

PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA TWITTER UNTUK ANALISIS SENTIMEN MASYARAKAT TERHADAP LGBT DI INDONESIA

Lulu Mamluatul Hikmah, Nining Rahaningsih, Raditya Damar Dana

Teknik Informatika, STMIK IKMI Cirebon

Jl. Perjuangan No. 10 B Majasem Kec. Kesambi Kota Cirebon

lulumh0172@gmail.com

ABSTRAK

Sampai saat ini, topik LGBT (*Lesbian, Gay, Bisexual, Transgender*) di Indonesia masih sering diperdebatkan karena kelompok ini mulai mengekspresikan diri secara langsung melalui aktivitas yang mereka lakukan dan sering mengomentari aktivitas mereka di media sosial. Hasil laporan Badan Narkotika Nasional (BNN) tercatat tahun 2022 jumlah kasus HIV/AIDS ada 62.856 kasus dengan rincian 9.901 kasus AIDS dan 52.955 kasus HIV. Hal ini menimbulkan perbincangan publik melalui media sosial salah satunya Twitter dengan memberikan berbagai macam komentar berupa dukungan, penolakan dan pandangan netral terhadap fenomena tersebut. Banyaknya jumlah dan ragam respon akan menjadi big data sehingga dapat menghasilkan informasi yang berharga. Data yang digunakan sebanyak 1501 tweet dengan kata kunci yang digunakan adalah "LGBT". Penelitian ini bertujuan untuk mengetahui sentimen masyarakat terhadap kasus LGBT. Metode yang digunakan dalam penelitian ini adalah algoritma *K-Nearest Neighbor*. Penelitian dilakukan untuk mengklasifikasi sentimen ke dalam tiga kelas yaitu positif, negatif dan netral dengan pemodelan 10 *k-folds cross validation*. Hasil klasifikasi yang diperoleh menunjukkan pandangan masyarakat terhadap LGBT cenderung netral ditunjukkan dengan polaritas sentimen netral 86.94% sebanyak 1305 tweets, sentimen positif 8.39% sebanyak 126 tweets dan sentimen negatif 4.66% sebanyak 70 tweets. Hasil akurasi tertinggi dari algoritma *K-Nearest Neighbor* dengan nilai $k=6$ yaitu 85.90%.

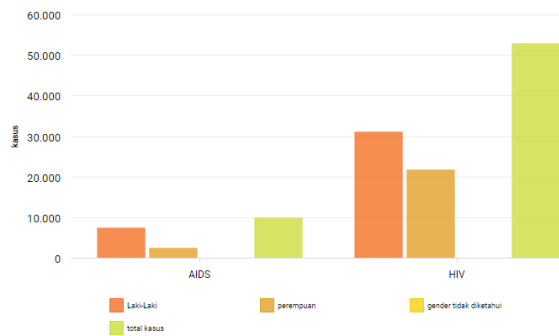
Kata kunci : Analisis sentimen, Twitter, LGBT, *K-nearest Neighbor*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi memberikan banyak perubahan dalam kehidupan masyarakat. Orang menggunakan media sosial sebagai bentuk eksistensinya di dunia maya [1]. Di area digital ini, banyak ragam jenis sosial media salah satunya Twitter. Pengguna Twitter bebas mengungkapkan apa saja baik perasaan atau pendapat. Bagian dari ilmu *text mining* yang dikenal sebagai analisis sentimen ini bertujuan untuk menganalisis dan mengekstrak data teks yang berisi pendapat, evaluasi, sikap, emosi, penilaian, dan sentimen seseorang terhadap objek, individu, organisasi, dan masalah. [2].

LGBT merupakan terjadinya ketidaksesuaian dalam kodrat seksualitas manusia, yang pada hakikatnya memiliki fitrah untuk menyukai sesama jenis. Di zaman sekarang fenomena tentang LGBT sangat diperbincangkan karena kontroversi yang mereka timbulkan di masyarakat. Mulai dari kodratnya yang dipertanyakan, keturunannya atau kaumnya, hingga keselamatannya yang dipertanyakan [3]. Karena kaum LGBT dianggap sebagai aib sosial yang tidak wajar dan membahayakan orang lain, mereka biasanya menghindari tampil dalam masyarakat [3]. Tetapi, seiring berkembangnya trend zaman dan teknologi informasi, LGBT mulai menyebar luas. Kaum ini menampakkan diri di media sosial maupun di lingkungannya. Secara langsung

mereka tidak malu mengekspresikan diri dengan menungghah kegiatan yang mereka lakukan. Masyarakat Indonesia saat ini mengalami perubahan budaya, yang memiliki efek positif dan negatif. Salah satu efek negatif yang paling terlihat adalah kehancuran nilai-nilai, etika, dan keyakinan agama. Keberadaan LGBT itu sendiri tidak karena genetika, ketidakseimbangan hormon, masalah psikologis, atau pelanggaran hukum, sebaliknya, itu adalah hasil dari pengaruh pengalaman pribadi dan lingkungan [4]. Kebebasan kaum LGBT untuk berbicara di media sosial seperti *Twitter* memiliki efek yang signifikan terhadap masyarakat. Media sosial memungkinkan kaum LGBT untuk dengan mudah berinteraksi dengan sesama LGBT [4]. Dengan demikian, gerakan LGBT berkembang melalui media sosial, tempat komunitas LGBT dapat dengan bebas melakukan tujuan dan misi mereka. Menurut Laporan Badan Narkotika Nasional menunjukkan bahwa pada tahun 2022, ada 62.856 kasus HIV/AIDS di Indonesia. Laki-laki mendominasi kasus HIV dan AIDS di Indonesia. Tercatat jumlah kasus HIV pada laki-laki mencapai 31.218 kasus sedangkan kasus HIV perempuan 21.737 kasus, dan 0 kasus yang gendernya tidak diketahui. Adapun jumlah kasus AIDS pada laki-laki sebanyak 7.375 kasus sedangkan kasus AIDS pada perempuan 2.521 kasus, dan 5 kasus untuk gendernya tidak diketahui.



