3rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
30 Agustus 2023 – Jakarta, Indonesia
Volume 2, Nomor 2, September 2023 - ISSN 2962-8628 (*online*)

SENTIMEN ANALISIS TENTANG HILIRISASI INDUSTRI BERDASARKAN OPINI MASYARAKAT DI TWITTER MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*

Marlina Hidayat[1*], Utomo Budiyanto[2]

[1*,2]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1*]linahdyt2030@yahoo.com, [2]utomo.budiyanto@budiluhur.ac.id
(* : corresponding author)

Abstrak- Perkembangan teknologi informasi telah mengalami kemajuan yang signifikan, terutama dalam memungkinkan masyarakat untuk menyampaikan opini mereka tanpa batasan waktu melalui media sosial. Salah satu platform media sosial yang populer untuk menyampaikan opini adalah Twitter. Banyak orang menggunakan Twitter untuk berbagi pendapat dan berpartisipasi dalam diskusi mengenai berbagai topik yang beragam, termasuk isu kebijakan hilirisasi. Hilirisasi merupakan salah satu strategi untuk meningkatkan nilai tambah komoditas yang dimiliki oleh negara. Topik ini memicu berbagai respons dari masyarakat di Twitter. Dengan banyaknya cuitan yang mengulas mengenai kebijakan hilirisasi, terdapat beragam opini dan sentimen yang terkandung di dalamnya. Opini yang diungkapkan oleh masyarakat di Twitter dapat menjadi referensi penting untuk memahami tingkat kepuasan mereka terhadap strategi dan kebijakan yang diusulkan. Namun, menafsirkan opini-opini ini secara langsung bisa menjadi sulit. Oleh karena itu, pendekatan alternatif yang dapat digunakan adalah proses text mining. Dalam penelitian ini, dilakukan analisis sentimen terhadap data cuitan di Twitter yang telah dikumpulkan selama 1 Mei – 30 Juni 2023 menggunakan *tweepy*. Selanjutnya, dilakukan ekstraksi fitur menggunakan metode TF-IDF, dan data tersebut diklasifikasikan menggunakan algoritme *K-Nearest Neighbor* (K-NN) menjadi beberapa label sentimen, yaitu sentimen positif dan negatif. Hal ini dilakukan untuk mengevaluasi performa algoritme K-Nearest Neighbor (K-NN) dengan ekstraksi fitur TF-IDF dalam mengklasifikasikan sentimen opini masyarakat tentang hilirisasi di Twitter dan mengukur akurasi analisis sentimen algoritme K-NN pada opini tersebut. Hasil pengujian menunjukkan bahwa algoritme *K-Nearest Neighbor* menghasilkan akurasi optimal sebesar 80,77%, presisi 84,38%, dan *recall* 84,38% dengan menggunakan nilai k=7. Dimana total sentimen positif yang dihasilkan yaitu sebanyak 32 data dan 20 data merupakan sentimen negatif.

Kata Kunci: Hilirisasi, Twitter, Analisis Sentimen, *K-Nearest Neighbor*

SENTIMENT ANALYSIS OF INDUSTRIAL DOWNSTREAMING BASED ON PUBLIC OPINION ON TWITTER USING K-NEAREST NEIGHBOR METHOD

Abstract- *The development of information technology has made significant progress, especially in enabling people to express their opinions without time limits through social media. One of the popular social media platforms for expressing opinions is Twitter. Many people use Twitter to share opinions and participate in discussions on a wide variety of topics, including downstream policy issues. Downstreaming is one of the strategies to increase the added value of commodities owned by the state. This topic sparked various responses from the public on Twitter. With so many tweets discussing downstream policies, there are various opinions and sentiments contained in them. The opinions expressed by the public on Twitter can be an important reference to understand their level of satisfaction with the proposed strategies and policies. However, interpreting these opinions directly can be difficult. Therefore, an alternative approach that can be used is the process of text mining. In this study, sentiment analysis was carried out on tweet data on Twitter that had been collected from 1 May to 30 June 2023 using tweepy. Furthermore, feature extraction was carried out using the TF-IDF method, and the data was classified using the K-Nearest Neighbor (K-NN) algorithm into several sentiment labels, namely positive and negative sentiments. This is done to evaluate the performance of the K-Nearest Neighbor (K-NN) algorithm with TF-IDF feature extraction in classifying public opinion sentiment about downstream on Twitter and measuring the accuracy of the K-NN algorithm sentiment analysis on that opinion. The test results show that the K-Nearest Neighbor algorithm produces optimal accuracy of 80.77%, 84.38% precision, and 84.38% recall using a value of k = 7. Where the total positive sentiment generated is as much as 32 data and 20 data is a negative sentiment.*

Keywords*: Downstreaming, Twitter, Sentiment Analysis, K-Nearest Neighbor*

1. PENDAHULUAN

Perkembangan teknologi telah memberikan dampak besar terhadap kemudahan akses dan berbagi informasi, terutama melalui media sosial. Media sosial memungkinkan orang-orang untuk terhubung dan berinteraksi satu sama lain di seluruh dunia dengan cara yang lebih cepat dan mudah, seperti berbagi foto, video, dan lainnya. Salah satu media sosial yang populer adalah Twitter, yang pada tahun 2023 memiliki pengguna media sosial sebanyak 167 juta di Indonesia. Jumlah pengguna Twitter di Indonesia pada awal tahun 2023 mencapai 24 juta pengguna [1].

