Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 4, August 2024, pp. 1085-1091

DOI: https://doi.org/10.52436/1.jutif.2024.5.4.1908
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON OF NAÏVE BAYES AND INFORMATION GAIN ALGORITHMS IN CYBERBULLYING SENTIMENT ANALYSIS ON TWITTER

Dinda Septia Ningsih[1], Ryan Randy Suryono[*2]

[1,2]Information System, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Indonesia
Email: [1]dinda_septia_ningsih@teknokrat.ac.id, [2]ryan@teknokrat.ac.id

(Article received: February 28, 2024; Revision: March 22, 2024; published: July 29, 2024)

Abstract

In the current digital era, cyberbullying is very easy to do because access to various social media platforms is very easy to obtain. Generation Z is a generation born in the era of digital technology advancement, being one of the parties that plays a role in the increasing cases of cyberbullying. The twitter social media platform is one of the platforms that is often used as a place for cyberbullying in Indonesia. With the alarming impact, this research aims to analyze cyberbullying cases on twitter. By comparing Naïve Bayes and Information Gain algorithms, this research will provide accuracy results from tweet data containing cyberbullying content. The dataset used comes from twitter with the time span of collecting the dataset is from January 05, 2024 to January 25, 2024. The dataset is then processed to produce a clean dataset that is ready to be tested using both algorithms. In this study, testing the two algorithms using the K-fold Cross Validation technique resulted in variations in each test. In testing both algorithms, an accuracy level is obtained that indicates how successful the model is in making predictions. In simple terms, this accuracy assesses how effective the model is in predicting cyberbullying sentiment in datasets from Indonesian twitter. Testing the Naïve Bayes algorithm obtained an accuracy of 92.3%. Testing the Information Gain algorithm has an accuracy of 97.8%. From the results obtained, it can be concluded that the Information Gain algorithm gets higher accuracy than the Naïve Bayes algorithm for cyberbullying sentiment analysis on Indonesian twitter.

Keywords: *Cyberbullying, Generation Z, Information Gain, Naïve Bayes.*

PERBANDINGAN ALGORITMA *NAÏVE BAYES* DAN *INFORMATION GAIN* DALAM ANALISIS SENTIMEN CYBERBULLYING DI TWITTER

Abstrak

Dalam era digital saat ini tindakan *cyberbullying* sangat mudah dilakukan karena akses ke berbagai *platform* media sosial yang sangat mudah didapatkan. Generasi Z ialah generasi yang dilahirkan di era kemajuan teknologi digital, menjadi salah satu pihak yang berperan dalam meningkatnya kasus *cyberbullying*. *Platform* media sosial *twitter* menjadi salah satu paltform yang yang sering digunakan sebagai tempat berlangsungnya tindakan *cyberbullying* di Indonesia. Dengan adanya dampak yang menghawatirkan penelitian ini bertujuan untuk manganalisis kasus *cyberbullying* yang ada di *twitter*. Dengan mambandingkan algoritma *Naïve Bayes* dan *Information Gain* penelitian ini akan memberikan hasil memberikan hasil akurasi dari data *tweet* yang mengandung konten *cyberbullying*. Dataset yang digunakan berasal dari *twitter* dengan rentang waktu pengumpulan dataset adalah dari 05 Januari 2024 sampai dengan 25 Januari 2024. Dataset lalu diproses untuk manghasilkan dataset bersih yang siap di uji menggunakan kedua algoritma. Dalam penelitian ini pengujian kedua algoritma menggunakan teknik *K-fold Cross Validation* yang meghasilkan variasi pada setiap pengujian. Dalam pengujian kedua algoritma, diperoleh tingkat akurasi yang mengindikasikan seberapa berhasilnya model dalam melakukan prediksi. Secara sederhana, akurasi ini menilai seberapa efektif model dalam memprediksi sentimen *cyberbullying* dalam dataset dari *twitter* Indonesia. Pengujian algoritma *Naïve Bayes* diperoleh akurasi sebesar 92,3%. Pengujian algoritma *Information Gain* akurasi sebesar 97,8%. Dari hasil yang diperoleh dapat disimpulkan bahwa algoritma *Information Gain* mendapatkan akurasi yang lebih tinggi dari pada algoritma *Naïve Bayes* untuk analisis sentimen *cyberbullying* di *twitter* Indonesia.

Kata kunci: *Cyberbullying, Generasi Z, Information Gain, Naïve Bayes.*.

