Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 1, No. 1, Januari 2017, hlm. x-x http://j-ptiik.ub.ac.id

Analisis Sentimen Opini Masyarakat Terhadap Fenomena TikTokShop di Indonesia Menggunakan Metode *K-Nearest Neighbor* berbasis *N-gram* dengan Seleksi Fitur *Information Gain*

Zianka Mahendra[1], Indriati[2], Achmad Ridok[2]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]Zianmahendra@student.ub.ac.id, [2]tif.indriati@ub.ac.id, [3]acridokb@ub.ac.id

Abstrak

TikTokShop merupakan fitur terbaru yang diperkenalkan dalam platform TikTok. Di tengah popularitasnya yang sedang melonjak pemerintah Indonesia secara mendadak mengambil keputusan untuk menutup akses ke fitur ini. Keputusan ini telah menyebabkan masyarakat memiliki pandangan tersendiri terhadap suatu kebijakan pemerintah baik itu positif (mendukung) ataupun negatif (menyangkal). Opini masyarakat terhadap TikTokShop tersebar luas di media sosial, termasuk dalam kolom komentar pada platform youtube yang sangat masif diperbincangkan. Analisis Sentimen menjadi kunci untuk memahami pandangan mendalam masyarakat terhadap kebijakan ini. Analisis Sentimen pada penelitian ini menggunakan kombinasi metode *K-Nearest Neighbors (KNN)* berbasis N-Gram dan *Information Gain* sebagai seleksi fitur. Fitur *N-Gram* yang digunakan dalam penelitian ini adalah fitur *Unigram*, *Bigram* dan Gabungan *Unigram-Bigram*. Berdasarkan pengujian yang telah dilakukan, didapatkan hasil bahwa nilai terbaik terdapat pada fitur *Unigram* dan nilai *threshold* yang digunakan adalah 100%, menghasilkan akurasi sebesar 89%, dengan recall 89%, *Precision* sebesar 89.00%, dan *F-Measure* sebesar 89.00%. Berdasarkan temuan tersebut, dapat disimpulkan bahwa dalam menganalisis sentimen opini masyarakat Indonesia terhadap TikTokShop, metode *K-Nearest Neighbor (KNN)* dengan fitur *Unigram* dan tanpa seleksi fitur *Information Gain* memberikan hasil terbaik.
Kata kunci: Analisis Sentimen, *K-Nearest Neighbor, N-Gram, Information Gain,* TikTokShop

Abstract

TikTokShop is the latest feature introduced on the TikTok platform. Amidst its rising popularity, the Indonesian government suddenly decided to close access to this feature. This decision has caused the public to form their own opinions about the government's policy, whether positive (supportive) or negative (opposed). Public opinions on TikTokShop have spread widely on social media, including in the comment sections on YouTube, where it is extensively discussed. Sentiment analysis becomes key to understanding the public's deep-seated views on this policy. Sentiment analysis in this study utilizes a combination of the K-Nearest Neighbors (KNN) method based on N-Gram and Information Gain as feature selection. The N-Gram features used in this study include Unigram, Bigram, and a combination of Unigram-Bigram. Based on the testing conducted, it was found that the best results were obtained with the Unigram feature and a threshold value of 100%, resulting in an accuracy of 89%, with a recall of 89%, precision of 89.00%, and an F-Measure of 89.00%. Based on these findings, it can be concluded that in analyzing the sentiment of Indonesian society towards TikTokShop, the K-Nearest Neighbor (KNN) method with Unigram features and without Information Gain feature selection yields the best results.
***Keywords**: Sentiment Analysis, K-Nearest Neighbor, N-Gram, Information Gain, TikTokShop*

1. PENDAHULUAN

Dalam era digital ini, fenomena media sosial menjadi salah satu aspek yang sangat mempengaruhi pola perilaku masyarakat. Salah satu platform yang memperoleh popularitas yang signifikan adalah TikTok. Menurut data terbaru yang diterbitkan oleh Goostats, Indonesia berada di posisi kedua dalam hal jumlah pengguna Tiktok tertinggi pada tahun 2023, dengan jumlah pengguna sebanyak 112 juta akun pengguna yang terdaftar. Karena daya tariknya yang luas dan tingkat minat publik yang banyak ini, Dalam beberapa tahun terakhir, TikTok telah membuat kemajuan signifikan dalam fungsinya dengan menambahkan fitur baru yang dikenal sebagai

TikTokShop. TikTokShop resmi diluncurkan pada April 2022, sejak diluncurkan TiktokShop mendapatkan popularitas yang besar di kalangan *e-commerce* dan penggemar belanja *online* karena antarmuka yang menarik dan mudah untuk memasarkan produk dan pembelian produk (Rosiyana et al., 2021).

Fitur ini dengan cepat mencapai tingkat popularitas yang sangat tinggi di kalangan pengguna TikTok. Di tengah popularitasnya yang sedang melonjak pemerintah Indonesia tiba-tiba mengambil kebijakan untuk menutup akses ke fitur ini. Kebijakan ini telah menimbulkan berbagai ragam pandangan di kalangan masyarakat Indonesia. Beberapa mendukung keputusan pemerintah, merasa bahwa ini adalah tindakan yang perlu dilakukan demi kepentingan publik. Namun, tidak sedikit pula yang menentang, beranggapan bahwa penutupan TikTokShop adalah bentuk pembatasan yang tidak perlu dan bahkan bisa jadi merugikan. Opini masyarakat terhadap TikTokShop tersebar luas di media sosial, termasuk dalam kolom komentar pada platform youtube yang sangat masif diperbincangkan. Itulah sebabnya, penting untuk dilakukan analisis sentimen opini masyarakat terhadap fenomena Tiktokshop guna memahami berbagai sudut pandang yang ada, dan dapat digunakan sebagai salah satu variabel evaluasi pemerintah untuk mengubah atau melanjutkan kebijakan publik yang telah diimplementasikan oleh pemerintah. Analisis sentimen merupakan sebuah ranah penelitian yang mengkaji pendapat, evaluasi, penilaian, sikap, dan emosi individu terhadap entitas tertentu seperti produk, topik, masalah, peristiwa, organisasi, layanan, dan aspek lainnya. (Liu, 2012). Analisis sentimen dapat mengelompokkan polaritas dari teks dalam suatu kalimat untuk mengetahui opini dari suatu kalimat apakah termasuk positif atau negatif (Nafan & Amalia, 2019).

