5th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
7 September 2024 – Jakarta, Indonesia
Volume 3, Nomor 2, September 2024 - ISSN 2962-8628 (*online*)

ALGORITMA *K-NEAREST NEIGHBOUR* PADA ANALISIS SENTIMEN KEPUASAN WARGA X TERHADAP BERITA KOMPAS

Abid Rafi Nur Hastama[1*], Arief Wibowo[2]

[1,2]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta Selatan, Indonesia

Email: [1*]2011501802@student.budiluhur.ac.id, [2]arief.wibowo@budiluhur.ac.id
(* : corresponding author)

Abstrak-Penelitian ini mengkaji sentimen masyarakat Indonesia terhadap berita yang dipublikasikan oleh Kompas, sebuah institusi media massa besar yang dikenal dengan mottonya 'terpercaya'. Meskipun Kompas memiliki reputasi yang kuat, berbagai pendapat negatif dari masyarakat tentang berita yang mereka terbitkan tersebar di berbagai *platform* media sosial seperti *Instagram*, *Facebook*, dan X. Tantangan utama bagi publik adalah menentukan apakah tanggapan terhadap berita yang diterbitkan oleh Kompas bersifat netral, positif, atau negatif. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap berita Kompas menggunakan algoritma *K-Nearest Neighbour* (KNN) karena algoritma tersebut cocok digunakan pada *dataset* yang sedikit. Data untuk penelitian ini dikumpulkan dari 513 *tweets* yang diambil dari X antara 16 Februari hingga 5 April 2024. Setelah melalui tahap *preprocessing*, jumlah data yang dianalisis berkurang menjadi 403 *tweets*. Kata kunci yang digunakan dalam proses *crawling* adalah 'berita kompas'. Dalam analisis sentimen, algoritma KNN yang dipadukan dengan pembobotan kata TF-IDF menunjukkan hasil yang cukup baik dengan akurasi 86%, *precision* 86%, *recall* 97%, dan *f1-score* 90% pada data *training*. Untuk data *testing*, hasil pengujian menunjukkan akurasi 86%, *precision* 86%, *recall* 100%, dan *f1-score* 93%. Analisis lebih lanjut mengungkapkan bahwa mayoritas sentimen terhadap berita Kompas bersifat netral dengan persentase 97.77%, sementara sentimen positif hanya 1.49%, dan sentimen negatif sebesar 0.74%. Kesimpulannya, mayoritas masyarakat Indonesia di X memiliki pandangan netral terhadap berita yang dipublikasikan oleh Kompas selama periode penelitian.

Kata Kunci: *Text Mining*, Analisis Sentimen, Kompas, TF-IDF

K-NEAREST NEIGHBOUR ALGORITHM ON SENTIMENT ANALYSIS OF X RESIDENTS' SATISFACTION WITH KOMPAS NEWS

Abstract-*This study examines the sentiment of the Indonesian public towards news published by Kompas, a major mass media institution known for its motto 'trustworthy'. Despite Kompas's strong reputation, various negative opinions from the public about the news they publish have spread across social media platforms such as Instagram, Facebook, and X. The main challenge for the public is determining whether the response to news published by Kompas is neutral, positive, or negative. This research aims to analyze public sentiment towards Kompas news using the K-Nearest Neighbour (KNN) algorithm. The data for this study was collected from 513 tweets retrieved from X between February 16 and April 5, 2024. After preprocessing, the number of tweets analyzed was reduced to 403. The keyword used in the crawling process was 'berita kompas'. In sentiment analysis, the KNN algorithm combined with TF-IDF word weighting showed promising results, with an accuracy of 86%, precision of 86%, recall of 97%, and an f1-score of 90% on the training data. For the testing data, the results showed an accuracy of 86%, precision of 86%, recall of 100%, and an f1-score of 93%. Further analysis revealed that the majority of sentiments towards Kompas news were neutral, with a percentage of 97.77%, while positive sentiment accounted for only 1.49%, and negative sentiment was 0.74%. In conclusion, the majority of the Indonesian public on X has a neutral view of the news published by Kompas during the study period.*

Keywords*: Text Mining, Sentiment Analysis, Kompas, TF-IDF*

1. PENDAHULUAN

Kemajuan teknologi informasi dan komunikasi, terutama media sosial, telah memberikan pengaruh yang besar kepada kita sebagai penggunanya, termasuk dalam hal menyebarkan informasi dan juga pendapat pribadi. X, sebagai salah satu *platform* media sosial terkemuka menjadi tempat dimana penggunanya dapat mengekspresikan atau berbagi cerita tentang apapun, seperti reaksi terhadap berita yang dikeluarkan oleh suatu media *online*. Salah satu dari banyaknya media *online* yang terkemuka di Indonesia adalah Kompas. Kompas menyediakan berita terkini dari dalam negeri maupun luar negeri, serta liputan mendalam di berbagai topik seperti ekonomi, politik, sosial, budaya, dan yang lainnya. Walaupun Kompas media yang terkemuka dan besar, perlu untuk memahami bagaimana persepsi dan sentimen masyarakat mengenai berita yang dikeluarkan oleh Kompas, khususnya melalui media sosial X.