1. PENDAHULUAN

Cyberbullying atau perundungan dunia maya ialah intimidasi atau perundungan yang dilakukan dengan menggunakan teknologi digital [1]. Tindakan yang dilakukan yakni menyebarkan pesan yang merendahkan, menghina, atau mengancam orang lain di internet. Dengan perkembangan teknologi yang pesat, serta mudahnya akses ke *platform* media sosial dan anonimitas yang memungkinkan pelaku untuk melakukan tindakan tanpa tertangkap atau diidentifikasi menjadi dampak meningkatnya *cyberbullying* di Indonesia [2].

Dalam era digital saat ini, pelaku pelaku *cyberbullying* dapat dengan mudah menyebarkan pesan negatif secara luas dan dalam waktu singkat tanpa mempertimbangkan dampaknya terhadap korban [3]. Berdasarkan data statistik dari Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menyebutkan jumlah penggunak internet di Indonesai pada tahun 2024 mencapai 221 juta jiwa, Menurut hasil survei penetrasi internet Indonesia tahun 2024 yang dipublikasikan oleh APJII, tingkat penggunaan internet di Indonesia mencapai 79,5%. Pengguna internet sebagian besar adalah Generasi Z (lahir antara 1997-2012), mencapai 34,40%[4]. Dari data tersebut pengguna internet didominasi oleh Generasi Z atau usia produktif. Dampak dari perilaku *online* Generasi Z yang mendominasi, terutama dalam hal potensi *cyberbullying* memicu kekhawatiran di Indonesia. Dengan adanya kekhawatiran terhadap kasus *cyberbullying* di Indonesia, pemerintah membuat kebijakan undang-undang untuk menangani masalah dalam *cyberbullying* yakni, Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik (ITE), menetapkan hukuman pelaku yang melakukan *cyberbullying* [5].

Twitter merupakan salah satu *platform* media sosial yang kerap dimanfaatkan sebagai tempat berlangsungnya tindakan *cyberbullying* di Indonesia [6]. Fitur-fitur *twitter* yang memungkinkan pengguna membuat postingan yang bersifat publik dan menyebar dengan cepat memungkinkan pelaku *cyberbullying* untuk menyebarkan pesan yang merendahkan atau mengancam orang lain secara luas [7]. Dengan penjelasan kasus diatas, peneliti memilih *platform* media sosial *twitter* karena paltform tersebut menyediakan kumpulan data teks sebagai sumber penelitian. Dari data teks dapat dianalisis menggunakan teknik *text mining*.

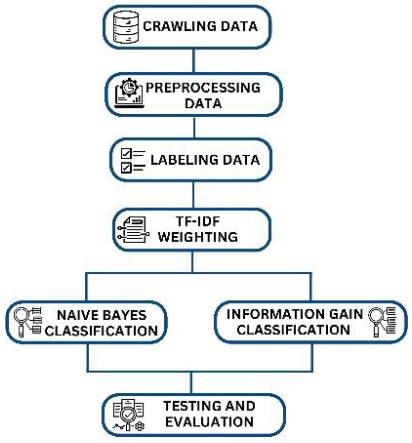
Analisis sentimen merupakan metode pemrosesan bahasa alami *(Natural Language Processing/NLP)* yang digunakan untuk mengidentifikasi dan menilai ekspresi emosional atau sentimen dalam suatu teks [8]. Analisis sentimen merupakan bagian dari *text mining*[9]. *Text mining* ialah teknik yang digunakan untuk mengekstraksi informasi baru dari kumpulan teks yang besar dan tidak terstruktur [10]. Analisis sentimen berguna dalam mengembangkan sistem untuk menganalisis, mengidentifikasi dan juga berpendapat. Proses ini bertujuan untuk mengenali opini atau sentimen dari konten teks mengenai suatu topik atau peristiwa, yang dapat berupa positif, negatif, atau netral [11].

Algoritma klasifikasi juga memiliki kelebihan dan kekurangan dalam mengklasifikasikan teks. algoritma *Naïve Bayes* dan *Information Gain* adalah algoritma yang memiliki kemampuan akurasi dan generalitas klasifikasi yang cukup tinggi[12]. Penggunaan *Naïve Bayes* dan *Information Gain* dalam tugas klasifikasi telah terbukti menghasilkan akurasi dan generalisasi yang tinggi. Pada penelitian Tsania Dzulkarnain, Dian Eka Ratnawati, dan Bayu Rahayudi mengenai penilaian masyarakat terhadap pelayanan rumah sakit di Malang, algoritma *Naïve Bayes Classifier* memperoleh akurasi sebesar 90% [13]. Pada penelitian Ahmad Wildan Attabi' , Lailil Muflikhah , Mochammad Ali Fauzi dengan penggunaan metode *Naïve Bayes Classifier* dan *Information Gain* digunakan untuk menganalisis sentimen guna mengevaluasi produk tertentu di *twitter* berbahasa indonesia memeperoleh akurasi sebesar 70% dan 74%[9].