Beberapa metode klasifikasi dapat digunakan dalam pemeriksaan proses analisis sentimen untuk data tekstual. Salah satu pendekatan yang dapat digunakan untuk analisis sentimen adalah algoritma *K-Nearest Neighbor* (KNN). K-Nearest Neighbor merupakan algoritma yang menggunakan jarak untuk mengidentifikasi data yang memiliki kesamaan fitur dengan data lainnya, sehingga data tersebut dapat diklasifikasikan dan ditentukan kelasnya berdasarkan kelas mayoritas dari data tetangga terdekat (Cunningham & Delany, 2021). Penelitian sebelumnya dilakukan oleh (Huq et al., 2017), yang melakukan perbandingan antara metode *K-Nearest Neighbor* dengan *Support Vector Machine* dengan menggunakan *Grid Search* untuk menemukan nilai terbaik dari *hyperparameter c* dan *gamma*, *dataset* yang digunakan diperoleh dari Twitter sebanyak 1000 tweets, Kemudian, hasil pengujian menunjukkan akurasi sebesar 80,60%, Presisi sebesar 85%, Recall sebesar 75%, dan F-Score sebesar 79% untuk algoritma *K-Nearest Neighbor*, sementara untuk algoritma *Support Vector Machine*, ditemukan akurasi sebesar 58,79%, Presisi 56%, Recall 69%, dan F-Score 61%. Terlihat bahwa algoritma *K-Nearest Neighbor* memiliki hasil yang lebih unggul dari hasil yang dihasilkan Support Vector Machine dengan perlakuan dan menggunakan jumlah data yang sama.

Dalam penelitian ini, akan diterapkan *Cosine Similarity* untuk mengukur jarak antara tetangga pada pendekatan klasifikasi *K-Nearest Neighbo*r dengan menggunakan metode *bag of words*. Penelitian sebelumnya yang dilakukan oleh (Fauzi et al., 2017), menyatakan bahwa terdapat masalah pada fitur *bag of word* yang tidak memperhatikan urutan kata dalam sebuah kalimat. Dua kalimat yang berbeda dalam sebuah komposisi yang sama akan dianggap mirip dan hal ini dapat mempengaruhi akurasi yang dihasilkan. Untuk menyempurnakan kekurangan yang disebutkan di atas maka penelitian ini akan menggunakan metode pendukung yang digunakan sebagai ekstraksi fitur yaitu *N-Gram*. Pada penelitian yang dilakukan oleh (Saputri et al., 2019) juga disebutkan bahwa penerapan *N-Gram* dapat menghasilkan informasi yang lebih beragam dan meningkatkan kinerja klasifikasi. Seperti pada penelitian yang dilakukan oleh (Pramono et al., 2019) mendapatkan peningkatan akurasi sebesar 6,7%.

Selain menggunakan ekstraksi fitur N-Gram, adapun pada penelitian ini digunakan juga metode seleksi fitur *Information Gain* yang digunakan untuk menghilangkan atau mengurangi fitur yang dianggap kurang relevan pada saat klasifikasi. *Information Gain* dipilih menjadi seleksi fitur pada penelitian ini karena penggunaan *Information Gain* sebagai seleksi fitur sering lebih unggul dibandingkan *Document Frequency*, *Mutual Information* dan *Chi-Square* (Muthia, 2016). Seperti pada penelitian yang telah dilakukan oleh (Jodha et al., 2018), dengan menggunakan 100 fitur teks yang dilakukan proses *stemming* terlebih dahulu lalu diperoleh bahwa penggunaan *Information*

Gain memperoleh nilai akurasi sebesar 74,47% sedangkan pada seleksi fitur *Chi-Square* diperoleh sebesar 70,11% dan pada *Mutual Information* diperoleh sebesar 38,22%.

Berdasarkan pemetaan masalah yang telah diuraikan serta studi litelatur yang telah dilakukan, penelitian ini bertujuan untuk menganalisis sentimen opini masyarakat terhadap fenomena TikTokShop. Sistem analisis sentimen akan dikembangkan menggunakan pendekatan klasifikasi *K-Nearest Neighbor* karena pendekatan ini terbukti memberikan kinerja yang unggul dibandingkan pendekatan metode lainnya dan mudah diimplementasikan dalam berbagai dataset. Selain itu, pendekatan ini akan menggunakan *N-Gram* sebagai metode ekstraksi fitur guna menghasilkan informasi yang lebih variatif dan meningkatkan kinerja klasifikasi serta metode seleksi fitur *Information Gain* juga akan diterapkan untuk mengurangi atau menghilangkan fitur yang dianggap tidak relevan dalam proses klasifikasi.

1.1 Text Preprocessing

Preprocessing text merujuk pada proses pengubahan bentuk teks yang tidak terstruktur sebelumnya menjadi bentuk teks yang terstruktur sesuai dengan kebutuhan untuk proses analisis yang lebih lanjut, seperti analisis sentimen, peringkasan teks, klasifikasi, clustering, topic modeling sebagainya.