Gambar 1. Grafik HIV/AIDS di Indonesia 2022
Sumber : (<https://databoks.katadata.co.id/>, 2023)

Berdasarkan penelitian yang dilakukan oleh [5], penelitian ini melakukan pengklasifikasian pendapat di Twitter mengenai pro dan kontra pada kaum LGBT menggunakan algoritma *Naïve Bayes*. Didapatkan hasil akurasi sebanyak 95%, dengan data polaritas positif 77%, dan data polaritas negatif sebanyak 150.2%. Kemudian penelitian [6] mengenai sentimen tentang kenaikan harga BBM menggunakan algoritma *K-Nearest Neighbor*, dengan hasil pengujian metode K-NN pada 10 nilai K yang berbeda (3,5,7,9,11,13,15,17,19,21) menggunakan mekanisme perbandingan 70:30, 80:20, dan 90:10. Akurasi paling tinggi sebesar 83.3% diperoleh untuk data pelatihan dan pengujian pada K=13 dan K=15. Kemudian penelitian [7] mengenai sentimen masyarakat terhadap pembelajaran online, dilakukan pengujian dengan K sebanyak 20 didapatkan hasil akurasi tertinggi terdapat pada saat K = 10 dengan nilai akurasi 84,65% dengan presisi mencapai 87%, *recall* 86% *f measure* 87% serta *error rate* mencapai 0,12% dan di dapatkan pula kecenderungan opini publik terhadap pembelajaran daring cenderung positif.

Penelitian ini bertujuan untuk menentukan persepsi masyarakat Indonesia tentang gerakan LGBT. Dalam penelitian ini, fakta yang diambil dari Twitter dengan teknik *text mining* digunakan untuk menunjukkan seberapa banyak masyarakat yang mendukung, menolak, atau memilih untuk tidak peduli dengan adanya LGBT ini. Penelitian ini menunjukkan tingkat akurasi algoritma *K-Nearest Neighbor* yang digunakan dalam kasus *text mining*. *Text mining* menambang data teks (kata atau kalimat) yang diambil dari dokumen dengan tujuan menemukan kata-kata yang dapat mewakili isi dokumen, yang memungkinkan analisis hubungan antar dokumen [8].

Penelitian ini fokus pada analisis dan pemahaman sentimen atau perasaan yang terkandung dalam data Twitter. Tujuan utamanya adalah untuk mengklasifikasikan reaksi terhadap berbagai jenis sentimen, seperti positif, negatif, atau netral. Penelitian ini menggunakan algoritma *K-Nearest Neighbor* dengan *tools Rapidminer* sebagai alat untuk menampilkan hasil klasifikasi data. Dataset yang

digunakan pada penelitian ini yaitu dari tweet para pengguna Twitter. Pengambilan data di Twitter menggunakan teknik *crawling* yang menghasilkan sebanyak 1501 tweet.

Dengan menggunakan metode *K-Nearest Neighbor*, penelitian ini akan menganalisis sentimen masyarakat terhadap LGBT berdasarkan klasifikasi big data pada media sosial Twitter. Analisis tren informasi yang menjadi fokus penelitian ini adalah analisis sentimen. Penelitian ini dilakukan untuk memahami bagaimana pengguna Twitter dapat memengaruhi pendapat orang lain untuk mendukung, menolak, atau tidak peduli dengan peristiwa yang terjadi di dunia nyata, yang berdampak pada individu atau kelompok yang menjadi fokusnya. Hasil penelitian ini dapat memberikan insentif kepada masyarakat Indonesia untuk menggunakan media sosial dengan bijak dan untuk tujuan positif.

2. TINJAUAN PUSTAKA

Paper berjudul “Analisis Sentimen Kaum LGBT Pada Media Sosial Twitter Menggunakan Algoritma *Naïve Bayes*”, penelitian ini berfokus pada analisis sentimen pro dan kontra komunitas LGBT di Twitter menggunakan algoritma *Naïve Bayes*. Penelitian ini menggunakan teknik *scrapping* untuk mengambil data 1579 tweet dari Twitter. Hasilnya menunjukkan bahwa analisis sentimen kaum LGBT menunjukkan akurasi sebesar 95%, polaritas positif sebesar 77%, dan polaritas negatif sebesar 150.2% [5].

Paper berjudul “Klasifikasi Sentimen Masyarakat di Twitter Terhadap Kenaikan Harga BBM dengan Metode K-NN”, data yang dikumpulkan dari 3000 tweet di Twitter digunakan untuk mengklasifikasikan sentimen masyarakat terhadap kenaikan harga BBM. Metode ini digunakan dengan *K-Nearest Neighbor*, *Feature Weighting* (TF-IDF), dan *Feature Selection* (Threshold), dan diimplementasikan dengan *Google Collab*. Hasil pengujian metode K-NN pada 10 nilai K yang berbeda (3,5,7,9,11,13,15,17,19,21) menggunakan mekanisme perbandingan 70:30, 80:20, dan 90:10. Akurasi paling tinggi sebesar 83.3% diperoleh untuk data pelatihan dan pengujian pada K=13 dan K=15 [6].

Paper berjudul “Sentiment Analysis of Post-Covid-19 Inflation Based On Twitter Using The K-Nearest Neighbor and Support Vector Machine Classification Methods”, tujuan dari penelitian ini adalah untuk melihat dan membandingkan akurasi dari dua metode klasifikasi SVM dan K-NN pada proses klasifikasi analisis sentimen. Sebanyak 5989 tweet digunakan. Hasil klasifikasi menunjukkan bahwa metode SVM memiliki akurasi sebesar 79% lebih tinggi daripada K-NN [9].