Analisis Sentimen adalah mengekstrak opini publik tentang topik tertentu, produk atau jasa yang di dalam nya terkandung teks-teks yang tidak terstruktur [1]. Metode *K-Nearest Neighbour* digunakan untuk proses klasifikasi dalam sebuah penelitian karena memiliki kesederhanaan dimana prosesnya berdasarkan pada pendekatan pembobotan yang sederhana dan kemudahan dalam implementasi, adaptasi dan proses *learning* serta memiliki nilai akurasi yang tinggi [2]. KNN menghitung kedekatan antara kasus baru (*data test*) dan kasus lama (*data training*) yang berdasarkan pencocokan bobot dari sejumlah fitur yang ada. Setelah jarak terhitung, maka jarak yang paling dekat dengan data latih dianggap memiliki kesamaan [3].

Proses dalam analisis sentimen ini selain menerapkan algoritma *K-Nearest Neighbour* (KNN) juga memanfaatkan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF merupakan metode untuk menentukan frekuensi relatif dari setiap kata atau token di mana setiap kata tersebut akan diberikan pembobotan berupa nilai berdasarkan penting atau tidaknya suatu kata dalam dokumen berdasarkan jumlah kemunculan kata dalam suatu dokumen dan mengukur kata-kata tersebut terhadap keseluruhan dokumen yang ada [4].

Selain itu terdapat beberapa penelitian sebelumnya tentang analisis sentimen, yaitu "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN" [1]. Penelitian ini mengolah *dataset* sebanyak 998 data untuk aplikasi Bareksa (484 sentimen positif dan 514 sentimen negatif) dan 1063 data untuk aplikasi Bibit (514 sentimen positif dan 522 sentimen negatif). Menggunakan model CRISP-DM (*Cross Industry Standard Process for Data Mining*) dan algoritma KNN, diperoleh akurasi, *precision*, dan *recall* masing-masing sebesar 85,14%, 91,91%, dan 76,44% untuk Bibit, serta 81,70%, 87,15%, dan 75,73% untuk Bareksa.

Selanjutnya, untuk penelitian "Penerapan Algoritma K-*Nearest Neighbors* Pada Analisis Sentimen Metode Pembelajaran Dalam Jaringan (DARING) Di Universitas Kristen Wira Wacana Sumba" [5]. Pada penelitian ini *dataset* diperoleh dari komentar grup *facebook* dan didapatkan data sebanyak 200. Kemudian hasil dari penelitian tersebut didapatkan akurasi yang baik sebesar 87.00% dan nilai AUC yang cukup besar, yaitu 0.19 dari 26 data *review* positif dan 174 *review* negatif.

Pada penelitian "Implementasi Metode K-*Nearest Neighbor* (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP" [3], *dataset* berupa ulasan atau *review* yang diberikan oleh pengguna FLIP dari *Google Play Store* diperoleh dengan teknik *scraping* dan didapatkan data sebanyak 11.002. Hasil penelitian dengan algoritma KNN bersama dengan pembobotan TF-IDF, didapatkan akurasi sebesar 76.68% dan 77.67% dari data uji telah benar terklasifikasi ke dalam kelas ulasan positif dengan nilai *precision* dan *recall* yang sangat tinggi yaitu sebesar 82.67% dan 86.92%.

Penelitian ini diharapkan dapat memberikan kontribusi dalam pemahaman tentang hubungan antara media *online* dan respons masyarakat di media sosial, serta relevansinya dalam konteks analisis sentimen. Selain itu, diharapkan hasil penelitian ini dapat memberikan kontribusi untuk Kompas sebagai bahan evaluasi terhadap berita-berita yang telah dikeluarkan.

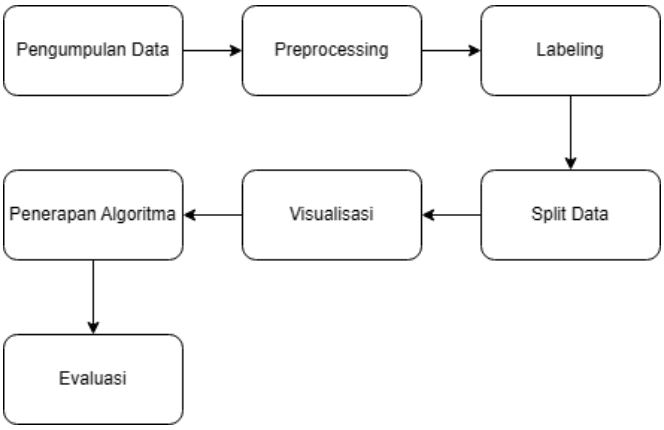
2. METODE PENELITIAN

2.1 Data Penelitian

Pada hari Jumat, 5 April 2024 dilakukan *crawling* yang memperoleh total 513 *tweets* dari rentang tanggal 16 Februari hingga 5 April 2024. Berdasarkan dari data tersebut akan diimplementasikan teknik *text mining* menggunakan metode KNN dengan TF-IDF untuk menemukan hasil dari sentimen warga X terkait berita kompas. Data ini di*crawling* menggunakan aplikasi *tweet-harvest* yang tersedia di GitHub (https://github.com/helmisatria/tweet-harvest).