Tujuan dari preprocessing adalah untuk membersihkan data teks dari karakter-karakter yang tidak relevan seperti tanda baca, menghapus kata-kata yang tidak relevan, mengubah huruf menjadi huruf kecil dan lain sebagainya. *Preprocessing* perlu dilakukan karena data teks sering kali mengandung noise yang dapat menurunkan performa, oleh karena itu, dengan melakukan preprocessing dapat membersihkan data dari elemen tersebut dan dapat mempercepat proses klasifikasi (Chuzaimah Zulkifli, 2018). Terdapat beberapa tahapan dalam *Preprocessing text* ini, antara lain:

1. *Data Cleaning Data Cleaning* adalah tahap awal *Preprocessing* teks yang bertujuan menghilangkan noise pada data. Tahapan ini melibatkan penghapusan karakter selain huruf, seperti simbol, tanda baca dan angka (Sabily et al., 2019).
2. *Case Folding* adalah proses perubahan huruf awal pada suatu kata dalam dokumen menjadi huruf kecil (*lower case*).
3. *Tokenizing* tahapan pemotongan kalimat berdasarkan tiap kata yang menyusunnya dari dokumen menjadi kata atau *token* (Onantya et al., 2019).
4. *Stopword removal* adalah tahapan mengambil kata-kata penting dari hasil *tokenization* dan menghapus kata yang tidak memiliki arti (Onantya et al., 2019).
5. *Stemming* adalah tahapan proses untuk mencari *root* kata (kata dasar) dari setiap kata yang harus sesuai dengan struktur morfologi bahasa Indonesia yang benar (Indriati & Ridok, 2016).

1.2 N-Gram

N-Gram didefinisikan sebagai urutan yang berdekatan yang terdiri dari *n* karakter dari sebuah *string* yang harus berurutan dalam rangkaian teks atau kata yang diberikan (Windisch & Csink, 2005). Contoh pada kata "TEKS".

Unigram: T, E, K, S
Bigram: TE, EK, KS
Gabungan: T, E, K, S, TE, EK, KS

1.3 Information Gain

Information Gain adalah suatu metode seleksi fitur pada *Machine Learning* yang paling umum digunakan untuk melakukan perangkingan atribut dalam aplikasi kategorisasi teks, analisis data, dan analisis data citra (Chormunge & Jena, 2016). *Information Gain* juga sering digunakan untuk mengurangi dimensi fitur pada suatu data atau mengurangi *noise* yang ditimbulkan oleh fitur-fitur yang tidak relevan. *Information Gain* mencari fitur-fitur yang memberikan informasi tertinggi terkait suatu kelas tertentu.

Penerapan seleksi fitur *Information Gain* akan efektif untuk menghapus nilai-nilai yang lebih kecil daripada nilai *threshold* yang telah ditentukan sebelumnya (Wang et al., 2006). *Information Gain* dapat dituliskan secara matematis melalui persamaan 2.1 sebagai berikut (Uğuz, 2011):

$$IG(t) = -\sum_{i=1}^{c} P(C_i) + P(t)\sum_{i=1}^{c} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{i=1}^{c} P(C_i|\bar{t}) log P(C_i|\bar{t}) \quad (1)$$

keterangan:

- c adalah representasi untuk kategori kelas pada dokumen.

- i adalah indeks atau urutan pada dokumen.
- $P(c_i)$ adalah representasi dari probabilitas untuk kategori ke-i.
- $P(t)$ adalah representasi untuk probabilitas term t muncul pada dokumen.
- $P(\overline{t})$ adalah representasi untuk probabilitas term t tidak muncul pada dokumen.
- $P(c_i|t)$ adalah representasi untuk kondisi probabilitas bersyarat di mana term t muncul dalam kategori.
- $P(c_i|\bar{t})$ dalah representasi untuk kondisi probabilitas bersyarat di mana term t tidak muncul dalam kategori.

1.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah salah satu metode yang efektif dan sederhana yang dapat digunakan secara baik untuk mengklasifikasi data berupa teks atau data tekstual (Indriati et al., 2021). K-Nearest Neighbor merupakan algoritma yang menggunakan jarak untuk mengidentifikasi data yang memiliki kesamaan fitur dengan data lainnya, sehingga data tersebut dapat diklasifikasikan dan ditentukan kelasnya berdasarkan kelas mayoritas dari data tetangga terdekat (Cunningham & Delany, 2021). Jumlah data tetangga ini, dilambangkan sebagai *k*, digunakan untuk memprediksi label kelas data uji. Terdapat beberapa metode untuk mengukur kedekatan antara data baru dan data lama (data latih), salah satunya dengan metode *Cosine Similarity*.

Cosine Similarity merupakan sebuah metode yang digunakan untuk mengetahui jarak dan kesamaan antara dua vektor *n* dimensi dengan mencari cosinus dari sudut di antara keduanya. Ini dapat digunakan untuk membandingkan dua dokumen dengan mencari titik antara dua identitas (Singh et al., 2020). Persamaan matematis *Cosine Similarity* dapat dilihat pada persamaan 2.5 (Nurjanah et al., 2017):

$$CosSim(q, d_j) = \frac{d_j \cdot q}{|d_j| \cdot |q|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (2)$$

Keterangan:

- $CosSim(q, d_j)$ adalah nilai kemiripan antara dokumen uji (q) dengan dokumen latih ke j (d_j).
- *t* adalah jumlah kata (term).
- *d* adalah document.
- *q* adalah query (kata kunci).
- w_{ij} adalah bobot kata (term) ke *i* pada dokumen latih *j*.
- w_{iq} adalah kata (term) ke *i* pada dokumen uji *q*.