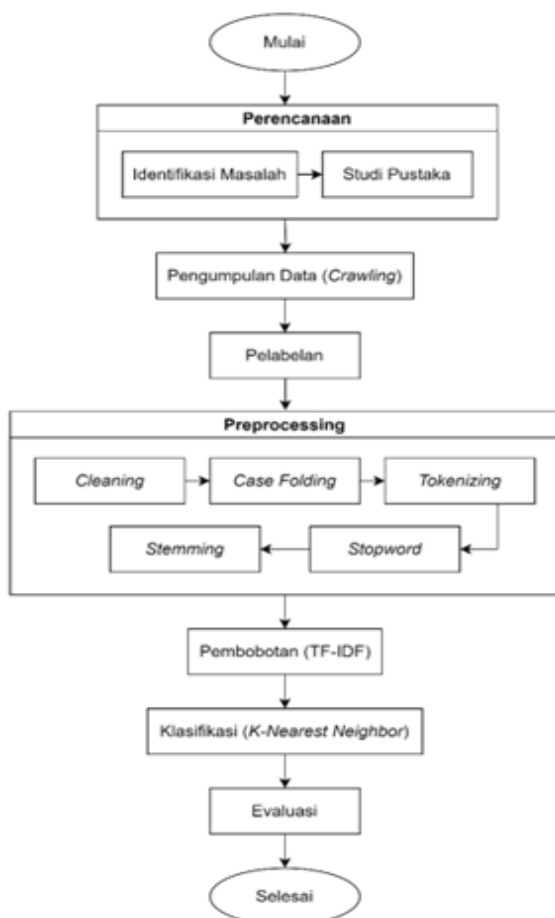
Paper berjudul “Text Classification on Sentiment Analysis of Marketplace SHOPEE Reviews On Twitter Using K-Nearest Neighbor (KNN) Method” membahas tentang gambaran dan hasil klasifikasi sentimen pengguna Twitter terhadap

Shopee. Metode yang digunakan adalah *K-Nearest Neighbor*. Diperoleh 150 data tweet dari Twitter API. Hasil yang diperoleh dari penelitian ini adalah tingkat akurasi klasifikasi teks tertinggi adalah 90% pada data latih dan perbandingan data uji 80% : 20% [10].

Paper berjudul “Visualisasi Data Analisa Sentimen RUU Omnibus Law Kesehatan Menggunakan KNN dengan Software RapidMiner” membahas tentang visualisasi data analisis sentimen terhadap topik kontroversial RUU Omnibus law kesehatan dengan menggunakan algoritma KNN dan *software RapidMiner*. Sebanyak 2.406 data tweet digunakan sebagai dataset dengan pembagian data training 1.688 tweet dan data testing sebanyak 722 tweet. Hasil penelitian menunjukkan bahwa pengujian K-Nearest Neighbors (KNN) menghasilkan nilai akurasi yang tinggi sebesar 84,58%. Visualisasi data penelitian ditampilkan dalam bentuk Cloud of Word dan Confusion Matrix [11].

3. METODE PENELITIAN

Penelitian ini menggunakan analisis sentimen untuk mengelompokkan tweet yang memiliki nilai positif, negatif, atau netral. Metode *K-Nearest Neighbor* digunakan untuk menganalisis sentimen tweet kaum LGBT.



Gambar 2. Tahapan metode penelitian

3.1. Pengumpulan Data

Dalam penelitian ini, teknik crawling digunakan untuk pengambilan data dengan bantuan *Google Colab*. Data yang dikumpulkan mencakup tweet yang menunjukkan sentimen netral, positif, dan negatif terhadap LGBT. Data ini mencakup teks tweet, tanggal dan waktu, serta metadata tambahan lainnya. Berdasarkan kata kunci "LGBT", 1501 data tweet digunakan dari pencarian dengan filter tambahan *lang-id*. Data yang dikumpulkan dari *crawling* Twitter dalam penelitian disimpan dalam format csv dan dikumpulkan dari 26 Oktober hingga 30 Oktober 2023.

3.2. Pelabelan

Untuk mengatasi masalah pengklasifikasian data tweet, pelabelan digunakan untuk memastikan bahwa data tersebut masuk ke dalam kelompok yang sesuai dengan informasinya. Pelabelan data ini menggunakan *library textblob* dengan bahasa pemrograman *python* yang bekerja secara otomatis agar data sentimen dikelompokkan dengan benar berdasarkan nilai polaritas dan subjektivitas dalam kalimat [12].

3.3. Preprocessing

Data yang telah dilabeli perlu diolah sebelum digunakan untuk analisis sentimen. Langkah-langkah *preprocessing* mencakup penghapusan karakter khusus, penghilangan tanda baca, *case folding* (mengubah teks menjadi huruf kecil), *tokenizing* (pemisahan kata-kata), *stopword* (penghilangan kata-kata berhenti), dan *stemming* (pengubahan kata menjadi bentuk dasar).

3.4. Pembobotan

Selanjutnya, penelitian ini akan membagi kalimat menjadi beberapa kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk menilai setiap kata. Tahap ini merupakan statistik numerik yang menunjukkan seberapa penting sebuah kata untuk dokumen dalam koleksi [12]. Pengukuran bobot dilakukan dengan memastikan bahwa setiap kata dalam tweet muncul dalam daftar fitur positif, negatif, dan netral. Tweet ini akan digunakan sebagai data latih dan data uji [11].

3.5. Klasifikasi K-Nearest Neighbor

Penelitian ini melakukan klasifikasi menggunakan algoritma *K-Nearest Neighbor* dengan bantuan *tools Rapidminer*. K-NN yang juga dikenal sebagai metode berbasis jarak, adalah algoritma klasifikasi supervised learning [13]. Metode K-NN mengklasifikasikan kumpulan data dengan data latih yang telah terklasifikasi sebelumnya [14]. Label kelas titik data uji diprediksi dengan KNN berdasarkan mayoritas label kelas dari k tetangga terdekatnya [11]. Kelas yang paling sering muncul adalah hasil dari klasifikasinya.

3.6. Evaluasi

Hasil dari tahapan klasifikasi algoritma *K-Nearest Neighbor* adalah kumpulan pendapat masyarakat berdasarkan frekuensi kata yang muncul sesuai dengan kelas masing-masing, yaitu positif, negatif, dan netral. Kemudian melakukan evaluasi hasil akurasi dari pengujian algoritma *K-Nearest Neighbor* menggunakan *Word Cloud*.

4. HASIL DAN PEMBAHASAN

Pengujian dilakukan sepuluh kali percobaan untuk mengevaluasi nilai *k* algoritma *K-Nearest Neighbor* menggunakan Rapidminer. Hasil dari analisis didapatkan nilai akurasi yaitu sebesar 85.90% saat menggunakan model pengujian *cross validation* dengan nilai *k*=6. Dari hasil yang didapatkan, LGBT merupakan topik yang biasa dimana masyarakat berpendapat netral. Adapun dalam hasil tersebut yang beropini netral 86.94% sebanyak 1305 tweets, 8.39% positif sebanyak 126 tweets dan 4.66% negatif sebanyak 70 tweets dari 1501 data. Hasilnya menunjukkan bahwa masyarakat Indonesia sebagian besar memiliki respon netral atau tidak peduli terhadap isu LGBT di Indonesia.