2.2 Penerapan Metode

Penelitian ini menggunakan teknik *text mining* dengan beberapa tahapan yang di jelaskan di gambar 1 untuk menghasilkan hasil akurasi yang baik, dalam kasus ini peneliti menggunakan algoritma KNN. Pada tahapan pertama ini, peneliti melakukan *crawling* data yang bersumber dari sosial media X yang dilaksanakan pada hari Jumat, 5 April 2024. Kemudian untuk *keyword* peneliti menggunakan kata 'berita kompas' dan mendapatkan data sebanyak 513 *record*. Lalu, data tersebut di *export* ke csv dan dilakukan *convert* ke xlsx.



Gambar 1. Tahapan Keseluruhan Metode Dari Penelitian

Selanjutnya di tahapan kedua adalah *preprocessing*. Proses *text mining* memerlukan tahap *preprocessing* yang dilakukan setelah pengumpulan data untuk memproses data agar data bisa diolah ke tahap yang berikutnya. Berikut adalah langkah-langkah dari tahap *preprocessing:*

a. *Filtering*
   Dalam proses *filtering* ini, tweet dengan *username* seperti @KompasTV dan yang bersangkutan dihapus dikarenakan penelitian ini bertujuan untuk meneliti sentimen warga X terhadap berita kompas, dan akun dengan *username* tersebut adalah subjek dari penelitian ini.
b. *Case Folding*
   Dalam proses ini, teks dibuat menjadi huruf kecil (*lowercase*) secara keseluruhan. Seperti contoh, kata ‘Saya’ atau ‘SAYA’ akan diubah menjadi ‘saya’.
c. *Cleansing*
   Setelah data telah terkumpul, tahap pertama untuk *preprocessing* adalah *cleansing*. *Cleansing* adalah pembersihan data dari suatu simbol-simbol yang tidak diperlukan, seperti karakter spesial, menghapus emoji, menghapus *link*, menghapus *mention*, menghapus *hashtag*, dan menghapus karakter spsial / angka.
d. Tokenisasi
   Setelah pembersihan data, dilakukan tokenisasi untuk memisahkan kata dalam kalimat guna memudahkan penggalian makna. Contoh: "kompas hari ini beritanya oke banget" menjadi ['kompas', 'hari', 'ini', 'beritanya', 'oke', 'banget'].
e. Normalisasi
   Tahap selanjutnya setelah tokenisasi adalah normalisasi data, yaitu mengubah kata informal menjadi formal untuk memudahkan pengolahan. Contoh: 'kompas, apa, gak, berita, yg, lain' diubah menjadi 'kompas apa tidak berita yang lain'. Peneliti menggunakan dataset kamus informal dan formal dari *repository* GitHub (https://github.com/louisowen6/NLP_bahasa_resources?tab=readme-ov-file#stop-words).
f. *Stopword Removal*
   Setelah melakukan normalisasi data, tahap selanjutnya adalah *stopword removal*. Tahap ini adalah tahap untuk menghilangkan kata yang tidak diperlukan agar pengolahan data menjadi lebih optimal. Sebagai contoh kalimat ini ‘baru tahu hari ini skripsiku’ akan diubah menjadi seperti ini ‘baru tahu hari skripsiku’. Jika dilihat, kata ‘ini’ dihilangkan dari kalimat tersebut karena kata tersebut tidak perlukan untuk pengolahan data.
g. *Stemming*
   Tahap terakhir adalah *stemming* data, yaitu mereduksi kata berimbuhan menjadi kata dasar untuk memudahkan analisis. Contoh: 'cara berpikir orang baca kompas liputan berita' diubah menjadi 'cara pikir orang baca kompas liput berita'.

Kemudian tahapan selanjutnya adalah *labeling*. Penelitian yang dilakukan saat ini menggunakan kamus *Vader Lexicon* untuk tahap *labeling*. Hasil *compound* dapat diklasifikasikan menjadi kategori positif, netral, atau negatif berdasarkan kriteria yakni jika nilai *compound* lebih dari 0,5, maka dimasukkan ke dalam kategori positif dengan angka 1, jika nilai *compound* antara -0,5 dan 0,5, maka dimasukkan ke dalam kategori netral dengan angka 0 dan jika nilai *compound* di bawah -0,5, maka dimasukkan ke dalam kategori negatif dengan angka -1 [6].

Pada tahap *split* data, tweet yang sudah diberi label akan dipecah menjadi dua bagian: *training* data dan *testing* data. Proses ini menggunakan teknik *stratified sampling*. *Stratified sampling* merupakan metode pengambilan sampel proporsional yang didasarkan pada strata atau tingkatan data populasi secara proporsional.

Metode ini termasuk dalam kelompok *probability sampling* di mana setiap bagian dari populasi memiliki kesempatan yang sama untuk dipilih, sehingga sampel yang dipilih dapat mewakili populasi secara keseluruhan [7]. Penelitian ini menghitung persentase label positif, negatif, dan netral untuk setiap *tweet*, lalu membagi dataset menjadi 80% *training* data dan 20% *testing* data dengan persentase label yang sama.