Twitter tidak hanya menjadi tempat untuk berkomunikasi, tetapi juga menjadi platform di mana masyarakat dapat beropini dan menyampaikan pendapat tentang berbagai topik. Dalam beberapa waktu terakhir, isu kebijakan hilirisasi industri menjadi topik yang banyak dibicarakan di Indonesia, termasuk di lingkungan Twitter. Hilirisasi juga disebut sebagai *downstreaming* atau peningkatan nilai, merujuk pada usaha untuk mengurangi ekspor bahan mentah dan sebaliknya mendorong pemanfaatan bahan tersebut di dalam negeri dalam sektor industri. Tujuannya adalah untuk meningkatkan nilai tambah dalam negeri dan menciptakan peluang kerja. Jika terjadi kebutuhan untuk melakukan ekspor, fokusnya adalah pada pengiriman barang jadi yang merupakan hasil dari proses olahan bahan baku tersebut [2].

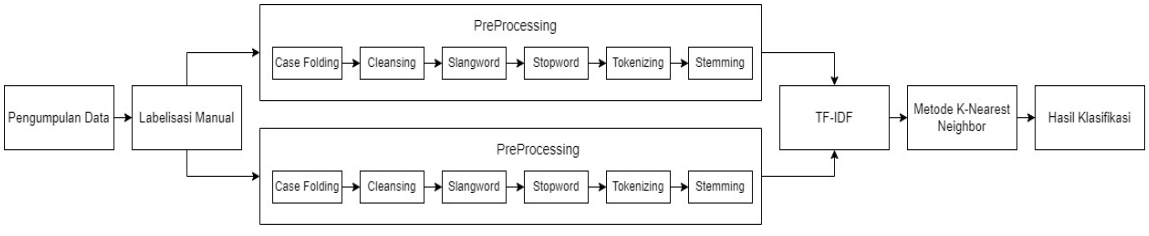
Topik hilirisasi menarik beragam reaksi dari masyarakat di Twitter, sehingga terdapat berbagai opini dan sentimen yang terungkap dalam cuitan-cuitan tersebut. Dengan banyaknya pengguna Twitter yang mengungkapkan opini-opini mereka, platform ini dapat dimanfaatkan untuk mencari informasi tentang polaritas positif dan negatif dari opini tersebut. Analisis informasi ini memerlukan penerapan teknik analisis yang tepat agar hasilnya dapat memberikan manfaat bagi berbagai pihak. Salah satu metode penelitian yang cocok untuk menganalisis opini-opini pada cuitan tersebut adalah analisis sentimen [3]. Analisis sentimen merupakan proses penerapan analisis teks untuk mengumpulkan berbagai data dari internet dan berbagai platform media sosial, dengan tujuan untuk memahami opini pengguna yang terdapat pada platform-platform tersebut. Analisis sentimen dapat dianggap sebagai aktivitas klasifikasi, di mana sentimen dalam data teks dikelompokkan ke dalam kelasnya masing-masing [4].

Sejumlah penelitian sebelumnya telah melakukan analisis sentimen pada Twitter dengan menggunakan berbagai algoritme. Penelitian sebelumnya tentang *fine-grained sentiment analysis* yang mengekstrak sentimen dari twitter dengan studi kasus pada PT. Indosat Tbk dengan dataset dari twitter sebanyak 19.948 data menunjukkan hasil akurasi sebesar 78% untuk *Naïve Bayes* dan 81% untuk *K-Nearest Neighbor* (KNN) dengan k=7 [5]. Selain itu, terdapat penelitian lain menggunakan algoritme *K-Nearest Neighbor* (KNN) dan ekstraksi fitur TF-IDF untuk menganalisis sentimen pengguna Twitter tentang kebijakan Pemerintah terkait Pembelajaran Daring. Studi ini memakai 1825 data cuitan berbahasa Indonesia dari 1 Februari - 30 September 2020, dengan akurasi rata-rata 84,65% [6]. Penelitian lain mengenai evaluasi sentimen pendapat publik terhadap kinerja KPK menggunakan data dari platform Twitter. Data ini kemudian dianalisis menggunakan algoritma *Support Vector Machine* (SVM) pada 2000 cuitan, dan hasilnya menunjukkan akurasi rata-rata sekitar 82%, dengan dominasi label negatif mencapai 77%.[7].

