengan *stemming*	0.81	0.75	0.92	0.83
2		MNB dengan *lemmatization*	0.80	0.75	0.91	0.82
3		MNB tanpa *stemming*	0.80	0.75	0.91	0.82
4		**MNB tanpa *stopword***	**0.83**	**0.77**	**0.93**	**0.84**
5	10	MNB dengan *stemming*	0.81	0.76	0.92	0.83
6		MNB dengan *lemmatization*	0.81	0.76	0.91	0.83
7		MNB tanpa *stemming*	0.81	0.76	0.91	0.83
8		**MNB tanpa *stopword***	**0.83**	**0.78**	**0.94**	**0.85**

Hasil akurasi tiap skenario menunjukkan tidak ada perbedaan secara signifikan. Pada pengujian skenario menggunakan k-fold 10, hasil perbandingan nilai akurasi terbaik dari kinerja masing-masing skenario masih diperoleh pada pengujian *Multinomial Naïve Bayes* dengan tanpa *stopword removal* memperoleh akurasi sebesar 83%. Sedangkan untuk tiga skenario lain akurasi yang diperoleh sama-sama bernilai 81%.

Dari rangkaian uji coba tersebut diketaui bahwa *Multinomial Naïve Bayes* tanpa *stopword removal* pada

penelitian ini memperoleh hasil lebih baik dibandingkan dengan skenario pengujian yang lainnya. Dengan kata lain, algoritma *Multinomial Naïve Bayes* dapat digunakan untuk melakukan klasifikasi teks dengan hasil yang baik meskipun tidak dilakukan *stopword removal* pada tahap preprocessing teks. Tanpa penggunaan tahap stopword removal maka waktu proses dalam melakukan klasifikasi pun dapat lebih cepat yaitu 3.3s sebagaimana yang ditunjukkan pada Tabel 2.

B. Pembahasan

Hasil uji coba menunjukkan bahwa skenario yang menggunakan kombinasi stemming tanpa stopword removal memberikan hasil terbaik dalam hal akurasi klasifikasi menggunakan algoritma Multinomial Naïve Bayes dibandingkan dengan skenario lainnya.

Berikut ini beberapa faktor yang menjadikan skenario MNB dengan stemming tanpa stopword removal memberikan hasil yang lebih baik:

1) *Pentingnya informasi kontekstual dalam Bahasa Inggris:* Dalam bahasa Inggris, stopword seperti "*the*", "*is*", "*at*", "*which*", dan "*on*" sering dianggap tidak penting dan biasanya dihapus selama preprocessing. Namun, dalam konteks pertanyaan *insincere*, *stopword* ini bisa memberikan petunjuk penting terkait tone atau maksud pertanyaan. Contohnya, pertanyaan seperti "*Why is this so unfair?*" atau "*How can anyone believe this?*" menggunakan kata-kata yang mungkin dianggap *stopword* tetapi memberikan indikasi ketidakjujuran atau provokasi.
2) *Stemming dan homogenisasi variasi kata:* Stemming mengurangi kata-kata ke bentuk dasarnya, seperti "*running*" menjadi "*run*", yang membantu dalam mengurangi variasi kata dan memperjelas pola dalam data. Ini sangat penting dalam bahasa Inggris, di mana variasi kata bisa sangat banyak dan dapat membingungkan model jika tidak ditangani dengan baik. Misalnya, kata "*criticize*", "*criticized*", dan "*criticizing*" akan di-stem menjadi "*critic*", sehingga mempermudah MNB dalam mengidentifikasi pola yang relevan untuk membedakan pertanyaan insincere dari yang sincere.
3) *Interaksi antara Stopword dan Kata Penting*: Dalam bahasa Inggris, stopword sering kali berinteraksi dengan kata-kata yang lebih bermakna untuk membentuk konteks yang lengkap. Menghapus stopword bisa mengurangi konteks penting ini. Sebagai contoh, pertanyaan "*How can someone not see this?*" kehilangan nuansa skeptis jika kata "*not*" dihapus. Nuansa ini penting untuk mengidentifikasi pertanyaan insincere yang sering kali memiliki nada provokatif atau meremehkan.