Setelah menyelesaikan perhitungan jarak antar dokumen, fase berikutnya yang dilakukan untuk metode *K-Nearest* Neighbor yaitu:

1. Menentukan nilai *k*
2. Mengurutkan jarak dari nilai tertinggi hingga terendah.
3. Setelah itu dilakukan pengambilan data sebanyak nilai *k* terdekat.
4. Menentukan hasil klasifikasi dari kelas mayoritas, jika hasil dari data latih dan data uji semakin mirip berdasarkan jarak, maka akan semakin sesuai.

1.5 Evaluasi

1.5.1 K-Fold Cross Validation

K-Fold Cross-Validation merupakan metode statistik yang digunakan untuk mengevaluasi kinerja model machine learning dengan membagi data menjadi *k* subset yang memiliki ukuran yang sama. Metode ini digunakan untuk menguji seberapa baik dan akurat performa sebuah model (Tempola et al., 2018). Tujuannya adalah untuk mendapatkan hasil prediksi atau klasifikasi yang tidak hanya memiliki akurasi tinggi, tetapi juga validitas yang baik. Hal ini memungkinkan untuk mengidentifikasi kelemahan dan kekuatan dari model yang digunakan, serta melihat potensi kesalahan yang mungkin terjadi dalam praktiknya.

Metode ini bekerja dengan membagi dataset menjadi beberapa lipatan(*fold*), di mana satu lipatan(*fold*) digunakan sebagai data uji, sedangkan lipatan(*fold)* lainnya sebagai data latih. Dalam setiap iterasi *fold* pengujian, data uji yang dipakai tidak boleh tumpang tindih antara satu *fold* dengan *fold* lainnya. (Oktaviani Putri & Cahya Wihandika, 2020).

1.5.2 Confusion Matrix

Confusion Matrix ini merupakan sebuah tabel yang digunakan untuk mengevaluasi performa model dari penelitian atau klasifikasi sentimen. Evaluasi dapat dilakukan dengan

beberapa indikator seperti *accuracy*, *precision*, *f-measure* dan *recall* (Pristiyanti et al., 2018). Tabel *Confusion Matrix* terdiri dari empat metrik utama, yang dapat dilihat pada Tabel 1.

Tabel 1 Confusion Matrix

	True (Prediksi)	*False* (Prediksi)
True (Aktual)	*True Positive* (TP)	*False Positive* (FP)
False (Aktual)	*False Negative* (FN)	*True Negative* (TN)

Keterangan:

- *True Positive* (TP) menyatakan jumlah data yang memiliki kelas *actual* positif yang diklasifikasi oleh sistem sebagai data positif
- *False Positive* (FP) menyatakan jumlah data yang memiliki kelas *actual* negatif yang diklasifikasi sistem sebagai data positif
- *False Negative* (FN) menyatakan jumlah data yang memiliki kelas *actual* positif yang diklasifikasi oleh sistem sebagai data negatif
- *True Negative* (TN) menyatakan jumlah data yang memiliki kelas *actual* negatif yang diklasifikasi oleh sistem sebagai data negatif

Perhitungan yang dapat dilakukan dengan menggunakan confusion matrix antara lain:

Accuracy menyatakan persentase dari jumlah data uji yang diklasifikasikan dengan benar oleh model klasifikasi (Suyanto, 2017). Nilai *accuracy* dapat dihitung menggunakan persamaan 2.6 (Cholissodin & Soebroto, 2019).

$$accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (3)$$

Precision memberi persentase seberapa baik model klasifikasi dapat memprediksi data positif benar pada kenyataannya. Nilai *precision* dapat dihitung menggunakan persamaan 2.7 (Cholissodin & Soebroto, 2019).

$$Precision = \frac{TP}{FP+TP} \quad (4)$$

Recall atau yang sering disebut *sensitiviy* ini mengukur persentase banyaknya data yang sukses diprediksi sebagai positif dibandingkan dengan keseluruhan data yang sebenarnya positif. Nilai *recall* dapat dihitung menggunakan persamaan 2.8 (Cholissodin & Soebroto, 2019).

$$Recall = \frac{TP}{FN+TP} \quad (5)$$

F-Measure merupakan rata-rata harmonis antara presisi dan sensitivitas dalam satu skor tunggal untuk memberikan pemahaman yang lebih lengkap tentang kinerja suatu sistem. Nilai *F-Measure* dapat dihitung menggunakan persamaan 2.9

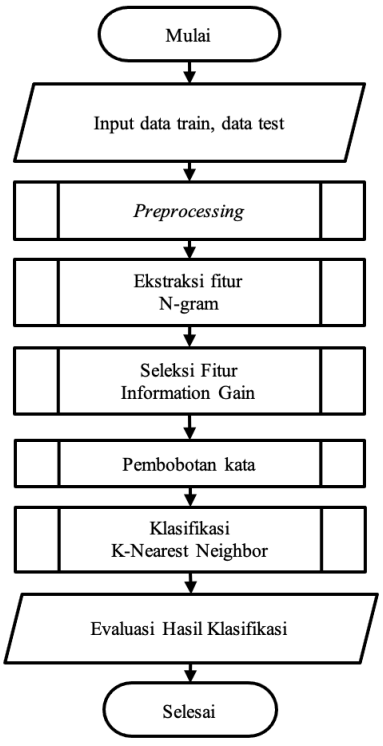
$$F - Measure = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (6)$$

2. METODE

2.1 Pengumpulan Data

Proses pengumpulan data dilakukan melalui Teknik crawling data, di mana peneliti menggunakan algoritma crawling data di pemrograman Python menggunakan YouTube Data API untuk mengakses dan mengambil informasi dari komentar dari vidio yang berkaitan dengan TikTokShop. Langkah pertama melibatkan pencarian video-video relevan dengan menggunakan kata kunci terkait TikTokShop melalui antarmuka YouTube. Setelah video-video tersebut diidentifikasi, peneliti menggunakan teknik crawling data untuk mengakses kolom komentar yang terdapat pada masing-masing video dan mengambil informasi dari kolom komentar, termasuk tanggapan dan pendapat yang diungkapkan oleh masyarakat Indonesia.