Berdasarkan hasil penelitian sebelumnya, maka penelitian ini difokuskan pada analisis sentimen menggunakan algoritme *K-Nearest Neighbor* (K-NN) untuk menganalisis opini masyarakat di Twitter tentang isu hilirisasi. Penggunaan algoritme *K-Nearest Neighbor* (KNN) dengan ekstraksi fitur TF-IDF diharapkan dapat mengklasifikasikan cuitan dari Twitter dengan baik, sehingga menghasilkan dua klasifikasi sentimen, yaitu positif dan negatif. Pengujian algoritme akan menunjukkan tingkat akurasi dan ketepatan algoritme *K-Nearest Neighbor* (KNN) dalam melakukan klasifikasi sentimen.

2. METODE PENELITIAN

Dalam menganalisa sentimen dan mengetahui akurasinya, ada beberapa tahapan untuk mendapatkan hasil yang terbaik. Keseluruhan metodologi analisis sentimen yang dilakukan dapat dilihat pada Gambar 1 berikut ini.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Penelitian ini mengumpulkan data dari Twitter dengan menggunakan metode *crawling* melalui API Twitter, Untuk mendapatkan API Key Twitter, peneliti melakukan pendaftaran akun Twitter di halaman *Twitter Developer*. Data tweet diambil dari API Twitter dengan menggunakan kata kunci "Hilirisasi" dan proses *crawling* dilakukan dengan memanfaatkan *library tweepy* dalam bahasa pemrograman Python. Sebanyak 260 tweet terkait kata kunci "hilirisasi" berhasil dikumpulkan dari Mei hingga Juni 2023. Data kemudian disimpan dalam format Excel untuk analisis lebih lanjut.

2.2 Labelisasi Manual

Pada tahap labelisasi manual, dataset yang telah terkumpul diberi label secara manual menggunakan dua jenis label, yaitu label positif dengan nilai 0 pada data yang mengandung kalimat positif, mengindikasikan dukungan masyarakat terhadap kebijakan hilirisasi tersebut, dan label negatif dengan nilai 1 pada data yang memuat kalimat negatif, menandakan bahwa masyarakat tidak mendukung kebijakan hilirisasi tersebut. Proses pelabelan manual dilakukan berdasarkan kalimat dari tweet tersebut oleh seorang ahli ekonomi, yang juga merupakan dosen di jurusan ekonomi dan hasilnya disimpan dalam format file excel sebelum diimpor ke dalam database melalui sistem.

2.3 *Preprocessing*

Setelah data terkumpul dan diberi label, langkah berikutnya adalah tahap *preprocessing*. *Text preprocessing* adalah langkah awal dalam text mining yang melibatkan serangkaian tahapan untuk mempersiapkan data sebelum dilakukan penemuan pengetahuan pada sistem text mining. Tujuan dari proses ini adalah untuk mengolah data sehingga siap digunakan dalam analisis lebih lanjut dan operasi text mining [8]. Tahapan *preprocessing* adalah sebagai berikut:

a. *Case Folding* adalah proses mengubah semua huruf kapital dalam suatu dokumen menjadi huruf kecil (*lowercase*).
b. *Cleansing* adalah langkah untuk menghilangkan karakter yang tidak diperlukan dalam data dokumen, termasuk *url*, *username*, *mention*, *hashtag*, dan *retweet* [7].
c. *Slangword* adalah proses mengubah kata-kata yang disingkat menjadi bentuk kata dengan arti yang sama sesuai dengan KBBI. Bertujuan untuk menghasilkan informasi yang lebih mudah diproses [9].
d. *Stopword* adalah tahap di mana kata-kata yang tidak memberikan makna penting dalam data dokumen dihapus [9].
e. *Tokenization* adalah proses seleksi, pemecahan, dan pemotongan kata dalam suatu dokumen berdasarkan spasi, sehingga menghasilkan *term-term* yang terpisah [7].
f. *Stemming* adalah proses untuk mendapatkan bentuk kata dasar dengan menghilangkan awalan, akhiran, sisipan, dan konfiks setiap kata [10].

2.4 TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah suatu metode untuk menghitung bobot kata-kata yang terdapat dalam data dokumen. Data dokumen akan diubah menjadi vektor berdasarkan kata-kata (*term*) yang ada, dan vektor tersebut akan digunakan untuk proses klasifikasi. Penggunaan metode TF-IDF memungkinkan analisis menggunakan algoritma *K-Nearest Neighbor* (K-NN) [11]. Berikut adalah persamaan perhitungan pembobotan TF-IDF yang digunakan dalam penelitian ini:

$$tf_{(k,d)} = \frac{jumlah\ frekuensi\ istilah\ k\ yang\ muncul\ dalam\ dokumen\ d}{jumlah\ istilah\ dalam\ dokumen} \quad (1)$$

$$idf_{(k)} = log\frac{N}{df_k} \quad (2)$$

$$tfidf_{(k,d)} = tf_{(k,d)} * idf_{(k)} \quad (3)$$

Dalam rumus tersebut, tfidf(*k*,*d*) adalah bobot kata (*term*) yang terdapat dalam dokumen, tf(*k*,*d*) merupakan jumlah frekuensi kata tersebut dalam dokumen, N adalah total jumlah dokumen yang ada dalam database, dan df adalah jumlah dokumen yang mengandung *term* tersebut.