Setelah *labeling*, visualisasi diperlukan untuk mengetahui berapa perbandingan dari setiap label, kemudian untuk mengetahui kata mana yang sering muncul, kata positif, dan juga kata negatif. Peneliti menggunakan Wordcloud untuk visualisasi pada penelitian kali ini. Setelah melalui tahapan *labeling* dan juga *split* data, data harus berbentuk numerik. Untuk mengubah data tersebut ke bentuk numerik dibutuhkan metode pembobotan yang bernama TF-IDF. Metode *Term Frequency Invers Document Frequency* (TF-IDF) merupakan metode yang digunakan menentukan seberapa jauh keterhubungan kata (term) terhadap dokumen dengan memberikan bobot setiap kata [8]. Metode ini menghitung bobot berdasarkan frekuensi kemunculan kata dalam dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut, menghasilkan bobot tinggi untuk kata yang sering muncul dalam satu dokumen tetapi jarang ditemukan di seluruh koleksi dokumen. Rumusnya adalah sebagai berikut:

$$\text{tf(t)} = \frac{f_{t,d}}{\max(tf)} \quad (1)$$

$$\text{idf}_t = \log\left(\frac{D}{dft}\right) \quad (2)$$

$$\text{TF-IDF} = tf_t \; x \; IDF_t \quad (3)$$

Keterangan:

tf	=	*Term Frequency*
$f_{t,d}$	=	Jumlah kata yang dicari dalam sebuah dokumen
max(tf)	=	Jumlah kemunculan terbanyak *term* pada dokumen yang sama
Nilai D	=	Jumlah keseluruhan dokumen
df_t	=	Jumlah dokumen yang mengandung *term* t
IDF	=	*Inversed Document Frequency*
d	=	Dokumen ke-d
t	=	Kata ke-t dari kata kunci
TF-IDF	=	*Term Frequency Invers Document Frequency*

K-Nearest Neighbour (KNN) adalah metode populer untuk klasifikasi dan regresi dalam *machine learning*. Dalam analisis sentimen, KNN mengklasifikasikan teks ke kategori seperti positif, negatif, atau netral berdasarkan kesamaan dengan dokumen yang telah diklasifikasikan sebelumnya. Dokumen dianggap sebagai titik dalam ruang fitur, dan KNN mencari K dokumen terdekat menggunakan jarak *euclidean* untuk menentukan kategori sentimen mayoritas sebagai prediksi. Berikut adalah rumus *Euclidean distance* untuk menentukan tetangga terdekat [9].

$$d(i)\sqrt{\sum_{i=1}^{p}(x_i - x_2)^2} \quad (4)$$

Keterangan:

d	=	Jarak
i	=	Variabel data
p	=	Dimensi data
x_1	=	Sampel data
x_2	=	Data uji

Pada tahapan terakhir yaitu evaluasi, diperlukan perhitungan *confusion matrix*. *Confusion matrix* adalah matriks yang menunjukkan kinerja algoritma dalam mengklasifikasi secara visual. Melalui matriks ini dapat diketahui perbandingan antara klasifikasi aktual dengan prediksinya [10]. Kemudian pada tabel 1, hasil dari perhitungan tersebut akan diolah menjadi *confusion matrix* yang membandingkan nilai *evaluation metrics*, yaitu akurasi, *precision*, *recall* dan *f1-score* dari hasil pemrosesan di bagian *training* juga *testing* pada program *web* yang dibuat.

Tabel 1. *Confusion Matrix*

Actual	*Predicted*		
	Negatif	Netral	Positif
Negatif	TP(Neg)	FN(Neg) & FP(Net)	FN(Neg) & FP(Pos)
Netral	FN(Net) & FP(Neg)	TP(Net)	FN(Net) & FP(Pos)
Positif	FN(Pos) & FP(Neg)	FN(Pos) & FP(Net)	TP(Pos)

Keterangan:
True Positive (TP) = Jumlah dokumen dari kelas *actual* dan kelas *predicted* berada di posisi yang sama.
False Negative (FN) = Jumlah baris (*row*) dokumen dari kelas *actual* yang keliru diklasifikasikan sebagai kelas lain.
False Positive (FP) = Jumlah kolom (*column*) dokumen dari kelas *actual* yang keliru diklasifikasikan sebagai kelas lain.

a. Akurasi
Akurasi adalah nilai yang mencerminkan apakah jumlah dokumen yang diprediksi atau diklasifikasikan dengan benar. Rumus yang digunakan adalah sebagai berikut:

$$accuracy = \frac{\sum TP}{Total} \tag{5}$$

b. *Precision*
Precision adalah metrik yang mengukur akurasi prediksi model dalam mengidentifikasi suatu kelas. Ini dihitung sebagai rasio antara prediksi benar (*True Positives*) dan total prediksi untuk kelas tersebut (*True Positives* + *False Positives*). *Precision* keseluruhan dihitung menggunakan metode *weighted*.

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$precision\ (weighted) = \frac{\sum precision * predicted\ class}{Total} \tag{7}$$

c. *Recall*
Recall adalah metrik yang mengukur kemampuan model untuk mendeteksi semua contoh sebenarnya dalam suatu kelas. Ini dihitung sebagai rasio antara prediksi benar (*True Positives*) dan total contoh sebenarnya dalam kelas tersebut (*True Positives* + *False Negatives*). Nilai recall keseluruhan dihitung dengan metode *weighted*.