2.2 Perancangan Algoritma



Gambar 1 Diagram Alir Sistem

Berdasarkan Gambar 1 menjelaskan tahapan-tahapan yang akan dilakukan pada penelitian analisis sentimen opini masyarakat

terhadap fenomena TikTokShop di Indonesia. Adapun tahapan-tahapan yang dilakukan meliputi *Preprocessing*, untuk mengolah kalimat dan membersihkannya dan dipotong menjadi kata per kata yang nantinya digunakan sebagai term. Setelah itu dilakukan kombinasi kata *N-gram* untuk membangkitkan kata dalam sebuah data, dilanjutkan dengan Perhitungan pembobotan kata dengan metode *Term Frequency – Inverse Document Frequency* (TF-ITF), setelah itu melakukan seleksi fitur dengan metode *Information Gain* dengan perhitungan sesuai dengan rumus 1. Hasil yang didapatkan dari seleksi fitur Information Gain kemudian dilakukan klasifikasi data menggunakan metode K-Nearest Neighbor.

3. PENGUJIAN DAN ANALISIS

Dalam penelitian ini, dilakukan pengujian menggunakan metode *K-Fold Cross Validation* dengan membagi menjadi 10 iterasi, proses ini melibatkan pembagian data menjadi sepuluh bagian sayang sama besar.

3.1 Penentuan Rata-rata Fold Terbaik

Setelah menguji dengan teliti menggunakan metrik Akurasi, *Recall*, *Precision* dan *F-measure* untuk setiap iterasi dan variasi *k*. Pada Tabel 1 menampilkan hasil rata-rata dari *Fold* terbaik untuk mendapatkan hasil terbaik dari semua *Fold*.

Tabel 1 Penentuan Fold Terbaik

Fold	Akurasi	*Recall*	*Precision*	*F-Measure*
1	70.40%	77.60%	67.84%	72.36%
2	78.30%	78.30%	78.62%	78.24%
3	77.60%	77.60%	77.65%	77.59%
4	76.40%	76.40%	77.00%	76.25%
5	81.40%	81.40%	81.46%	81.39%
6	80.80%	80.80%	81.34%	80.70%
7	79.60%	79.60%	80.52%	79.43%
8	78.30%	78.30%	79.08%	77.28%
9	82.40%	82.40%	83.86%	83.21%
10	85.30%	85.30%	85.75%	85.25%

hasil rata-rata terbaik ditemukan pada *fold* ke-10 dengan akurasi, *Recall*, *Precision*, dan *F-Measure*, yakni masing-masing sebesar 85,30%, 85,30%, 85.75%, dan 85.25%.

3.2 Penentuan Rara-rata kinerja seluruh nilai K terbaik dari *10-Fold*

Setelah melakukan *10-Fold Cross-Validation* dengan berbagai nilai *k*, diperoleh hasil rata-rata. Hasil rata-rata dari setiap nilai *k* tersebut menunjukkan nilai *Fold* terbaik. Tabel 2 memperlihatkan hasil rata-rata pengujian dari semua nilai *k*.

Tabel 2 rata-rata k terbaik

k	Akurasi	*Recall*	*Precision*	*F-Measure*
3	74.10%	75.83%	76.54%	74.16%
5	76.20%	78.00%	78.83%	77.83%
9	79.40%	81.00%	81.50%	80.93%
11	78.90%	80.00%	80.70%	79.90%
13	80.60%	81.33%	81.79%	82.94%
15	80.30%	81.50%	81.98%	81.44%
17	79.50%	80.33%	80.86%	80.26%
19	80.20%	80.67%	81.16%	80.59%
21	80.50%	81.33%	81.90%	81.26%
23	80.80%	82.33%	82.76%	82.30%

Pada Tabel 2 menampilkan hasil rata-rata dari Semua nilai *k*, dengan pengujian akurasi, *Precision*, *Recall*, dan *F-Measure*. Hasil evaluasi menunjukkan bahwa nilai rata-rata tertinggi diperoleh saat $k = 23$, dengan akurasi mencapai 80.80%, *F-Measure* mencapai 82.30%, *Recall* mencapai 82.33%, dan *Precision* mencapai 82.30%.

3.3 Pengujian Pengaruh Penggunaan Ngram

Pengujian pengaruh penggunaan *N-Gram* ini dilakukan dengan menggunakan data yang berasal dari *Fold* ke-10. *Fold* ke-10 dipilih sebagai data uji karena memiliki rata-rata akurasi yang tertinggi dibandingkan dengan fold lainnya dan pengujian ini juga menggunakan nilai $k = 23$. Nilai *k* ini dipilih karena, berdasarkan pada pengujian 10-*Fold Cross Validation*, $k = 23$ memberikan hasil rata-rata akurasi yang paling tinggi. Tabel 3 memperlihatkan hasil dari pengujian pengaruh penggunaan *N-Gram*.