2.5 Metode *K-Nearest Neighbor*

K-Nearest Neighbor (K-NN) adalah metode yang sederhana dan mudah diimplementasikan. Keberadaan label pada data memudahkan proses pengelompokkan ke dalam kelas yang paling sesuai. Metode ini memiliki

keunggulan dalam mengklasifikasikan data menggunakan data latih dan data uji. Selain itu, hasil dan akurasi prediksi dapat diinterpretasikan dengan akurat dengan memperhatikan nilai k terdekat dengan tepat [11]. Adapun pada penelitian ini K-NN dihitung menggunakan perhitungan jarak *Euclidean Distance* pada persamaan berikut [12]:

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n} (x_{training}^{i} - y_{testing}^{i})^2} \quad (4)$$

Keterangan:

$d_{(x,y)}$: Jarak
$x_{training}^{i}$: *Training data*
$y_{testing}^{i}$: *Testing data*
i : variabel data
n : dimensi data

2.6 Hasil Klasifikasi

Hasil kasifikasi merupakan tampilan atau visualisasi berdasarkan perhitungan yang telah dilakukan. Visualisasi ini didapatkan setelah pengujian dengan algoritme K-NN dan juga *confusion matrix* yang digunakan untuk menghitung dan membuat kesimpulan dari hasil penelitian yang telah dilakukan. *Confusion matrix* merupakan sebuah tabel yang menggambarkan jumlah data uji yang diklasifikasikan dengan benar dan jumlah data uji yang diklasifikasikan dengan salah [5]. Pada *confusion matrix*, akan dilakukan perhitungan untuk akurasi, *precision*, dan *recall*. Contoh *confusion matrix* dapat dilihat pada Gambar 2 [13]:

	Prediksi Ya	Prediksi Tidak
Sebenarnya Ya	TP	FN
Sebenarnya Tidak	FP	TN

Gambar 2. Contoh *confusion matrix*

Berdasarkan nilai pada confusion matrix, digunakan persamaan berikut untuk menghitung nilai akurasi, presisi, dan *recall* dari model yang diuji [13].

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (5)$$

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (6)$$

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (7)$$

3. HASIL DAN PEMBAHASAN

Bagian ini mengungkapkan hasil dari penelitian yang telah dilakukan. Penelitian ini bertujuan untuk menganalisis sentimen dari opini masyarakat yang terdapat dalam cuitan-cuitan di Twitter terkait isu kebijakan hilirisasi. Selain itu, penelitian ini juga mengukur tingkat akurasi dari algoritme k-nearest neighbor untuk menilai seberapa efektifnya dalam melakukan klasifikasi sentimen.

3.1 Implementasi Metode

Implementasi metode melibatkan beberapa tahapan yang harus dilakukan secara berurutan. Tahapan-tahapan tersebut meliputi pengumpulan data, pemberian label, *preprocessing*, pembobotan kata dengan TF-IDF, dan klasifikasi menggunakan metode K-Nearest Neighbor (K-NN).

3.1.1 Pengumpulan Data

Pengumpulan data dilakukan dari Twitter menggunakan *library Tweepy* dalam bahasa pemrograman Python dengan kata kunci "hilirisasi". Total terkumpul 260 tweet dari tanggal 01 Mei 2023 hingga 30 Juni 2023. Data disimpan dalam format Excel untuk pelabelan manual dan dalam database MYSQL. Berikut ini adalah sampel data yang berhasil diambil dapat dilihat pada Tabel 1.

Tabel 1. Sampel data hasil *crawling*

publishedAt	authorDisplayName	textDisplay
2023-06-18 09:23:27	PakNasyah	@Heraloebss Kayaknya ada yang merasa dikadali, ini namanya hilirisasi anggaran, dari atas turun ke bawah mantul lagi ke atas kayak bola basket..
2023-06-15 23:06:03	wahananewsdotco	Proyek Andalan Jokowi Sudah Raup Cuan Rp 165 T #Wto #Hilirisasi #HilirisasiNikel #Jokowi https://t.co/utaKyUCrqt

3.1.2 Labelisasi Manual

Pada tahap ini, dataset yang terkumpul diberi label secara manual dengan menggunakan dua jenis label, yaitu positif (0) untuk data dengan sentimen positif, dan negatif (1) untuk data dengan sentimen negatif. Proses pelabelan dilakukan oleh seorang pakar, dan setelah itu, data disimpan dalam format file excel dan diimpor ke dalam database melalui sistem. Sampel dataset yang telah diberi label dapat dilihat pada Tabel 2.