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$recall\ (weighted) = \frac{\sum recall * predicted\ class}{Total} \tag{9}$$

d. *F1-Score*
F1-Score mengkombinasikan *precision* dan *recall* untuk menilai kinerja model secara menyeluruh, terutama berguna saat ada ketidakseimbangan kelas. *F1-score merupakan* rata-rata harmonis dari *precision* dan *recall*. Setelah didapatkan nilainya, kemudian keseluruhan kelas dihitung dengan perhitungan *weighted*.

$$f1\ score = 2 * \frac{precision + recall}{precision + recall} \tag{10}$$

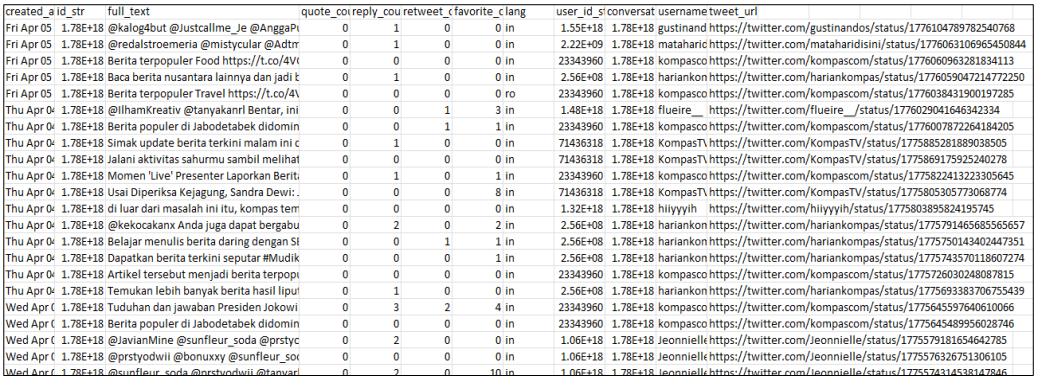
$$f1\ score\ (weighted) = \frac{\sum f1\ score * predicted\ class}{Total} \tag{11}$$

3. HASIL DAN PEMBAHASAN

Penelitian ini terdiri dari enam tahapan utama dalam mengimplementasikan metode yang digunakan. Tahapan-tahapan tersebut dilakukan secara berurutan seperti yang sudah ditampilkan pada gambar 1. Tahapan tersebut meliputi: Pengumpulan data (crawling), *preprocessing*, *labeling*, *split* data, *modeling* (penggunaan TF-IDF dengan algoritma KNN), dan yang terakhir *evaluate*. Penjelasan terperincinya adalah sebagai berikut.

3.1 Pengumpulan Data (*Crawling*)

Gambar 2 merupakan hasil dari *crawling* yang diperoleh dari X dalam bentuk *tweet*. *Crawling* dilakukan pada hari Jumat, 5 April 2024 dan didapatkan *tweet* sebanyak 513 *records* dari rentang tanggal 16 Februari hingga 5 April 2024. Kata kunci yang digunakan pada saat crawling adalah 'berita kompas'.



created_a	id_str	full_text	quote_co	reply_cou	retweet_c	favorite_c	lang	user_id_s	conversat	username	tweet_url
Fri Apr 05	1.78E+18	@kalog4but @Justcallme_Je @AnggaP	0	1	0	0	in	1.55E+18	1.78E+18	gustinand	https://twitter.com/gustinandos/status/1776104789782540768
Fri Apr 05	1.78E+18	@redalstroemeria @mistycular @Adtr	0	1	0	0	in	2.22E+09	1.78E+18	mataharic	https://twitter.com/mataharidisini/status/1776063106965450844
Fri Apr 05	1.78E+18	Berita terpopuler Food https://t.co/4V	0	0	0	0	in	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1776060963281834113
Fri Apr 05	1.78E+18	Baca berita nusantara lainnya dan jadi k	0	1	0	0	in	2.56E+08	1.78E+18	hariankon	https://twitter.com/hariankompas/status/1776059047214772250
Fri Apr 05	1.78E+18	Berita terpopuler Travel https://t.co/4\	0	0	0	0	ro	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1776038431900197285
Thu Apr 04	1.78E+18	@IlhamKreativ @tanyakanrl Bentar, ini	0	0	1	3	in	1.48E+18	1.78E+18	flueire__	https://twitter.com/flueire__/status/1776029041646342334
Thu Apr 04	1.78E+18	Berita populer di Jabodetabek didomin	0	0	1	1	in	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1776007872264184205
Thu Apr 04	1.78E+18	Simak update berita terkini malam ini c	0	1	0	0	in	71436318	1.78E+18	KompasTV	https://twitter.com/KompasTV/status/1775885281889038505
Thu Apr 04	1.78E+18	Jalani aktivitas sahurmu sambil melihat	0	0	0	0	in	71436318	1.78E+18	KompasTV	https://twitter.com/KompasTV/status/1775869175925240278
Thu Apr 04	1.78E+18	Momen 'Live' Presenter Laporkan Berita	0	1	0	1	in	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1775822413223305645
Thu Apr 04	1.78E+18	Usai Diperiksa Kejagung, Sandra Dewi:	0	0	0	8	in	71436318	1.78E+18	KompasTV	https://twitter.com/KompasTV/status/1775805305773068774
Thu Apr 04	1.78E+18	di luar dari masalah ini itu, kompas tem	0	0	0	0	in	1.32E+18	1.78E+18	hiiyyyih	https://twitter.com/hiiyyyih/status/1775803895824195745
Thu Apr 04	1.78E+18	@kekocakanx Anda juga dapat bergabu	0	2	0	2	in	2.56E+08	1.78E+18	hariankon	https://twitter.com/hariankompas/status/1775791465685565657
Thu Apr 04	1.78E+18	Belajar menulis berita daring dengan SI	0	0	1	1	in	2.56E+08	1.78E+18	hariankon	https://twitter.com/hariankompas/status/1775750143402447351
Thu Apr 04	1.78E+18	Dapatkan berita terkini seputar #Mudik	0	0	0	1	in	2.56E+08	1.78E+18	hariankon	https://twitter.com/hariankompas/status/1775743570118607274
Thu Apr 04	1.78E+18	Artikel tersebut menjadi berita terpopu	0	0	0	0	in	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1775726030248087815
Thu Apr 04	1.78E+18	Temukan lebih banyak berita hasil liputl	0	1	0	0	in	2.56E+08	1.78E+18	hariankon	https://twitter.com/hariankompas/status/1775693383706755439
Wed Apr 0	1.78E+18	Tuduhan dan jawaban Presiden Jokowi	0	3	2	4	in	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1775645597640610066
Wed Apr 0	1.78E+18	Berita populer di Jabodetabek didomin	0	0	0	0	in	23343960	1.78E+18	kompasco	https://twitter.com/kompascom/status/1775645489956028746
Wed Apr 0	1.78E+18	@JavianMine @sunfleur_soda @prstyc	0	2	0	0	in	1.06E+18	1.78E+18	Jeonnielle	https://twitter.com/Jeonnielle/status/1775579181654642785
Wed Apr 0	1.78E+18	@prstyodwii @bonuxxy @sunfleur_soc	0	0	0	0	in	1.06E+18	1.78E+18	Jeonnielle	https://twitter.com/Jeonnielle/status/1775576326751306105
Wed Apr 0	1.78E+18	@sunfleur_soda @prstyodwii @tanyakl	0	2	0	10	in	1.06E+18	1.78E+18	Jeonnielle	https://twitter.com/Jeonnielle/status/1775574314538147846