4.1. Pengumpulan data

Pada tahapan ini data yang digunakan adalah *crawling* data tweet dari hasil kata kunci “LGBT”. Sebelum pengambilan data, terlebih dahulu memasukan token yang didapat dari akun Twitter pribadi sebagai akses yang digunakan untuk syarat kunci pengambilan data. Pengambilan data menggunakan *tools google colab* dengan *library tweet-harvest* menghasilkan total data sebanyak 1501 tweet dari 26-30 Oktober 2023 dengan atribut diantaranya *created_at*, *id_str*, *full_text*, *quote_count*, *reply_count*, *retweet_count*, *favorite_count*, *lang*, dll. Bentuk data yang telah dikumpulkan dapat dilihat pada Gambar 3.

	created_at	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	lang
0	Mon Oct 30 23:58:51 +0000 2023	1719142098543521947	anak anak plss kalian jangan sampai dukung zio...	0	0	0	0	in
1	Mon Oct 30 23:58:08 +0000 2023	1719141920067424661	@Suchkijan @MamanSu30264482 @tubirfess Udah @...	0	2	0	0	in
2	Mon Oct 30 23:58:23 +0000 2023	1719141270566904077	@ddgafdat Latarbelakang lgbt, cepat taubat lah...	0	0	0	0	in
3	Mon Oct 30 23:49:01 +0000 2023	1719139375051940195	@ikhitiromal @akusiapa1192798 @andikamalreza @f...	0	1	0	0	in
4	Mon Oct 30 23:46:59 +0000 2023	171913882266285433	@fhnmandaxx @mefena_mate @tanyarifes Udah lama...	0	0	1	2	in
...	...	...	...	...	...	...	...	...
1496	Thu Oct 26 14:44:31 +0000 2023	1717552794880925884	@tanyakanrl Tuman emang LGBT harus di kasih pe...	0	0	0	0	in
1497	Thu Oct 26 14:42:47 +0000 2023	1717552358442602922	@riko_history Makanya jgn ganggu kaum normal j...	0	1	0	13	in
1498	Thu Oct 26 14:39:31 +0000 2023	1717551535457202684	@memefess aku fans berat lagu lagu mereka dan ...	0	0	0	0	in
1499	Thu Oct 26 14:38:09 +0000 2023	1717551443551678763	@apakitakeos lya ntar dengan alasan ngasih pel...	0	1	0	0	in
1500	Thu Oct 26 14:33:39 +0000 2023	1717550058911224064	@tanyakanrl lu pada anti lgbt anti lgbt tapi h...	0	0	0	1	in

Gambar 3. Hasil crawling data

Dari semua data yang ada akan diseleksi dan atribut yang digunakan untuk proses *K-Nearest Neighbor* yaitu *text* tweet, merupakan atribut yang

terdapat pada tabel hasil *crawling* yang berisi kalimat/komentar pengguna Twitter terkait *keywords* LGBT.

	full_text
0	anak anak plss kalian jangan sampai dukung zio...
1	@Suchkijan @MamanSu30264482 @tubirfess Udah @...
2	@ddgafdat Latarbelakang lgbt, cepat taubat lah...
3	@ikhitiromal @akusiapa1192798 @andikamalreza @f...
4	@fhnmandaxx @mefena_mate @tanyarifes Udah lama...
...	...
1496	@tanyakanrl Tuman emang LGBT harus di kasih pe...
1497	@riko_history Makanya jgn ganggu kaum normal j...
1498	@memefess aku fans berat lagu lagu mereka dan ...
1499	@apakitakeos lya ntar dengan alasan ngasih pel...
1500	@tanyakanrl lu pada anti lgbt anti lgbt tapi h...

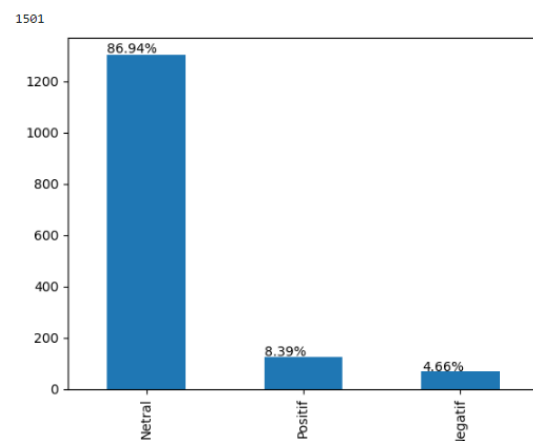
Gambar 4. Seleksi data

4.2. Pelabelan

Data yang telah berhasil dikumpulkan dan diseleksi selanjutnya diberi label. Pelabelan dilakukan secara otomatis menggunakan *library textblob*. Karena penelitian ini menggunakan metode *K-Nearest Neighbor* untuk analisis sentimen, data harus memiliki label. Berikut contoh hasil pemberian label sentimen pada Gambar 5.

full_text	sentiment	polarity	subjectivity
anak anak plss kalian jangan sampai dukung zio...	Netral	0.0	0.0
@Suchkijan @MamanSu30264482 @tubirfess Udah @...	Netral	0.0	0.0
@ddgafdat Latarbelakang lgbt, cepat taubat lah...	Netral	0.0	0.0
@ikhitiromal @akusiapa1192798 @andikamalreza @f...	Netral	0.0	0.0
@fhnmandaxx @mefena_mate @tanyarifes Udah lama...	Negatif	-0.1	0.1
@elonmusk @StephenKing lgbt?	Netral	0.0	0.0
@sakurakmst @SoundOfYogi Mikir ga kalau genosi...	Netral	0.0	0.0
@IncognitoSpring @temanduit @yunartowijaya Doy...	Positif	0.5	1.0
@Estadao Petista LGBT	Netral	0.0	0.0
@SoundOfYogi Gue rada2 susah percaya sama oran...	Netral	0.0	0.0

Gambar 5. Contoh hasil pelabelan



Gambar 6. Diagram bar persentase sentimen

Dari 1501 tweet terdapat 1305 netral, 126 positif dan 70 negatif, sehingga dapat disimpulkan bahwa data dari hasil crawling Twitter ini cenderung netral. Adapun Gambar 6 menunjukkan proporsi data yang digunakan dalam penelitian ini.