Tabel 2. Sampel dataset setelah labelisasi manual

publishedAt	authorDisplayName	textDisplay	Label
2023-06-18 09:23:27	PakNasyah	@Heraloebss Kayaknya ada yang merasa dikadali, ini namanya hilirisasi anggaran, dari atas turun ke bawah mantul lagi ke atas kayak bola basket..	1
2023-06-15 23:06:03	wahananewsdotco	Proyek Andalan Jokowi Sudah Raup Cuan Rp 165 T #Wto #Hilirisasi #HilirisasiNikel #Jokowi https://t.co/utaKyUCrqt	0

3.1.3 Preprocessing

Selanjutnya, data yang telah diberi label akan mengalami tahap *preprocessing* untuk membersihkan data yang tidak terstruktur dan mengandung *noise*. Proses pembersihan data dilakukan sebelum analisis dengan beberapa tahapan, yaitu *case folding, cleansing, slangword, stopword, tokenization* dan *stemming*. Berikut contoh data hasil proses *preprocessing* dokumen ke-2 dari sampel labelisasi manual dilihat pada Tabel 3:

Tabel 3. Contoh hasil *preprocessing*

Tahapan	Hasil
Data Tweet Asli	Proyek Andalan Jokowi Sudah Raup Cuan Rp 165 T #Wto #Hilirisasi #HilirisasiNikel #Jokowi https://t.co/utaKyUCrqt
Case Folding	proyek andalan jokowi sudah raup cuan rp 165 t #wto #hilirisasi #hilirisasinikel #jokowi https://t.co/utakyucrqt
Cleansing	proyek andalan jokowi sudah raup cuan rp t
Slangword	proyek andalan jokowi sudah raup uang rupiah triliun
Stopword	proyek andalan jokowi raup uang rupiah triliun
Tokenization	proyek, andalan, jokowi, raup, uang, rupiah, triliun
Stemming	proyek andal jokowi raup uang rupiah triliun

3.1.4 TF-IDF

Setelah menyelesaikan proses pelabelan dan *preprocessing* , langkah selanjutnya adalah melakukan pembobotan menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Berikut ini merupakan contoh hasil perhitungan bobot kata dengan menggunakan TF-IDF dapat dilihat pada Tabel 4:

Tabel 4. Contoh hasil perhitungan TF-IDF

Kata (*term*)	TF		DF	IDF (log N/DF)	TF-IDF (TF*IDF)
	D1	D2			
proyek	0	1	1	0,30102999566398	0,30102999566398
andal	0	1	1	0,30102999566398	0,30102999566398
jokowi	0	1	1	0,30102999566398	0,30102999566398
raup	0	1	1	0,30102999566398	0,30102999566398
uang	0	1	1	0,30102999566398	0,30102999566398
rupiah	0	1	1	0,30102999566398	0,30102999566398
triliun	0	1	1	0,30102999566398	0,30102999566398

3.1.5 Klasifikasi *K-Nearest Neighbor* (K-NN)

Setelah proses pembobotan pada setiap tweet, tahap berikutnya adalah melakukan klasifikasi dengan menggunakan metode *K-Nearest Neighbor* (K-NN). Pada metode ini, kita akan memilih k tetangga terdekat dengan menggunakan jarak *Euclidean* sebagai acuan. Sebelum melakukan klasifikasi sudah dipersiapkan sampel data uji bersih setelah melalui tahap *preprocessing* yang akan digunakan pada Tabel 5 berikut:

Tabel 5. Contoh data uji setelah tahap *preprocessing*

Dataset	Label
ev murah hilir hilir arti cipta sektor kerja formal	positif

Kemudian menghitung nilai TF-IDF sampel data uji tersebut, hasil perhitungan dapat dilihat pada Tabel 6 berikut:

Tabel 6. Hasil perhitungan TF-IDF data uji

Kata (*term*)	TF	DF	IDF	TF-IDF
kayak	0	1	0,30102999566398	0
rasa	0	1	0,30102999566398	0
bohong	0	1	0,30102999566398	0
nama	0	1	0,30102999566398	0
hilir	2	1	0,30102999566398	0,60
anggar	0	1	0,30102999566398	0
atas	0	1	0,30102999566398	0
turun	0	1	0,30102999566398	0
bawah	0	1	0,30102999566398	0
pantul	0	1	0,30102999566398	0
bola	0	1	0,30102999566398	0
basket	0	1	0,30102999566398	0
proyek	0	1	0,30102999566398	0
andal	0	1	0,30102999566398	0
jokowi	0	1	0,30102999566398	0
raup	0	1	0,30102999566398	0
uang	0	1	0,30102999566398	0
rupiah	0	1	0,30102999566398	0
triliun	0	1	0,30102999566398	0

Selanjutnya dilakukan perhitungan jarak menggunakan persamaan *euclidean distance*. Hasil perhitungan jaraknya akan dilakukan pengurutan secara *ascending* dari yang terkecil hingga terbesar. Berikut hasil pengurutan jarak *euclidean* setelah dihitung nilainya dapat dilihat pada Tabel 7:

Tabel 7. Hasil pengurutan nilai jarak *euclidean*

Urutan	Jarak *Eculidean*	Data ke-

1	0,521399247	Uji 1, latih 2
2	1,126351107	Uji 1, latih 1

Jika jarak *euclidean* sudah dihitung, maka dapat diambil nilai teratas berdasarkan jumlah *k* yang telah ditentukan. Sebagai contoh nilai *k* = 1, maka pada Tabel 8 berikut ditunjukan hasil yang diperoleh berdasarkan nilai *k* yang ditentukan.