Adapun implikasi temuan dari penelitian evaluasi teknik preprocessing pada Multinomial Naïve Bayes untuk klasifikasi pertanyaan insincere diantaranya sebagaimana berikut ini:

1) *Strategi preprocessing yang disesuaikan dengan Bahasa Inggris*: Hasil uji coba menekankan bahwa dalam tugas klasifikasi teks berbahasa Inggris, seperti deteksi pertanyaan insincere, stopword dapat memiliki nilai kontekstual yang signifikan dan sebaiknya tidak dihapus secara otomatis. Ini menunjukkan bahwa pendekatan standar dalam penghapusan stopword mungkin tidak selalu sesuai, terutama dalam konteks yang memerlukan analisis nuansa bahasa yang lebih mendalam.
2) *Peningkatan model klasifikasi dalam Bahasa Inggris:* Dengan mempertahankan stopword dan menerapkan stemming, model MNB dapat dioptimalkan untuk menangkap pola bahasa yang lebih kompleks dan kontekstual dalam teks berbahasa Inggris. Ini dapat meningkatkan akurasi dan efektivitas model dalam mendeteksi pertanyaan insincere, yang sering kali menggunakan strategi bahasa yang halus untuk menyampaikan nada negatif atau meremehkan.
3) Aplikasi praktis untuk Platform Global: Mengingat data yang digunakan berasal dari Quora, sebuah platform tanya-jawab yang beroperasi secara global, hasil ini relevan bagi platform lain yang juga menghadapi masalah serupa dengan konten berbahasa Inggris. Implementasi teknik preprocessing yang tepat dapat membantu dalam menyaring pertanyaan-pertanyaan yang mengganggu atau tidak relevan, menjaga kualitas dan integritas platform, serta meningkatkan kepercayaan dan partisipasi pengguna.

IV. KESIMPULAN

Penelitian ini mengkaji efektivitas teknik preprocessing dalam klasifikasi pertanyaan insincere pada data dari Quora menggunakan model Multinomial Naïve Bayes (MNB). Hasil menunjukkan bahwa penggunaan teknik stemming tanpa stopword removal memberikan hasil terbaik dengan akurasi 83%, precision 77%, recall 93%, dan F1-score 84%. Hal ini menunjukkan bahwa meskipun stopwords tidak dihapus, penggunaan stemming mampu menyederhanakan kata tanpa kehilangan informasi penting, yang berkontribusi pada peningkatan performa model.

Berikut ini saran penelitian lebih lanjut agar dapat mengembangkan kontribusi

1. Ekspansi dataset: Menggunakan dataset yang lebih besar dan bervariasi untuk meningkatkan generalisasi model.
2. Eksplorasi teknik preprocessing lanjutan: Meneliti teknik preprocessing lainnya seperti word embedding atau metode pengurangan dimensi yang lebih canggih untuk menangani data teks.
3. Penggunaan model lain: Membandingkan hasil dengan model pembelajaran mesin atau deep learning lainnya seperti Random Forest, SVM, atau model transformer seperti BERT untuk melihat apakah ada peningkatan kinerja yang signifikan.
4. Aplikasi dalam deteksi konten: Menggunakan temuan ini untuk mengembangkan alat deteksi konten berbahaya atau tidak pantas di platform online, membantu menjaga integritas komunitas online.

DAFTAR PUSTAKA

[1] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, "Finding expert users in community question answering," in *Proceedings of the 21st International Conference on World Wide Web*, New York, NY, USA: ACM, Apr. 2012, pp. 791–798. doi: 10.1145/2187980.2188202.

[2] P. K. Roy, "Multilayer Convolutional Neural Network to Filter Low Quality Content from Quora," *Neural Process Lett*, vol. 52, no. 1, pp. 805–821, Aug. 2020, doi: 10.1007/s11063-020-10284-x.