Penelitian ini bertujuan untuk menganalisis sentimen melalui *twitter* guna memahami peran twitter sebagai platform yang memfasilitasi dan menjadi tempat berlangsungnya tindakan *cyberbullying* di Indonesia. Dengan menggunakan algoritma klasifikasi *Naïve Bayes* dan *Information Gain* penelitian ini akan memberikan hasil memberikan hasil akurasi dari data *tweet* yang mengandung konten *cyberbullying*.

2. METODE PENELITIAN

Tahap Penelitian



Gambar 1. Tahapan Penelitian

Berdasarkan gambar 1.1 metode penelitian ini melibatkan beberapa langkah. Pertama, data dikumpulkan melalui *crawling* data *twitter*. Selanjutnya *preprocessing* data, kemudian *labeling* data, pembobotan TF-Idf, klasifikasi pada kedua algoritma yakni *Naïve Bayes* dan Information Gain, serta pengujian dan evaluasi.

2.1. *Crawling* Data

Pada penelitian ini data dikumpulkan menggunakan metode *crawling* menggunakan *library Harvest*. Metode ini memungkinkan pengumpulan data dari *twitter* dengan mengakses dan mengekstrak teks dari postingan atau komentar. Rentang waktu pengumpulan data adalah dari 05 Januari 2024 sampai dengan 25 Januari 2024.

2.2. *Preprocessing* Data

Pada preprocessing data melibatkan beberapatahap yakni, *Cleansing, Case Folding, Tokenizing, Stopword Removal, Stemming*, dan *Drop Data Duplicate*. Serangkaian tahapan preprocessing dilakukan untuk mengolah data mentah dari *twitter* menjadi data yang siap digunakan.

2.3. TF-IDF *Weighting*

TF-IDF Weighting ialah metode statistik yang umum digunakan untuk menunjukkan seberapa signifikan setiap kata dalam suatu dokumen dibandingkan dengan seluruh koleksi dokumen. Teknik ini dimanfaatkan untuk memberikan nilai penting pada kata-kata dalam suatu dokumen berdasarkan seberapa sering kata tersebut muncul dalam dokumen tersebut.[14]. Berikut rumus *Term-Frequency Inverse Document Frequency* (TF-IDF) :

Persamaan 1

$$IDFt = log\left(\frac{D}{df}\right) \quad (1)$$

Dimana:
IDF : *inversed document frequency*
t : kata ke-*t* dari kata kunci
D : total dokumen
df : banyak dokumen yang mengandung kata dicari

Persamaan 2

$$Wd.t = tf\ d.t\ x\ IDFt \quad (2)$$

Dimana:
d : dokumen ke-*d*
t : kata ke-*t* dari dokumen kunci
W : bobot dokumen ke-d terhadap kata ke-*t*
tf : banyak kata yang dicari pada sebuah dokumen.

2.4. *Naïve Bayes*

Algoritma *Naïve Bayes* adalah teknik klasifikasi yang banyak dipakai dalam analisis sentimen untuk mengidentifikasi sentimen (positif, negatif, atau netral) dari teks [15]. Algoritma ini berasumsi bahwa setiap kata dalam teks bersifat independen terhadap sentimen, meskipun asumsi ini sederhana dan tidak selalu merepresentasikan realitas dengan akurat[16]. Dalam analisis sentimen, *Naïve Bayes* menggunakan probabilitas prior untuk mengestimasi probabilitas bahwa teks tertentu memiliki sentimen tertentu sebelum melihat teks baru. Probabilitas prior ini bergantung pada seberapa sering kata-kata tertentu muncul dalam dataset pelatihan untuk masing-masing sentimen. Selanjutnya, dengan menggunakan *Teorema Bayes*, algoritme *Naïve Bayes* menghitung probabilitas posterior untuk tiap sentimen berdasarkan kata-kata yang muncul dalam teks baru [17]. Sentimen dengan probabilitas posterior tertinggi dianggap sebagai hasil klasifikasi. Berikut persamaan *Teorema Bayes* secara matematis:

$$P(A|B) = \frac{P(B|A)\ x\ P(A)}{P(B)} \quad (3)$$

Di mana:
P(A|B) : adalah probabilitas posterior dari sentimen A (positif, negatif, atau netral) terhadap dokumen B
P(B|A) : adalah probabilitas likelihood dari dokumen B jika sentimennya adalah A
P(A) : adalah probabilitas prior dari sentiment A
P(D\B) : adalah probabilitas dari dokumen B

Asumsi *Naïve Bayes* adalah bahwa setiap kata dalam dokumen B bersifat independen terhadap sentimen, sehingga probabilitas *likelihood* P(A|B) dapat dipecah menjadi perkalian probabilitas kata-kata dalam dokumen B jika sentimennya adalah A.