Tabel 8. Hasil *k* tetangga terdekat

Urutan	Jarak *Eculidean*	Data ke-	Label
1	0,521399247	Uji 1, latih 2	Positif

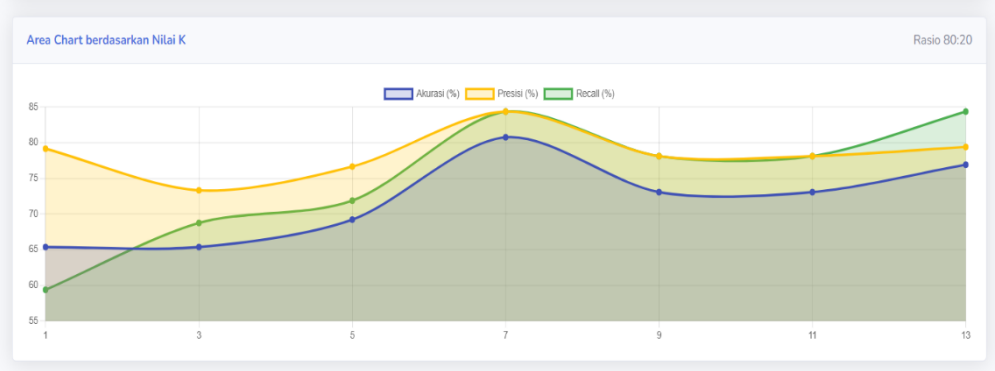
Hasil dari Tabel 8 mengindikasikan bahwa K tetangga terdekat adalah data latih 2. Dikarenakan nilai yang dipilih yaitu 1 sehingga tidak ada perbandingan dari data latih lainnya. Maka hasil pengujian pada data uji 1 diberi label positif.

3.2 Pengujian

Pada penelitian ini, dilakukan pengujian untuk mengevaluasi kinerja algoritme *K-Nearest Neighbor* (K-NN) dalam menentukan label kelas pada data uji. Pengujian dilakukan dengan menggunakan beberapa percobaan rasio data latih dan data uji, yaitu 80% data latih dan 20% data uji, 75% data latih dan 25% data uji, serta 90% data latih dan 10% data uji. Selain itu, juga dilakukan perbandingan hasil klasifikasi dengan menggunakan berbagai nilai *k* yang berbeda, seperti *k*=1, *k*=3, *k*=5, *k*=7, *k*=9, *k*=11, dan *k*=13. Tujuan dari perbandingan ini adalah untuk mengevaluasi bagaimana performa metode *K-Nearest Neighbor* (K-NN) berubah dengan variasi nilai *k* dan rasio data yang berbeda. Berikut adalah hasil pengujian untuk setiap nilai *k* pada 3 rasio yang tampilkan dalam bentuk grafik *k*:

a. Grafik nilai *k* pada rasio 80:20

Hasil pengujian pada rasio 80:20 menunjukkan bahwa nilai akurasi terbesar ada pada saat nilai *k*=7 seperti yang terlihat pada Gambar 3 Berikut:



Gambar 3. Grafik pengujian rasio 80:20

Berdasarkan Gambar 3 terlihat grafik hasil perhitungan akurasi, presisi dan *recall* berdasarkan rasio dataset 80:20 dengan 208 data latih dan 52 data uji. Grafik tersebut menunjukkan akurasi terbesar saat nilai *k*=7 sedangkan akurasi terendah saat nilai *k*=1 dan 3. Karena akurasi sendiri diukur dengan membagi jumlah prediksi label yang sesuai dengan label aktualnya berdasarkan nilai *k* masing-masing dengan jumlah data uji pada rasio yang digunakan. Maka akurasi tertinggi dan terendah ini dipengaruhi oleh jumlah pembagian dataset pada rasio 80:20 serta jumlah tetangga (*k*) yang digunakan.

b. Grafik nilai *k* pada rasio 75:25

Hasil pengujian pada rasio 75:25 menunjukkan bahwa nilai akurasi terbesar ada pada saat nilai *k*=1 seperti yang terlihat pada Gambar 4 Berikut:

Gambar 4. Grafik pengujian rasio 75:25

Berdasarkan Gambar 4 terlihat grafik hasil perhitungan akurasi, presisi dan *recall* berdasarkan rasio dataset 75:25 dengan 195 data latih dan 65 data uji. Akurasi tertinggi terjadi saat nilai *k*=1 sedangkan akurasi terendah saat nilai *k*=3 dan 5. Karena akurasi sendiri diukur dengan membagi jumlah prediksi label yang sesuai dengan label aktualnya berdasarkan nilai *k* masing-masing dengan jumlah data uji pada rasio yang digunakan. Maka akurasi tertinggi dan terendah ini dipengaruhi oleh jumlah pembagian dataset pada rasio 75:25 serta jumlah tetangga (*k*) yang digunakan.