[3] N. Ghasemi, R. Fatourechi, and S. Momtazi, "User Embedding for Expert Finding in Community Question Answering," *ACM Trans Knowl Discov Data*, vol. 15, no. 4, Jun. 2021, doi: 10.1145/3441302.

[4] M. Neshati, Z. Fallahnejad, and H. Beigy, "On dynamicity of expert finding in community question answering," *Inf Process Manag*, vol. 53, no. 5, pp. 1026–1042, Sep. 2017, doi: 10.1016/j.ipm.2017.04.002.

[5] P. K. Roy, J. P. Singh, A. M. Baabdullah, H. Kizgin, and N. P. Rana, "Identifying reputation collectors in community question answering (CQA) sites: Exploring the dark side of social media," *Int J Inf Manage*, vol. 42, pp. 25–35, Oct. 2018, doi: 10.1016/j.ijinfomgt.2018.05.003.

[6] D. Y. Kim, X. Li, S. Wang, Y. Zhuo, and R. K. W. Lee, "Topic enhanced word embedding for toxic content detection in Q&A sites," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, Association for Computing Machinery, Inc, Aug. 2019, pp. 1064–1071. doi: 10.1145/3341161.3345332.

[7] N. E. Febriyanty, M. A. Hariyadi, and C. Crysdian, "Hoax Detection News Using Naïve Bayes and Support Vector Machine Algorithm," *International Journal of Advances in Data and Information Systems*, vol. 4, no. 2, pp. 191–200, Oct. 2023, doi: 10.25008/ijadis.v4i2.1306.

[8] E. Fitri, F. R. Lumbanraja, and A. Ardiansyah, "KLASIFIKASI ABSTRAK JURNAL KOMPUTASI MENGGUNAKAN METODE TEXT MINING DAN ALGORITMA SUPPORT VECTOR MACHINE," *Jurnal Pepadun*, vol. 1, no. 1, pp. 83–88, Dec. 2020, doi: 10.23960/pepadun.v1i1.13.

[9] F. Shofiya, D. Arifianto, M. Kom, H. Azizah, A. Faruq, and M. Pd, "PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DAN MULTINOMIAL NAIVE BAYES (MNB) DALAM KLASIFIKASI ABSTRAK TUGAS AKHIR (STUDI KASUS: FAKULTAS TEKNIK UNIVERSITAS MUHAMMADIYAH JEMBER)," 2020.

[10] M. Rivest, E. Vignola-Gagné, and É. Archambault, "Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling," *PLoS One*, vol. 16, no. 5, p. e0251493, May 2021, doi: 10.1371/journal.pone.0251493.

[11] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes." [Online]. Available: www.kompas.com

[12] A. Prayoga Permana, K. Ainiyah, and K. Fahmi Hayati Holle, "Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up," 2021. [Online]. Available: https://www.kaggle.com/manishkc06/startup-success-prediction.

[13] Y. Romadhoni and K. F. H. Holle, "Analisis Sentimen Terhadap PERMENDIKBUD No.30 pada Media Sosial Twitter Menggunakan Metode Naive Bayes dan LSTM," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 7, no. 2, pp. 118–124, May 2022, doi: 10.30591/jpit.v7i2.3191.

[14] C. Dewi, R. C. Chen, H. J. Christanto, and F. Cauteruccio, "Multinomial Naïve Bayes Classifier for Sentiment Analysis of Internet Movie Database," *Vietnam Journal of Computer Science*, vol. 10, no. 4, pp. 485–498, Nov. 2023, doi: 10.1142/S2196888823500100.

[15] K. Ainiyah and K. F. H. Holle, "Analisis Sentimen Terhadap Permendikbud Ristek Nomor 30 Tahun 2021 pada Media Sosia Twitter Menggunakan Metode Lexicon-Based dan Multinomial Naïve Bayes," *Jurnal Ilmiah Informatika*, vol. 7, no. 1, pp. 29–40, Jun. 2022, doi: 10.35316/jimi.v7i1.29-40.