Tabel 3 pengaruh N-Gram

Uji	Akurasi	*Recall*	*Precision*	*F-Measure*
Unigram	89.00%	89.00%	89.02%	89.00%
Bigram	72.00%	72.00%	75.28%	71.06%
Gabungan	87.00%	87.00%	87.01%	87.00%

Berdasarkan Tabel 3 dapat diketahui bahwa penggunaan ekstraksi fitur *N-Gram* dalam analisis sentimen menggunakan metode K-Nearest Neighbor telah menghasilkan akurasi tertinggi, yaitu sebesar 89.00%, 89.00% pada *Recall*, 89,02% pada *Precision* serta pada *F-Measure* menghasilkan 89% pada variasi fitur *Unigram*.

3.4 Pengujian Pengaruh Information Gain

Pengujian ini dilakukan untuk melihat bagaimana penggunaan seleksi fitur *Information Gain* dan *N-gram* dapat mempengaruhi hasil klasifikasi. Klasifikasi ini akan dilakukan dengan menggunakan metode *K-Nearest Neighbor*. Pengujian ini dilakukan dengan menguji berbagai nilai *threshold* seleksi fitur *Information Gain* untuk mencari tahu *threshold* yang optimal untuk melakukan klasifikasi dalam sistem. Selain itu, penelitian ini juga melakukan pengujian tanpa menggunakan seleksi fitur *Information Gain* untuk mengevaluasi dampaknya terhadap hasil klasifikasi pada sistem. Rentang nilai *threshold* yang diuji adalah 25%, 50%, 75%, 90% dan 100% untuk tanpa menggunakan seleksi fitur *Information Gain*. Pada Tabel 4 berikut, ditampilkan hasil pengujian fitur pengaruh seleksi fitur *Information Gain* pada Fitur *Unigram*.

Tabel 4 Uji Pengaruh IG pada fitur Unigram

Pengujian Information Gain menggunakan Unigram				
Threshold	***Akurasi***	***Recall***	***Prescision***	***F-Measure***
25%	50.00%	50.00%	25.00%	33.33%
50%	50.00%	50.00%	25.00%	33.33%
75%	50.00%	50.00%	25.00%	33.33%
90%	55.00%	55.00%	65.26%	45.91%
100%	89.00%	89.00%	89.02%	89.00%

Hasil pengujian fitur pengaruh seleksi fitur Information Gain pada Fitur Bigram ditampilkan pada Tabel 5.

Tabel 5 Uji Pengaruh IG pada fitur Bigram

Pengujian Information Gain menggunakan Bigram				
Thres hold	***Akurasi***	***Recall***	***Prescision***	***F-Measure***
25%	50.00%	50.00%	25.00%	33.33%
50%	50.00%	50.00%	25.00%	33.33%
75%	50.00%	50.00%	25.00%	33.33%
90%	50.00%	50.00%	25.00%	33.33%
100%	72.00%	72.00%	75.28%	71.06%

Hasil pengujian fitur pengaruh seleksi fitur *Information Gain* pada Fitur Gabungan ditampilkan pada tabel Tabel 6.

Tabel 6 Uji Pengaruh IG pada fitur Gabungan

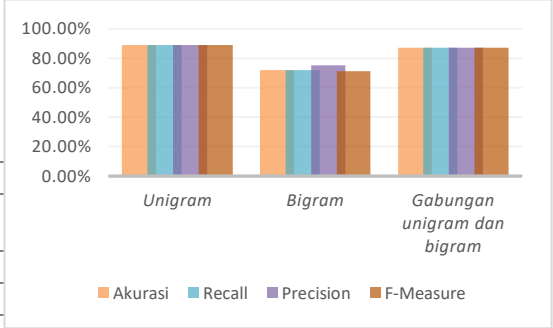
Pengujian Information Gain menggunakan Gabungan				
Thres hold	***Akurasi***	***Recall***	***Prescision***	***F-Measure***
25%	50.00%	50.00%	25.00%	33.33%
50%	50.00%	50.00%	25.00%	33.33%
75%	50.00%	50.00%	25.00%	33.33%
90%	51.00%	51.00%	75.25%	35.52%
100%	87.00%	87.00%	87.01%	87.00%

3.5 Analisis Pengujian Pengaruh Rata-Rata Akurasi Kinerja Pada 10-Fold Cross Validation

Gambar 2 Pengujian pengaruh Information Gain

Seperti yang tertera pada Gambar 2, dari hasil pengujian, didapatkan nilai *k* terbaik pada saat *k* = 23 karena memiliki rata-rata Akurasi tertinggi, hasil pengujian pada k = 23 menunjukan bahwa rata-rata Recall sebesar 80.80%, Precision sebesar 81.18%, akurasi sebesar 80.80%, dan pada F-Measure sebesar 80.76%. Dalam analisis ini, ditemukan bahwa hasil akurasi yang dihasilkan cenderung stabil dan bahkan meningkat sejalan dengan peningkatan nilai k. Faktor ini menunjukkan bahwa prediksi yang dihasilkan semakin konsisten seiring peningkatan jumlah tentangga yang dipertimbangkan dalam proses prediksi. Dengan mempertimbangkan lebih banyak tetangga, metode KNN meningkatkan kemungkinan dapat menghasilkan representasi struktur data yang lebih baik. Peningkatan hasil yang relatif konsisten ini juga dipengaruhi oleh penyebaran data yang merata dan sebaran data yang seimbang. Penelitian ini menunjukan bahwa akurasi, *Recall*, *Precision* dan *F-Measure* memiliki nilai yang stabil.