2.5. *Information Gain*

Information Gain merupakan teknik untuk mengevaluasi pentingnya fitur dalam klasifikasi data. Ailgoritma ini menggunakan konsep *entropi* untuk menentukan seberapa baik sebuah fitur dapat memisahkan kelas-kelas yang berbeda dalam dataset [18]. Fitur *Decision Tree* atau pohon keputusan dalam *Information Gain* dapat mengidentifikasi aspek aspek yang penting dalam menentukan sentiment suatu teks. Untuk membagi data menjadi subset yang lebih *homogen*, *classifier* secara otomatis memilih fitur yang memberikan *Information Gain* tertinggi saat membangun pohon keputusan (decision tree). Berikut rumus perhitungan *Information Gain*:

$$Informatoin\ Gain = Entropy\ (S) - \sum_{v \in Values(A)} \frac{S_v}{S}\ x\ Entropy(S_v) \quad (4)$$

Di mana:
S : Kumpulan data sebelum pemisahan
A : Fitur yang sedang di evaluasi
Values(A) : Nilai yang mungkin dari fitur *A*
|*S*| : Jumlah total *instance* dalam kumpulan data *S*
$|S_v|$: Jumlah instance dalam *S* yang memilioki nilai *v* untuk fitur *A*
Entropy (S): Entropi k umpulan data S, yang dihitung sebagai:

$$Entropy(S) = -\sum_{c \in Classes} p(c)\ x\ log_2\big(p(c)\big) \quad (5)$$

Dimana $p(c)$ adalah *instance* dalam *S* yang termasuk dalam *c*.

3. HASIL DAN PEMBAHASAN

3.1. *Crawling* Data

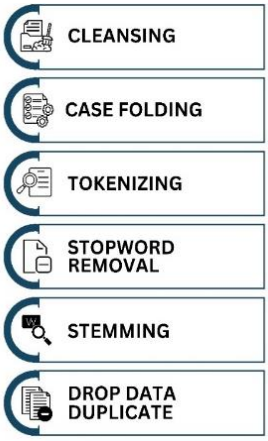
Data yang berhasil dikumpulkan berjumlah 6.014 komentar *twitter*. Pengumpulan data dilakukan pada rentang waktu 05 Januari 2024 sampai dengan 25 Januari 2024. contoh hasil *crawling* data dapat dilihat pada tabel 1.

Table 1. Hasil *Crawling*

Username	Tweet
andit002	Siapa sih yg nyaranin debat capres? Kenapa ga dibikin kaya ajang miss indonesia aja ya, kasih pertanyaan random hasil kocokan. Kalo debat kaya gini masyarakat diajarin menjatuhkan org untuk meninggikan diri, padahal banyak quotes yg bilang naiklah tanpa menjatuhkanðŸ˜®â€• ðŸ˙¨
reyanaaaya	@gibran_gen @prabowo @gibran_tweet Udah ready bgt nih mau liat debat capres kedua hari minggu nanti. Yakin banget pak prabowo bisa kuasiain materi yg ada

3.2. *Preprocessing* Data

Tahapan *Preprocessing* Data



Gambar 2. Tahapan *Preprocessing* Data

Pada tahap ini data komentar yang diperoleh dari crawling media sosial *twitter* dalam bentuk kalimat yang tidak terstruktur, tidak dapat digunakan. Oleh karena itu, data harus diubah menjadi data terstruktur yang memiliki nilai numerik. Ada beberapa tahapan pada tahap ini antara lain:

3.2.1. *Cleansing*

Pada tahap ini dilakukan penghapusan karakter-karakter yang tidak relevan seperti tanda baca, simbol, dan karakter khusus lainnya. Hasil *cleansing* dapat dilihat pada tabel 2.

Tabel 2. Hasil *Cleansing*

Username	Cleansing
andit002	Siapa sih yg nyaranin debat capres Kenapa ga dibikin kaya ajang miss indonesia aja ya kasih pertanyaan random hasil kocokan Kalo debat kaya gini masyarakat diajarin menjatuhkan org untuk meninggikan diri padahal banyak quotes yg bilang naiklah tanpa menjatuhkan
reyanaaaya	Udah ready bgt nih mau liat debat capres kedua hari minggu nanti Yakin banget pak prabowo bisa kuasiain materi yg ada

3.2.2. *Case Folding*

Pada tahap ini, teks akan diubah menjadi format di mana semua hurufnya berubah menjadi huruf kecil [19]. Proses ini disebut dengan "*lowercasing*." *Lowercasing* berguna untuk memastikan konsistensi dalam pemrosesan teks, terutama dalam analisis teks dan pemodelan bahasa. Hasil *Case Folding* dapat dilihat pada tabel 3.