Gambar 2. Data Hasil *Crawling* Yang Didapatkan Menggunakan Tweet-Harvest

3.2 *Preprocessing*

Pada tabel 2 menjelaskan tahapan *preprocessing* yang dilakukan setelah pengumpulan data untuk memproses data agar data bisa diolah ke tahap yang berikutnya.

Tabel 2. Tahapan *Preprocessing*

Tahap	Teks awal	Teks Akhir
Filtering	* Tweet dengan username 'hariankompas', 'kompascom', 'kompastv', 'kompasbola', 'kompasmuda', 'kompasklasika', 'kompasdata', 'kompastvjatim' *	-
Case Folding	@PebriyanDwi_P @rfzr007 @KompasTV Soal posisi Yusril ya itu urusan Yusril. Saya meluruskan judul berita Kompas yang sumir soal sumpah sebagai saksi di pengadilan. Judul yang menggiring bahkan menjebak pembaca. https://t.co/C31kHVe9rk	@pebriyandwi_p @rfzr007 @kompastv soal posisi yusril ya itu urusan yusril. saya meluruskan judul berita kompas yang sumir soal sumpah sebagai saksi di pengadilan. judul yang menggiring bahkan menjebak pembaca. https://t.co/c31khve9rk
Cleansing	@pebriyandwi_p @rfzr007 @kompastv soal posisi yusril ya itu urusan yusril. saya meluruskan judul berita kompas yang sumir soal sumpah sebagai saksi di pengadilan. judul yang menggiring bahkan menjebak pembaca. https://t.co/c31khve9rk	soal posisi yusril ya itu urusan yusril saya meluruskan judul berita kompas yang sumir soal sumpah sebagai saksi di pengadilan judul yang menggiring bahkan menjebak pembaca
Tokenisasi	soal posisi yusril ya itu urusan yusril saya meluruskan judul berita kompas yang sumir soal sumpah sebagai saksi di pengadilan judul yang menggiring bahkan menjebak pembaca	soal, posisi, yusril, ya, itu, urusan, yusril, saya, meluruskan, judul, berita, kompas, yang, sumir, soal, sumpah, sebagai, saksi, di, pengadilan, judul, yang, menggiring, bahkan, menjebak, pembaca
Normalisasi	soal, posisi, yusril, ya, itu, urusan, yusril, saya, meluruskan, judul, berita, kompas, yang, sumir, soal, sumpah, sebagai, saksi, di, pengadilan, judul, yang, menggiring, bahkan, menjebak, pembaca	soal posisi yusril iya itu urusan yusril saya meluruskan judul berita kompas yang sumir soal sumpah sebagai saksi di pengadilan judul yang menggiring bahkan menjebak pembaca
Stopword Removal	soal posisi yusril iya itu urusan yusril saya meluruskan judul berita kompas yang sumir soal sumpah sebagai saksi di pengadilan judul yang menggiring bahkan menjebak pembaca	soal posisi yusril iya urusan yusril meluruskan judul berita kompas sumir soal sumpah saksi pengadilan judul menggiring bahkan menjebak pembaca
Stemming	soal posisi yusril iya urusan yusril meluruskan judul berita kompas sumir soal sumpah saksi pengadilan judul menggiring bahkan menjebak pembaca	soal posisi yusril iya urus yusril lurus judul berita kompas sumir soal sumpah saksi adil judul giring bahkan jebak baca

3.3 *Labeling*

Setelah *preprocessing* dilakukan, data terpangkas dari 513 menjadi 403 dikarenakan tahapan *filtering* yang menghapus *tweet* dengan *username* yang ditentukan. Kemudian, pada tabel 3 di bawah ini merupakan hasil dari proses *labeling* otomatis menggunakan kamus *Vader Lexicon* pada data kepuasan warga X terhadap berita yang dikeluarkan oleh Kompas dari hasil *preprocessing* sebelumnya.