4.3. Preprocessing

Selanjutnya dilakukan tahap pengolahan data yang meliputi *cleaning*, *tokenizing*, *case folding*, *stopwords*, dan *stemming*.

4.3.1. Cleaning

Pada tahap *cleaning* dilakukan dengan menggunakan *library pandas* untuk menghilangkan karakter atau simbol, *emoticon*, *hashtag* (#), *mention* (@username), *retweet* (RT) dan url (<http://>). Setelah melewati proses *cleaning*, data yang awalnya terdapat 1501 data tweet berkurang menjadi 1305 data.

Tabel 1. Hasil cleaning

Sebelum	Sesudah
anak anak plss kalian jangan sampai dukung zionist ya. cukup pro lgbt aja 😊 tpi jangan dilakuin juga 😊	anak anak plss kalian jangan sampai dukung zionist ya cukup pro lgbt aja tpi jangan dilakuin juga
@Suchkijan @MamanSu30264482 @tubirfess Udah @Suchkijan jangan ditanggepin, percuma klo kalah debat omongannya bakal kotor kek otaknya 🤔. Isi like2 di twitnya aja kek gini. Sok anti Islam, sok pro LGBT, dll. Udah cukup tau kan isi otaknya si BANG MAMAN ini gimana?! Wkwk.. mending kita yg waras ngalah. https://t.co/lvSK9YzFlh	Udah jangan ditanggepin percuma klo kalah debat omongannya bakal kotor kek otaknya Isi like2 di twitnya aja kek gini Sok anti Islam sok pro LGBT dll Udah cukup tau kan isi otaknya si BANG MAMAN ini gimana Wkwk mending kita yg waras ngalah
@sakurakmst @SoundOfYogi Mikir ga kalau genosida dan LGBT itu sama2 berbahaya?? Dan harus dilenyapkan, gue ga bilang orgnya ya, tp penyimpangannya. Cuman manusia LGBT ini yg ga mau penyimpangannya dilenyapkan	Mikir ga kalau genosida dan LGBT itu sama berbahaya Dan harus dilenyapkan gue ga bilang orgnya ya tp penyimpangannya Cuman manusia LGBT ini yg ga mau penyimpangannya dilenyapkan

4.3.2. Tokenizing

Pada tahapan *tokenizing* dilakukan menggunakan bantuan *library NLTK (Natural Language Toolkit)* yaitu *library* yang memproses bahasa alami dengan membagi teks menjadi setiap bagian kalimat dan kata berdasarkan spasi yang ditemukan.

Tabel 2. Hasil tokenizing

Sebelum	Sesudah
anak anak plss kalian jangan sampai dukung zionist ya cukup pro lgbt aja tpi jangan dilakuin	['anak', 'anak', 'plss', 'kalian', 'jangan', 'sampai', 'dukung', 'zionist', 'ya', 'cukup', 'pro', 'lgbt', 'aja', 'tpi', 'jangan',

Sebelum	Sesudah
juga	['dilakuin', 'juga']
Udah jangan ditanggepin percuma klo kalah debat omongannya bakal kotor kek otaknya Isi like di twitnya aja kek gini Sok anti Islam sok pro LGBT dll Udah cukup tau kan isi otaknya si BANG MAMAN ini gimana Wkwk mending kita yg waras ngalah	['Udah', 'jangan', 'ditanggepin', 'percuma', 'klo', 'kalah', 'debat', 'omongannya', 'bakal', 'kotor', 'kek', 'otaknya', 'Isi', 'like', 'di', 'twitnya', 'aja', 'kek', 'gini', 'Sok', 'anti', 'Islam', 'sok', 'pro', 'LGBT', 'dll', 'Udah', 'cukup', 'tau', 'kan', 'isi', 'otaknya', 'si', 'BANG', 'MAMAN', 'ini', 'gimana', 'Wkwk', 'mending', 'kita', 'yg', 'waras', 'ngalah']
Mikir ga kalau genosida dan LGBT itu sama berbahaya Dan harus dilenyapkan gue ga bilang orgnya ya tp penyimpangannya Cuman manusia LGBT ini yg ga mau penyimpangannya dilenyapkan	['Mikir', 'ga', 'kalau', 'genosida', 'dan', 'LGBT', 'itu', 'sama', 'berbahaya', 'Dan', 'harus', 'dilenyapkan', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'penyimpangannya', 'Cuman', 'manusia', 'LGBT', 'ini', 'yg', 'ga', 'mau', 'penyimpangannya', 'dilenyapkan']

4.3.3. Case Folding

Pada tahap *case folding*, semua huruf diubah menjadi huruf kecil, tanda baca dan angka juga dihilangkan. Dapat dilihat pada Tabel 3.