Tabel 3. Hasil *Case Folding*

Username	Case Folding
andit002	siapa sih yg nyaranin debat capres kenapa ga dibikin kaya ajang miss indonesia aja ya kasih pertanyaan random hasil kocokan kalo debat kaya gini masyarakat diajarin menjatuhkan org untuk meninggikan diri padahal banyak quotes yg bilang naiklah tanpa menjatuhkan
reyanaaaya	udah ready bgt nih mau liat debat capres kedua hari minggu nanti yakin banget pak prabowo bisa kuasiain materi yg ada

3.2.3. *Tokenizing*

Tahapan *tokenizing* dilakukan pemecahan teks menjadi token atau unit-unit yang lebih kecil, seperti kata atau frasa [20]. Hasil *tokenizing* dapat dilihat pada tabel 4.

Tabel 4. Hasil *Tokenizing*

Username	***Tokenizing***
andit002	['siapa', 'sih', 'yg', 'nyaranin', 'debat', 'capres', 'kenapa', 'ga', 'dibikin', 'kaya', 'ajang', 'miss', 'indonesia', 'aja', 'ya', 'kasih', 'pertanyaan', 'random', 'hasil', 'kocokan', 'kalo', 'debat', 'kaya', 'gini', 'masyarakat', 'diajarin', 'menjatuhkan', 'org', 'untuk', 'meninggikan', 'diri', 'padahal', 'banyak', 'quotes', 'yg', 'bilang', 'naiklah', 'tanpa', 'menjatuhkan']
reyanaaaya	['udah', 'ready', 'bgt', 'nih', 'mau', 'liat', 'debat', 'capres', 'kedua', 'hari', 'minggu', 'nanti', 'yakin', 'banget', 'pak', 'prabowo', 'bisa', 'kuasiain', 'materi', 'yg', 'ada']

3.2.4. *Stopword Removal*

Stopword Removal ialah tahap penghapusan kata-kata pengisi (*stop words*) dari teks yang akan dianalisis. *Stop words* adalah kata-kata umum yang kerap muncul dalam teks namun tidak memiliki dampak signifikan terhadap makna atau sentimen dalam analisis yang dilakukan[21]. Hasil *stopword removal* dapat dilihat pada tabel 5.

Tabel 5. Hasil *Stopword Removal*

Username	***Stopword Removal***
andit002	['siapa', 'sih', 'nyaranin', 'debat', 'capres', 'ga', 'dibikin', 'kaya', 'ajang', 'miss', 'indonesia', 'aja', 'kasih', 'pertanyaan', 'random', 'hasil', 'kocokan', 'kalo', 'debat', 'kaya', 'gini', 'masyarakat', 'diajarin', 'menjatuhkan', 'org', 'meninggikan', 'diri', 'padahal', 'banyak', 'quotes', 'bilang', 'naiklah', 'menjatuhkan']
reyanaaaya	['udah', 'ready', 'bgt', 'nih', 'mau', 'liat', 'debat', 'capres', 'kedua', 'hari', 'minggu', 'yakin', 'banget', 'pak', 'prabowo', 'kuasiain', 'materi']

3.2.5. *Stemming*

Pada tahapan ini dilakukan pengubahan kata-kata dalam teks menjadi bentuk dasar atau akar kata, dengan cara menghapus imbuhan atau awalan yang tidak penting. Hasil *stemming* dapat dilihat pada tabel 6.

Tabel 6. Hasil *Stemming*

Username	*Stemming*
andit002	siapa sih nyaranin debat capres ga bikin kaya ajang miss indonesia aja kasih tanya random hasil kocok kalo debat kaya gin masyarakat diajarin jatuh org tinggi diri padahal banyak quotes bilang naik jatuh
reyanaaaya	udah ready bgt nih mau liat debat capres dua hari minggu yakin banget pak prabowo kuasiain materi

3.2.6. *Drop Data Duplicate*

Pada tahapan ini dilakukan penghapusan data yang duplikat atau identik dari dataset yang digunakan untuk analisis sehingga membantu keakuratan dengan memastikan bahwa setiap baris data unik dan kontribusi informasi yang berbeda untuk analisis. Setelah dilakukan *Drop data Duplicate* dataset berubah menjadi 4.531.