3.6 Analisis Pengujian Pengaruh Variasi *N-Gram*



Gambar 3 Pengujian Variasi N-Gram

Berdasarkan Gambar 3 hasil pengujian variasi *N-Gram* menunjukkan bahwa akurasi tertinggi terjadi pada variasi *Unigram*, dengan persentase sebesar 89% pada akurasi, 89% pada *Recall*, *Precision* sebesar 89.02%, dan *F-Measure* sebesar 89.00%. Hal tersebut disebabkan karena *Unigram* memperlakukan setiap kata sebagai entitas tunggal sehingga membuat representasi teks menjadi sederhana dan memungkinkan sistem untuk lebih fokus pada frekuensi kemunculan kata tunggal, sehingga lebih mudah dipelajari dan dianalisis.

Selain itu, *Unigram* dapat menangkap pola-pola yang lebih umum dan penting, hal tersebut karena pada penelitian ini melakukan sentimen analisis opini masyarakat terhadap persetujuan atau ketidaksetujuan masyarakat terhadap fenomena TiktokShop. Setiap kata tunggal tertentu dapat memiliki makna yang cukup kuat untuk menggambarkan sentimen secara langsung tanpa memerlukan pasangan kata untuk memberikan konteks tambahan.

Sedangkan akurasi yang dihasilkan variasi bigram dan variasi Gabungan cenderung lebih rendah, persentase akurasi pada variasi bigram sebesar 74% dan 87% pada variasi gabungan. Hal ini disebabkan karena pada variasi bigram hanya berisi gabungan dari kata-kata tunggal yang sebelumnya dan sesudahnya saling berdekatan yang membentuk term baru menjadi term *bigram*, sehingga kata tunggal tertentu yang dapat memiliki makna yang cukup kuat untuk menggambarkan sentimen seperti term pada term unigram akan membentuk term baru, sehingga term bigram akan sedikit ditemukan pada dokumen latih dan kebanyakan satu term bigram hanya terdapat pada satu dokumen.

Penurunan akurasi yang rendah pada variasi bigram juga disebabkan juga oleh term kunci yang secara khas diasosiasikan dengan kelas positif juga seringkali muncul dalam kelas negatif. Kehadiran term positif yang signifikan dalam kelas negatif dapat menyebabkan model mengalami kebingungan dan kesulitan untuk membedakan antara kedua kelas tersebut secara akurat.

3.7 Hasil pengujian pengaruh Information Gain

Berdasarkan pengujian pengujian seleksi fitur *Information Gain* terhadap variasi *N-Gram* yang telah dilakukan, didapatkan analisis bahwa *Information Gain* pada penelian ini menghasilkan akurasi yang kurang baik tidak perperan secara optimal, ini dikarenakan pemotongan term dalam proses seleksi fitur menggunakan metode *Information Gain* dianggap terlalu agresif, yang berdampak pada kehilangan sejumlah informasi penting dalam dataset. Informasi ini seharusnya dapat menjadi bahan yang digunakan sistem untuk membuat prediksi yang akurat. Efek dari pemotongan term yang agresif ini menyebabkan banyak dokumen menjadi kosong, dan dalam beberapa kasus, dokumen-dokumen tersebut bahkan tidak mempunyai term sama sekali. Hal ini mengakibatkan sistem tidak mampu mengklasifikasi dokumen-dokumen tersebut yang berpotensi mengurasi efektivitas dan akurasi sistem.

Selain itu, hal lain yang mengindikasikan bahwa metode seleksi fitur Information Gain mungkin kurang ideal untuk digunakan dalam penelitian ini karena fokus penelitian yang digunakan pada penelitian ini, fokus penelitian ini adalah sentimen analisis opini masyarakat mengenai TikTokShop yang secara spesifik membahas tentang persetujuan dan ketidaksetujuan masyarakat terhadap keputusan penutupan TikTokShop yang dilakukan oleh Kementerian Perdagangan Indonesia seiring dengan perubahan regulasi yang berlaku. Dalam konteks opini tentang persetujuan dan ketidaksetujuan ini, terdapat term kunci “tuju” yang digunakan sebagai salah satu faktor dalam proses mengklasifikasi data. Term “tuju” ini sering muncul di setiap dokumen yang menunjukkan persetujuan terhadap kebijakan ini, mendandakan bahwa masyarakat secara umum setuju dengan kebijakan tersebut. Namu, dalam proses seleksi fitur menggunakan metode information gain, term “tuju” ini malah terhapus dikarenakan frekuensi kemunculannya yang sering dan ada pada setiap dokumen dengan kelas Positif. Penghasspusan term ini menyebabkan kata kunci pada dataset menjadi hilang dan sebagai akibatnya, sistem akan mengalami kesulitan dalam membedakan antara data kelas Positif dan kelas Negatif. Ini tentu saja menjadi masalah serius, mengingat pentingnya term “tuju” dalam kon teks penelitian ini.

4. Penutup

4.1 Kesimpulan

Berdasarkan hasil pengujian yang telah dilakukan dan analisis yang telah diuraikan pada bab sebelumnya, dapat ditarik kesimpulan bahwa analisis sentimen opini masyarakat

terhadap fenomena Tiktokshop Menggunakan Metode *K-Nearest Neighbor* Berbasis *N-Gram* Dengan Seleksi Fitur *Information Gain* menunjukkan performa yang baik dengan tingkat akurasi mencapai 89%, diikuti oleh nilai *Recall, Precision*, dan *F-Measure*, yaitu 90%, 88.24%, dan 89.11% secara berurutan. Hasil pengujian ini menunjukkan bahwa fitur *Unigram* adalah fitur yang paling cocok untuk menganalisis sentimen opini masyarakat terhadap fenomena TikTokShop dan dapat diambil kesimpulan bahwa seleksi fitur *Information Gain* dalam penelitian ini tidak perperan secara optimal dalam menangani kasus sentimen opini masyarakat mengenai opini masyarakat persetujuan dan ketidaksetujuan terhadap suatu fenomena.