c. Grafik nilai *k* pada rasio 90:10

Hasil pengujian pada rasio 90:10 menunjukkan bahwa nilai akurasi terbesar ada pada saat nilai *k*=1 seperti yang terlihat pada Gambar 5 Berikut:

Gambar 5. Grafik pengujian rasio 90:10

Pada Gambar 5 menunjukkan grafik hasil perhitungan akurasi, presisi dan *recall* berdasarkan rasio dataset 75:25 dengan 234 data latih dan 26 data uji. Akurasi tertinggi terjadi saat nilai *k*=1 sedangkan akurasi terendah saat nilai *k*=11. Karena perhitungan akurasi dengan cara membagi jumlah prediksi label yang sesuai dengan label aktualnya berdasarkan nilai *k* masing-masing dengan jumlah data uji pada rasio yang digunakan. Maka akurasi tertinggi dan terendah ini dipengaruhi oleh jumlah pembagian dataset pada rasio 90:10 serta jumlah tetangga (*k*) yang digunakan.

Kemudian disajikan juga nilai akurasi, presisi, dan *recall* dari semua nilai *k* pada semua rasio yang ada pada Tabel 9 berikut:

Tabel 9. Hasil akurasi, presisi dan *recall* semua nilai *k*

		Nilai *k*						
		k =1	*k* =3	*k* =5	*k* =7	*k* =9	*k* =11	*k* =13
Rasio 80:20	Akurasi	65,38%	65,38%	69,23%	80,77%	73,08%	73,08%	76,92%
	Presisi	79,17%	73,33%	76,67%	84,38%	78,13%	78,13%	79,41%
	Recall	59,38%	68,75%	71,88%	84,38%	78,13%	78,13%	84,38%

Rasio 75:25	Akurasi	73,85%	64,62%	64,62%	69,23%	67,69%	69,23%	66,15%
	Presisi	82,35%	72,22%	72,22%	75,68%	72,50%	73,17%	71,79%
	Recall	71,79%	66,67%	66,67%	71,79%	74,36%	76,92%	71,79%
Rasio 90:10	Akurasi	73,08%	61,54%	69,23%	65,38%	61,54%	57,69%	61,54%
	Presisi	84,62%	71,43%	75%	70,59%	68,74%	64,71%	68,75%
	Recall	68,75%	62,50%	75%	75%	68,75%	68,75%	68,75%

Setelah melakukan pengujian pada berbagai variasi nilai *k* dan rasio dataset, ditemukan nilai *k* terbaik untuk setiap rasio seperti yang tercantum pada Tabel 10 berikut:

Tabel 10. Hasil akurasi, presisi dan *recall* terbaik

Rasio	Nilai *k*	Akurasi	Presisi	*Recall*
80:20	7	80,77%	84,38%	84,38%
75:25	1	73,85%	82,35%	71,79%
90:10	1	73,08%	84,62%	68,75%

Berdasarkan Tabel 10, dapat disimpulkan bahwa dalam pengujian menggunakan algoritme *K-Nearest Neighbor* (K-NN), nilai *k* terbaik untuk mencapai hasil tertinggi adalah *k*=7 pada rasio 80:20. Pada pengujian ini, nilai *k*=7 pada rasio 80:20 memberikan akurasi paling optimal. Hal ini disimpulkan setelah melakukan pengujian dengan berbagai rasio dan nilai *k* yang berbeda, kemudian membandingkan hasil-hasil pengujian tersebut. Perbandingan dilakukan dengan mempertimbangkan akurasi tertinggi yang dihitung menggunakan rumus pada persamaan 5 subbab 2.6. Tinggi atau rendahnya akurasi dipengaruhi oleh berbagai faktor, salah satunya adalah pembagian dataset pada rasio dan jumlah tetangga (*k*) yang digunakan. Pada nilai *k*=7 tersebut, model klasifikasi berhasil mencapai akurasi sebesar 80,77%. Hasil ini menunjukkan bahwa model K-NN memberikan performa terbaik dalam memprediksi klasifikasi data dengan menggunakan tetangga terdekat (*k*=7) untuk metode K-NN.

4. KESIMPULAN

Berdasarkan pada hasil penelitian ini, menunjukkan keberhasilan dalam menerapkan metode klasifikasi teks untuk menganalisis opini masyarakat di Twitter mengenai hilirisasi industri. Data diambil dari Twitter selama periode 01 Mei 2023 sampai 30 Juni 2023 dengan total data yang digunakan yaitu 260 data tweet dengan menggunakan *library Tweepy* untuk *crawling* datanya dan kemudian dilakukan pembobotan kata menggunakan metode TF-IDF. Selanjutnya, metode *K-Nearest Neighbor* (K-NN) digunakan untuk mengklasifikasikan data dan menghasilkan nilai akurasi, presisi, dan *recall* sebagai evaluasi performa model klasifikasi. Hasil terbaik dicapai pada rasio dataset 80:20, dengan akurasi mencapai 80,77% dan nilai *k*=7. Penelitian ini juga mengidentifikasi bahwa beberapa faktor, seperti jumlah data dan fitur pembobotan, dapat mempengaruhi kinerja algoritme.