Tabel 3. Tahapan *Labeling*

Tweet	Sentimen
kompas tidak *neutral* tidak edukatif itu *block* banyak berita portal lebih *credible*	-1 (Negatif)
kalau berita pondok pesantren al hanifiyyah diri *source* kompas	0 (Netral)
kalau berita benar *courtesy* kompas loh	1 (Positif)

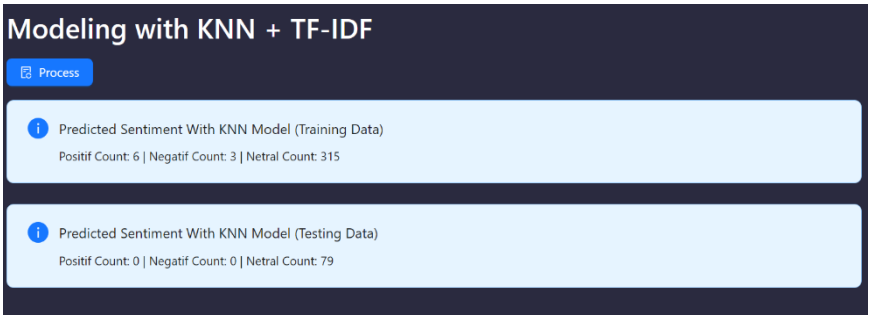
3.4 *Split* Data

Pada gambar 3 yaitu tahapan *split* data, tweet yang sudah diberi label akan dipecah menjadi dua bagian: *training* data dan *testing* data dengan rasio 80:20 menggunakan teknik *stratified sampling*.

Gambar 3. Hasil *split* data dengan rasio 80:20 dengan teknik *stratified sampling*

3.5 Penerapan Algoritma

Gambar 4 merupakan tahapan penerapan algoritma, data yang telah di *split* dilakukan perhitungan TF-IDF dan juga metode KNN dengan nilai K = 1.



Gambar 4. Hasil sentimen yang telah diprediksi menggunakan algoritma KNN + pembobotan TF-IDF

3.6 Evaluasi

Pada tabel 4 dan 5 tahapan evaluasi ini, dari nilai *predicted sentiment* yang didapatkan, kemudian dibentuk *confusion matrix* dan dihitung nilai *evaluation metrics* berdasarkan kriteria atau metrik yang telah ditentukan.

Tabel 4. *Confusion Matrix Training* Data

Actual	*Predicted*		
	Negatif	Netral	Positif
Negatif	2	20	0
Netral	1	272	1
Positif	0	23	5

Tabel 5. *Confusion Matrix Testing* Data

Actual	*Predicted*		
	Negatif	Netral	Positif
Negatif	0	5	0
Netral	0	68	0
Positif	0	6	0

Kemudian, dari *confusion matrix* tersebut dapat dihitung nilai *evaluation metrics*nya, yaitu akurasi, *precision*, *recall*, dan *f1-score*. *Training* data hasil *evaluation metrics*nya adalah sebagai berikut:

$$\textit{Precision (Weighted)} = \frac{(0.6666 * 3) + (0.8634 * 315) + (0.8333 * 6)}{3 + 315 + 6} = 0.8611\ (86\%)$$

$$\textit{Recall (Weighted)} = \frac{(0.0909 * 3) + (0.9927 * 315) + (0.1785 * 6)}{3 + 315 + 6} = 0.9692\ (97\%)$$

$$\textit{F1-Score (Weighted)} = \frac{(0.16 * 3) + (0.9235 * 315) + (0.2941 * 6)}{3 + 315 + 6} = 0.9048\ (90\%)$$

$$\text{Akurasi} = \frac{2 + 272 + 5}{3 + 315 + 6} = 0.8611\ (86\%)$$

Sedangkan, untuk *testing* data hasil *evaluation metrics*nya adalah sebagai berikut:

$$\textit{Precision (Weighted)} = \frac{(0.0 * 0) + (0.8634 * 79) + (0.0 * 0)}{5 + 68 + 6} = 0.8607\ (86\%)$$

$$\textit{Recall (Weighted)} = \frac{(0.0 * 0) + (1.0 * 79) + (0.0 * 0)}{5 + 68 + 6} = 1\ (100\%)$$