3.3. *Labeling* Data

Setelah tahap preprocessing dilakukan tahap selanjutnya adalah *labeling* data. Tahap ini dilakukan untuk memberikan labeling setiap isi komentar yang diposting oleh pengguna *twitter* berdasarkan polaritas sentimennya. Label ini menunjukkan apakah komentar tersebut positif, negatif, atau netral. Hasil *labeling* dapat dilihat pada tabel 7.

Tabel 7. Hasil *Labeling*

Username	*Stemming*	Label
andit002	siapa sih nyaranin debat capres ga bikin kaya ajang miss indonesia aja kasih tanya random hasil kocok kalo debat kaya gin masyarakat diajarin jatuh org tinggi diri padahal banyak quotes bilang naik jatuh	Negatif
reyanaaaya	udah ready bgt nih mau liat debat capres dua hari minggu yakin banget pak prabowo kuasiain materi	Positif

3.4. Pengujian dan Evaluasi

K-Fold Cross Validation merupakan metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan melakukan pembagian data menjadi dua bagian yakni data *training* dan data *testing*. Bagian pertama digunakan untuk mempelajari atau melatih data (*training*) dan bagian kedua digunakan untuk mengevaluasi model (*testing*). Data dibagi dengan rasio 80% data *training* dan 20% data *testing*. Pada metode ini bagian pelatihan dan validasi harus bertindak sebagai subset berurutan, memastikan bahwa setiap titik data diuji setidaknya satu kali.

3.4.1. Pengujian pengaruh *Naïve Bayes* pada *K-Fold Cross Validation*

Pada pengujian dengan menggunakan algoritma *MultinomialNB* data akan dibagi menjadi 5 fold untuk memperoleh *accuracy, precision, recall* dan *f1-score* yang dimana akan dilakukan 5 kali proses dengan tujuan mencari hasil rata- rata pengujian yang tertinggi. Hasil pengujian dapat dilihat pada tabel 8.

Tabel 8. Hasil pengujian *Naïve Bayes*

k	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1	0,923	0,933	0,926	0,920
2	0,915	0,926	0,916	0,913
3	0,914	0,926	0,913	0,910
4	0,915	0,923	0,916	0,913
5	0,910	0,920	0,913	0,910

Berdasarkan hasil pengujian menggunakan algoritma *MultinomialNB* rata-rata nilai tertinggi pada pengujian fold ke-1 dengan akurasi sebesar 0,923 atau 92,3%.

3.4.2. Pengujian pengaruh *Information Gain* pada *K-Fold Cross Validation*

Pada pengujian dengan menggunakan algoritma *DecisionTree Classifier* data akan dibagi menjadi 5 fold untuk memperoleh *accuracy, precision, recall dan f1-score* yang dimana akan dilakukan 5 kali proses dengan tujuan mencari hasil rata- rata pengujian yang tertinggi. Hasil pengujian dapat dilihat pada tabel 9.

Tabel 9. Hasil pengujian *Information Gain*

k	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1	0,978	0,980	0,980	0,980
2	0,976	0,976	0,976	0,973
3	0,973	0,976	0,973	0,973
4	0,976	0,980	0,976	0,976
5	0,977	0,980	0,976	0,973

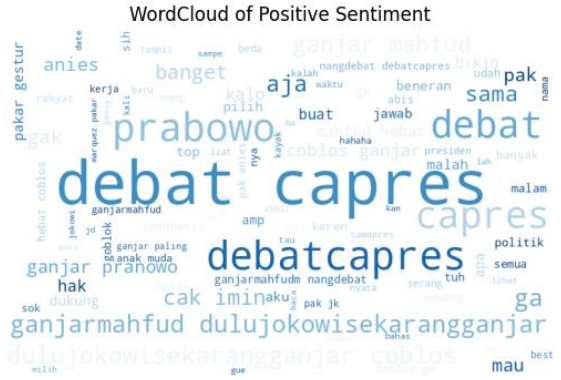
Berdasarkan hasil pengujian menggunakan algoritma *DecisionTree Classifier* rata-rata nilai tertinggi pada pengujian fold ke-1 dengan akurasi sebesar 0,978 atau 97,8%.

3.5. Visualisasi Data

Penggunaan *Wordcloud* bertujuan untuk menampilkan kata kata yang sering muncul dalam dataset yang dianalisis [22]. *Wordcloud* juga memberikan gambaran yang cepat dan mudah dipahami tentang tema atau topik yang dominan dalam penelitian. Proses pembuatan *WordCloud* dilakukan dengan menggunakan library matplotlib dalam bahasa pemrograman Python. Pada gambar 3 menunjukkan *wordcloud* dengan sentimen negatif atau akata kata yang mangandung unsur *cyberbullying*.