Tabel 3. Hasil case folding

Sebelum	Sesudah
['anak', 'anak', 'plss', 'kalian', 'jangan', 'sampai', 'dukung', 'zionist', 'ya', 'cukup', 'pro', 'lgbt', 'aja', 'tpi', 'jangan', 'dilakuin', 'juga']	['anak', 'anak', 'plss', 'kalian', 'jangan', 'sampai', 'dukung', 'zionist', 'ya', 'cukup', 'pro', 'lgbt', 'aja', 'tpi', 'jangan', 'dilakuin', 'juga']
['Udah', 'jangan', 'ditanggepin', 'percuma', 'klo', 'kalah', 'debat', 'omongannya', 'bakal', 'kotor', 'kek', 'otaknya', 'Isi', 'like', 'di', 'twitnya', 'aja', 'kek', 'gini', 'Sok', 'anti', 'Islam', 'sok', 'pro', 'LGBT', 'dll', 'Udah', 'cukup', 'tau', 'kan', 'isi', 'otaknya', 'si', 'BANG', 'MAMAN', 'ini', 'gimana', 'Wkwk', 'mending', 'kita', 'yg', 'waras', 'ngalah']	['udah', 'jangan', 'ditanggepin', 'percuma', 'klo', 'kalah', 'debat', 'omongannya', 'bakal', 'kotor', 'kek', 'otaknya', 'isi', 'like', 'di', 'twitnya', 'aja', 'kek', 'gini', 'sok', 'anti', 'islam', 'sok', 'pro', 'lgbt', 'dll', 'udah', 'cukup', 'tau', 'kan', 'isi', 'otaknya', 'si', 'bang', 'maman', 'ini', 'gimana', 'wkwk', 'mending', 'kita', 'yg', 'waras', 'ngalah']
['Mikir', 'ga', 'kalau', 'genosida', 'dan', 'LGBT', 'itu', 'sama', 'berbahaya', 'Dan', 'harus', 'dilenyapkan', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'penyimpangannya', 'Cuman', 'manusia', 'LGBT', 'ini', 'yg', 'ga', 'mau', 'penyimpangannya', 'dilenyapkan']	['mikir', 'ga', 'kalau', 'genosida', 'dan', 'lgbt', 'itu', 'sama', 'berbahaya', 'dan', 'harus', 'dilenyapkan', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'penyimpangannya', 'cuman', 'manusia', 'lgbt', 'ini', 'yg', 'ga', 'mau', 'penyimpangannya', 'dilenyapkan']

4.3.4. Stopword

Tahap ini dilakukan proses untuk menghilangkan kata yang kurang penting atau kata penghubung. pada *library NLTK* terdapat fitur

penghapusan *stopword* Bahasa Indonesia. Tabel 4 tampilan hasil *stopword*.

Tabel 4. Hasil remove *stopword*

Sebelum	Sesudah
['anak', 'anak', 'plss', 'kalian', 'jangan', 'sampai', 'dukung', 'zionist', 'ya', 'cukup', 'pro', 'lgbt', 'aja', 'tpi', 'jangan', 'dilakuin', 'juga']	['anak', 'anak', 'plss', 'dukung', 'zionist', 'ya', 'pro', 'lgbt', 'aja', 'tpi', 'dilakuin']
['udah', 'jangan', 'ditanggepin', 'percuma', 'klo', 'kalah', 'debat', 'omongannya', 'bakal', 'kotor', 'kek', 'otaknya', 'isi', 'like', 'di', 'twitnya', 'aja', 'kek', 'gini', 'sok', 'anti', 'islam', 'sok', 'pro', 'lgbt', 'dll', 'udah', 'cukup', 'tau', 'kan', 'isi', 'otaknya', 'si', 'bang', 'maman', 'ini', 'gimana', 'wkwk', 'mending', 'kita', 'yg', 'waras', 'ngalah']	['udah', 'ditanggepin', 'klo', 'kalah', 'debat', 'omongannya', 'kotor', 'kek', 'otaknya', 'isi', 'like', 'twitnya', 'aja', 'kek', 'gini', 'sok', 'anti', 'islam', 'sok', 'pro', 'lgbt', 'dll', 'udah', 'tau', 'isi', 'otaknya', 'si', 'bang', 'maman', 'gimana', 'wkwk', 'mending', 'yg', 'waras', 'ngalah']
['mikir', 'ga', 'kalau', 'genosida', 'dan', 'lgbt', 'itu', 'sama', 'berbahaya', 'dan', 'harus', 'dilenyapkan', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'penyimpangannya', 'cuman', 'manusia', 'lgbt', 'ini', 'yg', 'ga', 'mau', 'penyimpangannya', 'dilenyapkan']	['mikir', 'ga', 'genosida', 'lgbt', 'berbahaya', 'dilenyapkan', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'penyimpangannya', 'cuman', 'manusia', 'lgbt', 'yg', 'ga', 'penyimpangannya', 'dilenyapkan']

4.3.5. Stemming

Pada tahap *stemming*, kata berimbuhan dalam dokumen diubah menjadi kata dasar. *Stemming* yang digunakan yaitu algoritma Sastrawi, dapat dilihat pada Tabel 5.

Tabel 5. Hasil stemming

Sebelum	Sesudah
['anak', 'anak', 'plss', 'dukung', 'zionist', 'ya', 'pro', 'lgbt', 'aja', 'tpi', 'dilakuin']	['anak', 'anak', 'plss', 'dukung', 'zionist', 'ya', 'pro', 'lgbt', 'aja', 'tpi', 'dilakuin']
['udah', 'ditanggepin', 'klo', 'kalah', 'debat', 'omongannya', 'kotor', 'kek', 'otaknya', 'isi', 'like', 'twitnya', 'aja', 'kek', 'gini', 'sok', 'anti', 'islam', 'sok', 'pro', 'lgbt', 'dll', 'udah', 'tau', 'isi', 'otaknya', 'si', 'bang', 'maman', 'gimana', 'wkwk', 'mending', 'yg', 'waras', 'ngalah']	['udah', 'ditanggepin', 'klo', 'kalah', 'debat', 'omong', 'kotor', 'kek', 'otak', 'isi', 'like', 'twitnya', 'aja', 'kek', 'gin', 'sok', 'anti', 'islam', 'sok', 'pro', 'lgbt', 'dll', 'udah', 'tau', 'isi', 'otak', 'si', 'bang', 'maman', 'gimana', 'wkwk', 'mending', 'yg', 'waras', 'ngalah']
['mikir', 'ga', 'genosida', 'lgbt', 'berbahaya', 'dilenyapkan', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'penyimpangannya', 'cuman', 'manusia', 'lgbt', 'yg', 'ga', 'penyimpangannya', 'dilenyapkan']	['mikir', 'ga', 'genosida', 'lgbt', 'bahaya', 'lennyap', 'gue', 'ga', 'bilang', 'orgnya', 'ya', 'tp', 'simpang', 'cuman', 'manusia', 'lgbt', 'yg', 'ga', 'simpang', 'lennyap']

4.4. Pembobotan TF-IDF

Tahap selanjutnya adalah pembobotan kata dengan TF-IDF untuk mengetahui seberapa sering suatu kata muncul dalam dokumen. Contoh hasil pembobotan sebagaimana Gambar 6.