[16] A. Sabrani, I. W. Gede Putu Wirarama Wedashwara, and F. Bimantoro, "METODE MULTINOMIAL NAÏVE BAYES UNTUK KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA (Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia)." [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/

[17] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[18] M. Anandarajan, C. Hill, and T. Nolan, "Text Preprocessing," in *Practical Text Analytics, Advances in Analytics and Data Science*, vol. 2, 2019, pp. 45–59. doi: 10.1007/978-3-319-95663-3_4.

[19] N. E. Febriyanty, M. A. Hariyadi, and C. Crysdian, "Hoax Detection News Using Naïve Bayes and Support Vector Machine Algorithm," *International Journal of Advances in Data and Information Systems*, vol. 4, no. 2, pp. 191–200, Oct. 2023, doi: 10.25008/ijadis.v4i2.1306.

[20] A. Prayoga Permana, K. Ainiyah, and K. Fahmi Hayati Holle, "Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up," *JISKa*, vol. 6, no. 3, pp. 178–188, 2021.

[21] G. A. Dalaorao, A. M. Sison, E. Aguinaldo, and R. P. Medina, "Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy," in *2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2019, pp. 282–285. doi: 10.1109/TSSA48701.2019.8985458.

[22] A. Nur Khusna and I. Agustina, "Implementation of Information Retrieval Using Tf-Idf Weighting Method On Detik.Com's Website," in *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018. doi: 10.1109/TSSA.2018.8708744.

[23] M. Hanindia, P. Swari, D. Farrel, P. Rachmawan, and C. A. Putra, "Multinomial Optimization of Naïve Bayes Through the Implementation of Particle Swarm Optimization," *Technium: Romanian Journal of Applied Sciences and Technology*, vol. 16, pp. 169–175, 2023.

[24] A. Sabrani, I. W. Gede Putu Wirarama Wedashwara, and F. Bimantoro, "Metode Multinomial Naive Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia," *Jurnal Teknologi Informasi, Komputer, dan Aplikasinya*, vol. 2, no. 1, pp. 89–100, Mar. 2020, doi: https://doi.org/10.29303/jtika.v2i1.87.

[25] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 820–826, Aug. 2021, doi: 10.29207/resti.v5i4.3146.

[26] W. L. W. Foh, S. L. Ang, C. Y. Lim, A. A. L. Alaga, and G. H. Yeap, "Prediction of Tuberculosis Patients' Treatment Outcomes Using Multinomial Naive Bayes Algorithm and Class-Imbalanced Data," in *2023 IEEE IAS*

Global Conference on Emerging Technologies, GlobConET 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/GlobConET56651.2023.10150132.

[27] L. Mayasari and D. Indarti, "Klasifikasi Topik Tweet Mengenai COVID Menggunakan Metode Multinomial Naive Bayes dengan Pembobotan TF-IDF," *Jurnal Ilmiah Informatika Komputer*, vol. 27, no. 1, pp. 43–53, 2022, doi: 10.35760/ik.2022.v27i1.6184.

[28] F. W. Sembiring, R. A. Yusda, and S. Santoso, "Analysis Naive Bayes to Selection New Students for Superior Class STMIK Royal," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 9, no. 2, pp. 239–248, Mar. 2023, doi: 10.33330/jurteksi.v9i2.2216.

[29] C. Dewi, R.-C. Chen, H. J. Christanto, and F. Cauteruccio, "Multinomial Naïve Bayes Classifier for Sentiment Analysis of Internet Movie Database," *Vietnam Journal of Computer Science*, vol. 10, no. 4, pp. 1–14, Aug. 2023, doi: 10.1142/s2196888823500100.

[30] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *Online News Classification Using Multinomial Naive Bayes*, vol. 6, no. 1, pp. 32–38, 2017.

[31] K. M. Ting, *Confusion Matrix. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning*. Springer US, 2016. doi: 10.1007/978-1-4899-7502-7.