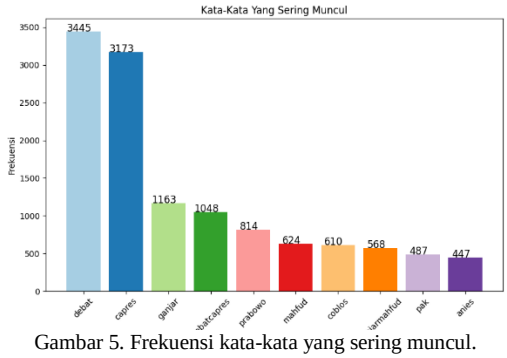
Gambar 3. Hasil *Wordcloud* dari sentimen negatif

Pada gambar 3 menunjukkan *wordcloud* dengan sentimen positif atau kata - kata yang tidak mangandung unsur *cyberbullying*.



Gambar 4. Hasil *Wordcloud* dari sentimen positif

Pada gambar 5 menunjukkan kata yang paling banyak muncul dari dataset yang digunakan.



Gambar 5. Frekuensi kata-kata yang sering muncul.

4. DISKUSI

Penelitian ini bertujuan untuk membandingkan hasil pengujian antara dua algoritma yakni *Naïve Bayes* dan *Information Gain* terkait *tweet* yang berisi *cyberbullying* di *twitter*. Pengujian pada kedua algoritma menggunakan algoritma *k-fold Cross Validation* yang hasil dari setiap pengujian berbeda beda. Pada pengujian algoritma *Naïve Bayes* didaptkan rata-rata tertinggi dengan akurasi 92,3% untuk pengujian fold ke 1. Pada pengujian algoritma *Information Gain* didaptkan rata-rata tertinggi dengan akurasi sebesar 97,8% untuk pengujian *fold* ke 1.

Dari penelitian Mochammad Ali Fauzi dkk terlihat bahwa hasil yang didapatkan berbeda dari penelitian ini. Penelitian ini juga menggunakan jumlah data yang lebih banyak.

5. KESIMPULAN

Berdasarkan hasil penelitian tentang analisis cyberbullying di media sosial *twitter* dan dengan membandingkan algoritma *Naïve Bayes* dan *Information Gain* maka didapatkan kedua hasil pengujian akurasi yakni 92,3% dan 97,8%. dari hasil ini dapat disimpulkan bahwa algoritma *Information Gain* mendapatkan akurasi yang lebih tinggi daripada algoritma *Naïve Bayes* . Dengan menggunakan *k-fold Cross Validation* dilakukan sebanyak 5 kali uji data untuk manghasilkan akurasi yang terbaik.

DAFTAR PUSTAKA

[1] F. N. Eleanora And R. Al Adawiah, "Perundungan Dunia Maya (Cyberbullying) Dan Upaya Preventif Di Kalangan Siswa Smk Bangun Persada Bekasi," *Jurnal Abdi Masyarakat Indonesia*, Vol. 1, No. 2, Pp. 203–208, Oct. 2021, Doi: 10.54082/Jamsi.67.

[2] D. Riswanto And R. Marsinun, "Perilaku Cyberbullying Remaja Di Media Sosial," *Analitika*, Vol. 12, No. 2, Pp. 98–111, Dec. 2020, Doi: 10.31289/Analitika.V12i2.3704.

[3] J. Pengabdian Kepada Masyarakat Hal And I. Fardian Anshori, "Jurnal Sosial & Abdimas Fenomena Cyber Bullying Dalam Kehidupan Remaja", [Online]. Available: Http://Ejurnal.Ars.Ac.Id/Index.Php/Jsa

[4] "Apjii Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang," Https://Apjii.Or.Id/Berita/D/Apjii-Jumlah-Pengguna-Internet-Indonesia-Tembus-221-Juta-Orang.

[5] D. Persetujuan Bersama, "Dewan Perwakilan Rakyat Republik Indonesia Dan Presiden Republik Indonesia." Accessed: Feb. 09, 2024. [Online]. Available: Https://Peraturan.Bpk.Go.Id/Details/37582/Uu-No-19-Tahun-2016

[6] M. I. Djamzuri And A. Putra Mulyana, "Fenomena Cyberbullying Pembiaran Juvenile Deliquency Dalam Teknologi Media Baru," *Jurnal Ilmu Sosial Dan Pendidikan (Jisip)*, Vol. 7, No. 1, Pp. 2598–9944, 2023, Doi: 10.58258/Jisip.V7i1.4801/Http.