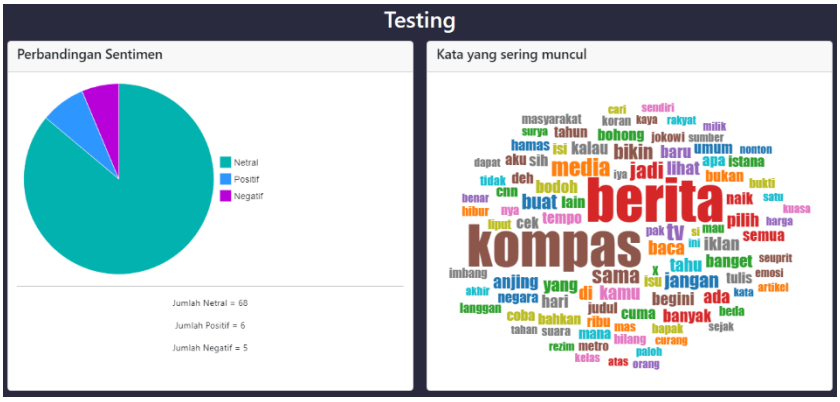
$$\textit{F1-Score (Weighted)} = \frac{(0.0 * 0) + (0.9251 * 79) + (0.0 * 0)}{5 + 68 + 6} = 0.9251\ (93\%)$$

$$\text{Akurasi} = \frac{0 + 79 + 0}{5 + 68 + 6} = 0.8607\ (86\%)$$

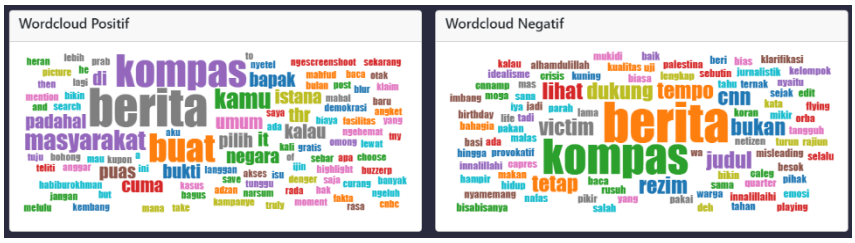
Berdasarkan evaluasi sentimen terhadap 403 data dari data *training* dan juga data *testing*, penelitian menunjukkan bahwa mayoritas warga X di Indonesia mengekspresikan sentimen netral sebesar 97.77%, sentimen positif sebesar 1.49%, dan sentimen negatif sebesar 0.74%.

3.7 Visualisasi

Pada gambar 5 dan 6 merupakan hasil dari penggunaan *Wordcloud* untuk mengetahui perbandingan sentimen, kata yang sering muncul, kata positif, dan kata negatif.



Gambar 5. Visualisasi Perbandingan Sentimen dan Kata Sering Muncul



Gambar 6. Visualisasi *Wordcloud* Positif dan Negatif

4. KESIMPULAN

Setelah evaluasi aplikasi dilakukan, dapat disimpulkan bahwa dari 403 dari 513 tweet, 97.77% sentimen netral, 1.49% positif, dan 0.74% negatif, menunjukkan mayoritas warga X di Indonesia netral terhadap berita Kompas. Metode TF-IDF dan KNN menghasilkan akurasi data *training* 86%, *precision* 86%, *recall* 97%, dan *f1-score* 90%, serta akurasi data *testing* 86%, *precision* 86%, *recall* 100%, dan *f1-score* 93%. Penelitian selanjutnya diharapkan meningkatkan *preprocessing* untuk data yang lebih bersih dan berkualitas, dan menambah kata dalam kamus informal dan formal untuk meningkatkan *evaluation metrics*.

DAFTAR PUSTAKA

[1] A. D. Adhi Putra, "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.

[2] S. S. Salim and J. Mayary, "Analisis Sentimen Pengguna Twitter Terhadap Dompet Elektronik Dengan Metode Lexicon Based Dan K – Nearest Neighbor," *J. Ilm. Inform. Komput.*, vol. 25, no. 1, pp. 1–17, 2020, doi: 10.35760/ik.2020.v25i1.2411.

[3] S. Rahayu, Y. MZ, J. E. Bororing, and R. Hadiyat, "Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP," *Edumatic J. Pendidik. Inform.*, vol. 6, no. 1, pp. 98–106, 2022, doi: 10.29408/edumatic.v6i1.5433.

[4] F. Rozi, F. Sukmana, and M. N. Adani, "Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Inverse Document Frequency (TF-IDF)," *JIMP J. Inform. Merdeka Pasuruan*, vol. 6, no. 3, pp. 1–5, 2021.

[5] A. Tanggu Mara, E. Sediyono, and H. Purnomo, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Metode Pembelajaran Dalam Jaringan (DARING) Di Universitas Kristen Wira Wacana Sumba," *Jointer - J. Informatics Eng.*, vol. 2, no. 01, pp. 24–31, 2021, doi: 10.53682/jointer.v2i01.30.

[6] D. Abimanyu, E. Budianita, E. P. Cynthia, F. Yanto, and Y. Yusra, "Analisis Sentimen Akun Twitter Apex Legends Menggunakan VADER," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 3, pp. 423–431, 2022, doi: 10.32672/jnkti.v5i3.4382.

[7] D. Safryda Putri and T. Ridwan, "Analisis Sentimen Ulasan Aplikasi Pospay Dengan Algoritma Support Vector Machine," *J. Ilm. Inform.*, vol. 11, no. 01, pp. 32–40, 2023, doi: 10.33884/jif.v11i01.6611.

[8] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, 2019, doi: 10.36294/jurti.v3i2.1077.

[9] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 2, pp. 118–127, 2021.

[10] H. Al Aziez and G. Anuraga, "Klasifikasi Daerah Tertinggal di Indonesia Menggunakan Algoritma SVM dan k-NN," *J. Ilmu Dasar*, vol. 22(1), no. 1, pp. 31–38, 2021.