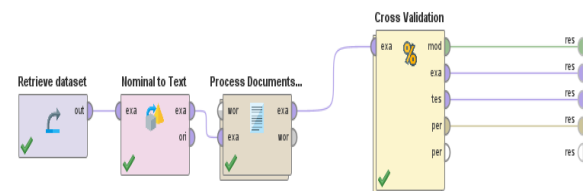
Row No.	sentimen	aamin	aana	aante	aspa	abad	abadi	abai
1	Neutral	0	0	0	0	0	0	0
2	Neutral	0	0	0	0	0	0	0
3	Neutral	0	0	0	0	0	0	0
4	Neutral	0	0	0	0	0	0	0
5	Negatif	0	0	0	0	0	0	0
6	Neutral	0	0	0	0	0	0	0
7	Neutral	0	0	0	0	0	0	0
8	Positif	0	0	0	0	0	0	0
9	Neutral	0	0	0	0	0	0	0
10	Neutral	0	0	0	0	0	0	0
11	Neutral	0	0	0	0	0	0	0
12	Neutral	0	0	0	0	0	0	0
13	Neutral	0	0	0	0	0	0	0
14	Neutral	0	0	0	0	0	0	0

Gambar 6. Hasil pembobotan

Sesuai dengan pembobotan TF-IDF, nilai setiap tweet yang dihasilkan dari pengolahan data dibandingkan dengan probabilitas masing-masing atribut. Hasil dari probabilitas atribut positif, negatif, dan netral dibandingkan untuk menentukan atribut dengan kemungkinan yang paling tinggi.

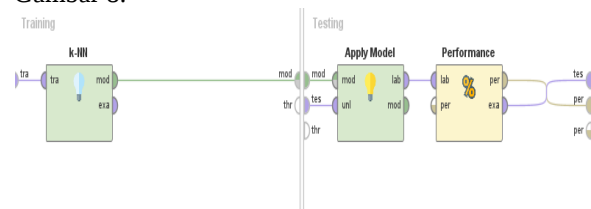
4.5. Klasifikasi K-Nearest Neighbor

Untuk melakukan klasifikasi K-NN menggunakan model 10 *k-fold validation* pada *cross validation*. Tahapan pemodelan dengan *k-fold validation* dapat dilihat pada Gambar 7.



Gambar 7. Model proses

Didalam operator *cross validation* terdapat pembagian data training dan data testing seperti pada Gambar 8.



Gambar 8. Model subprocess

Setelah pemodelan dilakukan, akurasi dapat dilihat pada Tabel 6.

Tabel 6. Perbandingan Hasil Akurasi K-Fold

K-Fold	K-NN
1	85.13%
2	85.13%
3	85.59%
4	85.82%
5	85.75%
6	85.90%
7	85.52%
8	85.67%
9	85.67%
10	85.60%

Tabel 6 menunjukkan hasil percobaan *k-fold validation* di mana K-NN mendapatkan nilai akurasi tertinggi 85.90 % pada nilai *k* = 6.

accuracy: 85.90% +/- 0.63% (micro average: 85.90%)

	true Netral	true Negatif	true Positif	class precision
pred. Netral	1110	61	121	85.91%
pred. Negatif	1	8	0	88.89%
pred. Positif	1	0	3	75.00%
class recall	99.82%	11.59%	2.42%	

Gambar 9. Hasil akurasi

Dari hasil uji coba data menggunakan algoritma K-NN dengan performa klasifikasi dapat dilihat pada Gambar 9 menunjukkan klasifikasi dikatakan netral 1110 *true* netral, 61 *true* negatif dan 121 *true* positif dengan *class precision* nya 85.91%, sedangkan klasifikasi negatif yaitu 1 *true* netral, 8 *true* negatif dan 0 *true* positif dengan *class precision* nya 88.89%, kemudian klasifikasi positif yaitu 1 *true* netral, 0 *true* negatif dan 3 *true* positif dengan *class precision* nya 75%. Untuk sentimen netral memiliki *class recall* 99.82%, sentimen negatif dengan *class recall* 11.59%, sedangkan pada sentimen positif memiliki *class recall* sebesar 2.42%. Nilai akurasi yang dihasilkan adalah 85.90%. Sehingga, algoritma *K-Nearest Neighbor* dapat digunakan dengan baik untuk klasifikasi sentimen.

Setelah semua proses analisa dilakukan untuk menghasilkan klasifikasi sentimen LGBT di Indonesia menggunakan algoritma *K-Nearest Neighbor*, maka hasil yang dicapai pada penelitian ini dari 26-30 Oktober 2023 melalui data hasil *crawling* 26-30 Oktober 2023 pada Twitter yaitu nilai *accuracy* nya 85.90%. Dapat dilihat pada gambar dibawah.

PerformanceVector

```
PerformanceVector:
accuracy: 85.90% +/- 0.63% (micro average: 85.90%)
ConfusionMatrix:
True:  Netral  Negatif  Positif
Netral: 1110   61      121
Negatif: 1     8       0
Positif: 1     0       3
kappa: 0.092 +/- 0.072 (micro average: 0.095)
ConfusionMatrix:
True:  Netral  Negatif  Positif
Netral: 1110   61      121
Negatif: 1     8       0
Positif: 1     0       3
```

Gambar 10. Performa vektor

4.6. Evaluasi

Visualisasi dengan *word cloud* digunakan untuk mengukur frekuensi kata yang sering muncul.

Open in Turbo Prep Auto Model Filter (20 / 20 example)

Row No.	word	in documents	total	in class (Netral)	in class (Negatif)	in class (Positif)
1	lgbt	1300	1404	1181	82	141
2	kaum	145	158	131	11	16
3	orang	116	151	123	13	15
4	dukung	109	133	114	8	11
5	kalo	91	107	71	13	23
6	udah	81	92	70	7	15
7	palestina	73	88	76	7	5
8	israel	66	76	65	6	5
9	emang	68	70	53	7	10
10	manusia	58	70	62	4	4
11	bela	43	60	54	4	2
12	agama	46	57	48	1	8
13	salah	44	55	45	3	7
14	suka	45	51	39	4	8
15	negara	42	50	40	2	8

ExampleSet (20 examples, 0 special attributes, 6 regular attributes)

Gambar 11. Filter 20 kata yang paling sering muncul

Berdasarkan Gambar 11 terlihat bahwa kata “LGBT” mendominasi *word cloud*, yang juga merupakan kata kunci dari pengambilan data penelitian ini. Kata “Kaum” juga muncul cukup besar menunjukkan tweet yang mendominasi di Twitter. Kemudian diikuti kata “Orang” dalam *word cloud* menunjukkan bahwa sentimen pada isu ini tergantung pada pandangan individu dan kelompok. Jika divisualisasikan dengan tampilan *word cloud* sebagaimana Gambar 12 berikut.