[7] N. Made Gita Dwi Purnamasari, M. Ali Fauzi, And L. Shinta Dewi, "Identifikasi Tweet Cyberbullying Pada Aplikasi Twitter Menggunakan Metode Support Vector Machine (Svm) Dan Information Gain (Ig) Sebagai Seleksi Fitur," 2018. [Online]. Available: Http://J-Ptiik.Ub.Ac.Id

[8] Y. Guo, S. Das, S. Lakamana, And A. Sarker, "An Aspect-Level Sentiment Analysis Dataset For Therapies On Twitter," *Data Brief*, Vol. 50, Oct. 2023, Doi: 10.1016/J.Dib.2023.109618.

[9] A. Wildan Attabi', L. Muflikhah, And M. A. Fauzi, "Penerapan Analisis Sentimen Untuk Menilai Suatu Produk Pada Twitter Berbahasa Indonesia Dengan Metode Naïve Bayes Classifier Dan Information Gain," 2018. [Online]. Available: Http://J-Ptiik.Ub.Ac.Id

[10] S.-J. Son, M. S. Do, G. Choi, And H.-K. Nam, "Identifying Research Trends In Avian Migration Tracking In Korea Using Text Mining," *J Asia Pac Biodivers*, Dec. 2023, Doi: 10.1016/J.Japb.2023.12.001.

[11] M. Qamal And W. Fuadi, "Analisis Sentimen Toko Online Menggunakan Algoritma Naive Bayes Classifier."

[12] J. Elektronik *Et Al.*, "Implementasi Algoritma Naive Bayes Classifier (Nbc) Dan Information Gain Untuk Mendeteksi Ddos," 2019, [Online]. Available: Https://Research.Unsw.Edu.Au/Projects/Unsw-Nb15-Dataset.

[13] T. Dzulkarnain, D. E. Ratnawati, B. Rahayudi, And P. Korespondensi, "Penggunaan Metode Naïve Bayes Classifier Pada Analisis Sentimen Penilaian Masyarakat Terhadap Pelayanan Rumah Sakit Di Malang The Use Of The Naïve Bayes Classifier Method In Sentiment Analysis Of The Community's Assessment Of Hospital Services In Malang," Vol. 10, No. 7, 2023, Doi: 10.25126/Jtiik.2023107979.

[14] A. Nursalim And R. Novita, "Sentiment Analysis Of Comments On Google Play Store, Twitter And Youtube To The Mypertamina Application With Support Vector Machine," *Jurnal Teknik Informatika (Jutif)*, Vol. 4, No. 6, Pp. 1305–1312, 2023, Doi: 10.52436/1.Jutif.2023.4.6.1059.

[15] M. Yunus, M. Husni, And M. M. Mufadhdhal, "Klasifikasi Sentimen Terhadap Badan Penyelenggara Jaminan Sosial (Bpjs) Pada Media Sosial Twitter Menggunakan Naive Bayes," *Smatika Jurnal*, Vol. 11, No. 02, Pp. 81–91, Dec. 2021, Doi: 10.32664/Smatika.V11i02.577.

[16] P. Sofyan Zakaria, R. Julianto, And R. Surya Bernada, "Implementasi Naive Bayes Menggunakan Python Dalam Klasifikasi Data." [Online]. Available: Https://Jurnalmahasiswa.Com/Index.Php/Biikma

[17] R. Fajar, S. Program, P. Rekayasa, N. Lunak, And R. Bengkalis, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," Vol. 3, No. 1.

[18] C. Destitus, "Support Vector Machine Vs Information Gain: Analisis Sentimen Cyberbullying Di Twitter Indonesia," *Ultima Infosys*, Vol. Xi, No. 2, P. 107, 2020.

[19] D. Darwis, E. Shintya Pratiwi, A. Ferico, And O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," 2020.

[20] P. Kumala Sari And R. Randy Suryono, "Komparasi Algoritma Support Vector Machine Dan Random Forest Untuk Analisis Sentimen Metaverse," 2024.

[21] I. Nur Fakhri And R. Febrian Umbara, "Analisis Sentimen Pada Kuisioner Kepuasan Terhadap Layanan Dan Fasilitas Kampus Universitas Dengan Menggunakan Klasifikasi Support Vector Machine (Svm)".

[22] B. Indra Kusuma And A. Nugroho, "Cyberbullying Detection On Twitter Uses The Support Vector Machine Method," *Jurnal Teknik Informatika (Jutif)*, Vol. 5, No. 1, Pp. 11–17, 2024, Doi: 10.52436/1.Jutif.2024.5.1.809.