5. DAFTAR PUSTAKA

Cholissodin, I., & Soebroto, A. (2019). *Buku Ajar AI, Machine Learning & Deep Learning*.

Chormunge, S., & Jena, S. (2016). Efficient Feature Subset Selection Algorithm for High Dimensional Data. *International Journal of Electrical and Computer Engineering (IJECE)*, *6*, 1880–1888. https://doi.org/10.11591/ijece.v6i4.9800

Chuzaimah Zulkifli, U. (2018). Pengembangan Modul PreprocessingTeks untuk Kasus Formalisasi dan Pengecekan Ejaan Bahasa Indonesia pada Aplikasi Web Mining Simple Solution (WMSS). *Jurnal Matematika Statistika Dan Komputasi*, *15*(2). https://doi.org/10.20956/jmsk.v15i2.5718

Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3459665

Fauzi, M. A., Utomo, D. C., Pramukantoro, E. S., & Setiawan, B. D. (2017). Automatic essay scoring system using N-GRAM and cosine similarity for gamification based elearning. *ACM International Conference Proceeding Series*, *Part F1312*(October), 151–155. https://doi.org/10.1145/3133264.3133303

Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment Analysis on Twitter Data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, *8*(6). https://doi.org/10.14569/ijacsa.2017.080603

Indriati, I.-, Rahayudi, B., & Dewi, C. (2021). Analisis Sentimen Mengenai Moda Raya Terpadu (MRT) Jakarta dengan Metode BM25 dan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *8*(2). https://doi.org/10.25126/jtiik.2021824508

Indriati, I., & Ridok, A. (2016a). SENTIMENT ANALYSIS FOR REVIEW MOBILE APPLICATIONS USING NEIGHBOR METHOD WEIGHTED K-NEAREST NEIGHBOR (NWKNN). *Journal of Enviromental Engineering and Sustainable Technology*, *3*(1). https://doi.org/10.21776/ub.jeest.2016.003.01.4

Indriati, I., & Ridok, A. (2016b). Sentiment Analysis for Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn). *Journal of Enviromental Engineering and Sustainable Technology*, *3*(1), 23–32. https://doi.org/10.21776/ub.jeest.2016.003.01.4

Jodha, R., Sanjay Bc, G., & Chowdhary, K. R. (2018). Text Classification using KNN with different Feature Selection Methods. *International Journal of Research*, *9*(1).

Liu, B. (2012). *Sentiment analysis and opinion mining*. Springer.

Muthia, D. A. (2016). Opinion Mining Pada Review Buku Menggunakan Algoritma Naive Bayes. *Jurnal TEKNIK KOMPUTER*, *II*(1), 1–8.

Nafan, M. Z., & Amalia, A. E. (2019). Kecenderungan Tanggapan Masyarakat terhadap Ekonomi Indonesia berbasis Lexicon Based Sentiment Analysis. *JURNAL MEDIA INFORMATIKA BUDIDARMA*. https://doi.org/10.30865/mib.v3i4.1283

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(12), 1750–1757.

Oktaviani Putri, F., & Cahya Wihandika, R. (2020). Analisis Sentimen pada Ulasan Pengguna MRT Jakarta Menggunakan Metode Neighbor-Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain. *J-Ptiik.Ub.Ac.Id*, *4*(7).

Onantya, I. D., Indriati, & Adikara, P. P. (2019). Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan BM25 Dan Improved K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), 2575–2580.

Pramono, F., Didi Rosiyadi, & Windu Gata.

(2019). Integrasi N-gram, Information Gain, Particle Swarm Optimation di Naïve Bayes untuk Optimasi Sentimen Google Classroom. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(3), 383–388. https://doi.org/10.29207/resti.v3i3.1119

Pristiyanti, R. I., Fauzi, M. A., & Muflikhah, L. (2018). Sentiment Analysis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(3).

Rosiyana, R. N., Agustin, M., Kalka Iskandar, I., & Luckyardi, S. (2021). A NEW DIGITAL MARKETING AREA FOR E-COMMERCE BUSINESS. In *International Journal of Research and Applied Technology* (Vol. 1, Issue 2).

Sabily, A. F., Adikara, P. P., & Fauzi, M. A. (2019). Analisis Sentimen Pemilihan Presiden 2019 pada Twitter menggunakan Metode Maximum Entropy. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(5), 4204–4209.

Saputri, R. P., Winahju, W. S., Fithriasari, K., Statistika, D., Matematika, F., & Data, S. (2019). *Klasifikasi Sentimen Wisatawan Candi Borobudur pada Situs TripAdvisor Menggunakan Support Vector Machine dan K-Nearest Neighbor*. *8*(2).

Singh, R. H., Maurya, S., Tripathi, T., Narula, T., & Srivastav, G. (2020). Movie Recommendation System using Cosine Similarity and KNN. *International Journal of Engineering and Advanced Technology*, *9*(5), 556–559. https://doi.org/10.35940/ijeat.e9666.069520

Suyanto. (2017). *DATA MINING: untuk klasfikasi dan klasterisasi data*. Informatika.

Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(5). https://doi.org/10.25126/jtiik.201855983

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, *24*(7). https://doi.org/10.1016/j.knosys.2011.04.014

Wang, Q., Guan, Y., Wang, X. L., & Xu, Z. M. (2006). A Novel Feature Selection Method Based on Category Information Analysis for Class Prejudging in Text Classification. *International Journal of Computer Science and Network Security*, *6*(1a).

Windisch, G., & Csink, L. (2005). Language identification using global statistics of natural languages. *Proceedings of the 2nd Romanian-Hungarian Joint Symposium on Applied Computational Intelligence (SACI)*, 243--255.