Gambar 12. Visualisasi word cloud

Gambar diatas menunjukkan bahwa kata-kata yang paling sering digunakan ditampilkan dalam *word cloud*. Semakin sering suatu kata muncul, semakin menonjol kata tersebut dibandingkan dengan kata yang lain. Kata-kata yang paling banyak digunakan di Twitter adalah LGBT, Kaum, dan Orang. Ini menunjukkan bahwa pengguna memberi tanggapan positif, negatif, atau netral.

5. KESIMPULAN DAN SARAN

Diketahuinya pandangan masyarakat terhadap LGBT pada Twitter dilihat dari hasil pelabelan tertinggi terdapat pada sentimen netral yang terprediksi dengan nilai 86.94% daripada sentimen positif dan negatif. Masyarakat memilih tidak memihak, mencoba memahami berbagai perspektif tanpa menilai atau menghakimi. Meskipun sentimen

netral dapat mencerminkan nilai-nilai seperti keberagaman dan hak asasi manusia, penting untuk diingat bahwa ada beragam perbedaan pendapat dalam masyarakat tentang masalah ini. Berdasarkan hasil analisis menggunakan algoritma *K-Nearest Neighbor* hasil akurasi yang didapatkan yaitu 85.90%. Sehingga hal ini menunjukkan bahwa algoritma *K-Nearest Neighbor* dapat digunakan dengan baik untuk klasifikasi sentimen terhadap data Twitter yang termasuk dalam kategori kelas netral, positif, atau negatif. Saran untuk penelitian selanjutnya dapat dicoba penggunaan pelabelan manual ataupun juga menggunakan *platform* lain seperti *vader* ataupun *spacy*.

DAFTAR PUSTAKA

- [1] A. K. Fauziyyah, "Analisis Sentimen Pandemi Covid19 Pada Streaming Twitter Dengan Text Mining Python," *J. Ilm. SINUS*, vol. 18, no. 2, p. 31, 2020, doi: 10.30646/sinus.v18i2.491.
- [2] D. A. Lestari and D. Mahdiana, "Penerapan Algoritma K-Nearest Neighbor pada Twitter untuk Analisis Sentimen Masyarakat Terhadap Larangan Mudik 2021," *Inform. J. Ilmu Komput.*, vol. 17, no. 2, p. 123, 2021, doi: 10.52958/iftk.v17i2.3629.
- [3] E. Charitoun and S. Gawa, "Penggunaan Media Sosial Sebagai Simbol dalam Mendukung Hubungan LGBT," *J. Educ.*, vol. 05, no. 04, pp. 15598–15608, 2023, [Online]. Available: <http://jonedu.org/index.php/joe>
- [4] D. W. Ardras and A. Voutama, "Analisis Sentimen Anti Lgbt Di Indonesia Melalui Media Sosial Twitter," *J. Tek.*, vol. 15, no. 1, pp. 23–28, 2023, doi: 10.30736/jt.v15i1.926.
- [5] A. Anjani, A. Chamid, and A. Murti, "Analisis Sentimen Kaum LGBT pada Media Sosial Twitter Menggunakan Algoritma Naïve Bayes," *JTINFO J. Tek. Inform.*, vol. 1, no. 2, pp. 1–8, 2022.
- [6] T. D. Arista, Y. Yusra, M. Fikry, and L. Oktavia, "Klasifikasi Sentimen Masyarakat di Twitter terhadap Kenaikan Harga BBM dengan Metode K-NN," *JUKI J. Komput. dan ...*, 2023, [Online]. Available: <https://www.ioinformatic.org/index.php/JUKI/article/view/189>
- [7] A. R. Isnain, J. Supriyanto, and ..., "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian J. ...*, 2021, [Online]. Available: <https://journal.ugm.ac.id/ijccs/article/view/65176>
- [8] M. Mustakim, M. Ridwan, and ..., "Implementasi Algoritma Support Vector Machine untuk Analisis Sentimen LGBT di Indonesia," *Bull. Artif. ...*, 2022, [Online]. Available: <https://journal.grahamitra.id/index.php/buai/article/view/31>
- [9] R. Puspitasari, Y. Findawati, and ..., "SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION BASED ON TWITTER USING THE K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE ...," *J. Tek. Inform. ...*, 2023, [Online]. Available: <http://www.jutif.if.unsoed.ac.id/index.php/jurnal/article/view/801>
- [10] Z. Rais, R. N. Said, and R. Ruliana, "Text Classification on Sentiment Analysis of Marketplace SHOPEE Reviews On Twitter Using K-Nearest Neighbor (KNN) Method," *JINAV J. Inf. ...*, 2022, [Online]. Available: <https://sainsmat.org/index.php/jinav/article/view/1389>
- [11] S. Abdullah, F. I. Komputer, M. T. Informatika, I. Informatika, and B. Lampung, "Visualisasi Data Analisa Sentimen RUU Omnibus law Kesehatan Menggunakan KNN dengan Software RapidMiner," vol. 8, no. 3, pp. 261–268, 2023.
- [12] M. Ridwan, "Analisis Sentimen Masyarakat Pada Twitter Terhadap Kasus Lgbt Di Indonesia Menggunakan Algoritma Support Vector Machine," pp. 1–69, 2019.
- [13] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *INSYST J. ...*, 2019, [Online]. Available: <https://jurnal.istts.ac.id/index.php/INSYST/article/view/36>
- [14] S. Sayed and M. Fikry, "Klasifikasi Sentimen Masyarakat di Twitter terhadap Ganjar Pranowo dengan Metode K-Nearest Neighbor," *JSAI (Journal Sci. Appl. ...*, 2023, [Online]. Available: <http://jurnal.umb.ac.id/index.php/JSAI/article/view/5414>