Adapun sebagai pertimbangan untuk penelitian selanjutnya dapat dilakukan dengan menggabungkan metode lain guna meningkatkan kualitas hasil yang dicapai. Selain itu, menggunakan pembobotan kata yang lebih ringkas agar dapat meningkatkan kinerja dan efisiensi pemrosesan data, sehingga sistem dapat berjalan lebih cepat dan optimal. Selanjutnya, penambahan jumlah data yang digunakan diharapkan dapat memperbaiki akurasi dan juga menambah keberagaman dataset yang digunakan dalam analisis sentimen opini masyarakat di Twitter mengenai isu hilirisasi. Dengan demikian, kombinasi langkah-langkah tersebut diharapkan dapat menghasilkan sistem yang lebih efektif dan akurat dalam melakukan klasifikasi sentimen.

DAFTAR PUSTAKA

[1] S. Kemp, "Digital 2023: Indonesia — DataReportal – Global Digital Insights." https://datareportal.com/reports/digital-2023-indonesia (accessed Jul. 03, 2023).

[2] M. Agung and E. A. W. Adi, "Peningkatan Investasi Dan Hilirisasi Nikel Di Indonesia," *JISIP (Jurnal Ilmu Sos. dan Pendidikan)*, vol. 6, no. 2, pp. 4009–4020, 2022, doi: 10.58258/jisip.v6i2.3085.

[3] T. T. Widowati and M. Sadikin, "Analisis Sentimen Twitter terhadap Tokoh Publik dengan Algoritma Naive Bayes dan Support Vector Machine," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 11, no. 2, pp. 626–636, 2021, doi: 10.24176/simet.v11i2.4568.

[4] R. Prabowo, H. Sujaini, T. Rismawan, J. H. Rekayasa Sistem Komputer Universitas Tanjungpura Jl Hadari Nawawi, K. Barat, and J. H. Hadari Nawawi, "Analisis Sentimen Pengguna Twitter Terhadap Kasus COVID-19 di Indonesia Menggunakan Metode Regresi Logistik Multinomial The Sentiment Analysis of Twitter User about COVID-19 Cases in Indonesia Using the Multinomial Logistics Regression Method," vol. 11, no. 1, pp. 85–90, 2023, doi: 10.26418/justin.v11i1.57450.

[5] W. Afifi, "Analisis sentimen pengguna twitter terhadap layanan internet pt indosat tbk dengan metode k-nearest neighbor (k-nn) dan naive bayes classifier (nbc)," *Repository.Uinjkt.Ac.Id*, 2022, [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/67033%0Ahttps://repository.uinjkt.ac.id/dspace/bitstream/123456789/67033/1/WAKHID AFIFI-FST.pdf

[6] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 121, 2021, doi: 10.22146/ijccs.65176.

[7] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 1–11, 2020, doi: 10.21107/edutic.v7i1.8779.

[8] M. Fadhillah, "Tugas Akhir," *J. Ekon. Vol. 18, Nomor 1 Maret201*, vol. 2, no. 1, pp. 41–49, 2020.

[9] S. M. Sari, "Analisis Sentimen Terhadap New Normal Di Era Covid-19 Menggunakan Algoritma K-Nearest Neighbor (K-NN)," pp. 1–80, 2021, [Online]. Available: http://repository.uinsu.ac.id/id/eprint/14945%0Ahttp://repository.uinsu.ac.id/14945/1/SKRIPSI Susan Mayang Sari %28ILKOMP NIM. 0701162003%29.pdf

[10] S. S. Utami, "Analisis Sentimen Pengguna Twitter Mengenai 'Sedotan Plastik' Dengan Metode K-Nearest Neighbor (KNN) Dan Neighbor-Weighted K-Nearest Neighbor …," 2019, [Online]. Available: https://repository.its.ac.id/64038/1/06211540000005-Undergraduate_Theses.pdf

[11] M. Furqan, S. Sriani, and S. M. Sari, "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia," *Techno.Com*, vol. 21, no. 1, pp. 51–60, 2022, doi: 10.33633/tc.v21i1.5446.

[12] A. Yudhana, S. Sunardi, and A. J. S. Hartanta, "Algoritma K-Nn Dengan Euclidean Distance Untuk Prediksi Hasil Penggergajian Kayu Sengon," *Transmisi*, vol. 22, no. 4, pp. 123–129, 2020, doi: 10.14710/transmisi.22.4.123-129.

[13] T. S. Sabrila, V. R. Sari, and A. E. Minarno, "Analisis Sentimen Pada Tweet Tentang Penanganan Covid-19 Menggunakan Word Embedding Pada Algoritma Support Vector Machine Dan K-Nearest Neighbor," *Fountain Informatics J.*, vol. 6, no. 2, p. 69, 2021, doi: 10.21111/fij.v6i2.5536.