

PERBANDINGAN LEXICON BASED DAN NAÏVE BAYES CLASSIFIER PADA ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP GEMPA TURKI

COMPARISON OF LEXICON-BASED AND NAIVE BAYES CLASSIFIER METHODS ON SENTIMENT ANALYSIS OF TWITTER USERS TO THE TURKEY EARTHQUAKE

Ahmad Faizal¹, Agung Susilo Yuda Irawan², Didi Juardi³

^{1,2,3}Universitas Singaperbangsa Karawang
1910631170156@student.unsika.ac.id

ABSTRACT

The Turkish Earthquake disaster that claimed many lives is currently busy both in the national and international media, which has led to the emergence of many opinions among social media users, especially on the Twitter Platform. Tweets posted by Twitter social media users can then be used as a useful source of information. Due to this, sentiment analysis can be used as a solution to process these voices using the Lexicon-Based Approach and Naïve Bayes Classifier. The purpose of this study is to classify opinions about the Earthquake Disaster event that occurred in Turkey on February 6, 2023, based on positive sentiment, neutral sentiment, and negative sentiment classes. A 90:10 scenario was used for testing. The evaluation results show that testing the Lexicon-Based Approach and Naïve Bayes Classifier produces an accuracy value of 65%. While Naïve Bayes Classifier without Lexicon-Based Approach produces an accuracy value of 64%.

Keyword: *Sentiment Analysis, Turkey Earthquake, Lexicon Based, Naïve Bayes Classifier*

ABSTRAK

Peristiwa bencana Gempa Turki yang menelan banyak korban jiwa sedang ramai saat ini baik di media nasional maupun media internasional, hal ini menyebabkan munculnya banyak opini pengguna sosial media terutama dalam Platform Twitter. Tweet yang diposting oleh pengguna sosial media Twitter tersebut kemudian dapat dijadikan sumber informasi yang bermanfaat. Dikarenakan hal tersebut, analisis sentimen dapat digunakan sebagai solusi untuk mengolah suara tersebut dengan menggunakan pendekatan Lexicon Based dan Naïve Bayes Classifier. Tujuan dari penelitian ini adalah untuk mengklasifikasikan pendapat tentang peristiwa Bencana Gempa yang terjadi di Turki pada 6 Februari 2023 berdasarkan kelas sentimen positif, sentimen netral dan sentimen negatif. Skenario 90:10 digunakan untuk pengujian. Hasil evaluasi menunjukkan bahwa pengujian pendekatan Lexicon Based dan Naïve Bayes Classifier menghasilkan nilai akurasi sebesar 65%. Sedangkan Naïve Bayes Classifier tanpa pendekatan Lexicon Based menghasilkan nilai akurasi sebesar 64%.

Kata Kunci: *Analisis Sentimen, Gempa Turki, Lexicon Based, Naïve Bayes Classifier*

PENDAHULUAN

Perkembangan teknologi internet seperti saat ini data digital terus meningkat, seiring dengan hal tersebut data digital harus diolah menjadi informasi yang berguna. Media sosial dapat benar-benar dimanfaatkan untuk menginformasikan masalah, mengekspresikan dukungan atau memberi tahu pihak berwenang selama keadaan darurat dan pasca bencana. Peristiwa bencana alam adalah keanehan yang disebabkan oleh serangkaian peristiwa yang terjadi secara alami, termasuk gempa bumi, letusan gunung berapi, banjir, musim kemarau, badai, dan longsor salju. Gempa bumi atau gempa

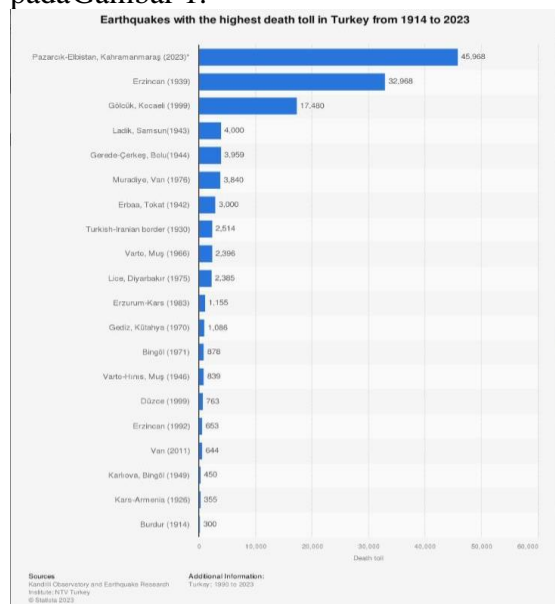
adalah getaran atau goyangan tanah yang disebabkan oleh pelepasan energi dari dalam bumi, terutama dari dalam lapisan-lapisan kerak bumi. Gempa dapat terjadi di mana saja di dunia, namun lebih sering terjadi di wilayah-wilayah di sekitar perbatasan lempeng tektonik di mana lempeng tersebut bertabrakan atau bergerak melewati satu sama lain (Fan et al., 2019).

Analisis sentimen terhadap postingan terkait bencana di media sosial dapat membantu dalam mendeteksi postingan yang berkontribusi terhadap kesadaran situasional dan lebih memahami dinamika jaringan termasuk perasaan, kepanikan, dan kekhawatiran pengguna dengan

mengidentifikasi polaritas sentiment yang diekspresikan oleh pengguna selama peristiwa bencana untuk meningkatkan pengambilan keputusan (Beigi et al., 2016).

Awal tahun 2023, dunia dikagetkan dengan kejadian bencana gempa bumi yang melanda Turki. Pada tanggal 6 Februari 2023, dua gempa bumi besar melanda provinsi Turki, Kahramanmaraş. Gempa dahsyat berkekuatan 7,8 magnitudo tersebut menghancurkan bangunan-bangunan di bagian kota Turki Selatan hingga Suriah, gempa tersebut diikuti oleh gempa berkekuatan 7,5 beberapa jam kemudian pada pukul 13:24 waktu setempat. World Health Organization (WHO) melaporkan bahwa 15 juta orang terkena dampak di Turki dan 10,87 juta orang terkena dampak di Suriah (World Health Organization, 2023).

Menurut data pada Statista yang diperoleh dari *NTV Turkey; Kandilli Observatory and Earthquake Research Institute* (2023), gempa tersebut tercatat menjadi gempa di Turki yang menelan korban jiwa terbanyak dari tahun 1914 hingga tahun 2023, hal tersebut dibuktikan pada Gambar 1.



Gambar 1. Grafik Korban Jiwa Gempa Turki Tahun 1914 -2023

Sumber : BBC

Melihat kondisi tersebut tentu memunculkan banyak pertanyaan tentang bagaimana sentiment masyarakat dalam

platform media sosial Twitter mengenai peristiwa gempa bumi di Turki tersebut. Penelitian analisis sentimen terhadap gempa Turki melalui media sosial Twitter ini bertujuan untuk memahami berbagai pendapat yang muncul dari para pengguna internet, salah satunya dari media sosial Twitter mengenai bencana gempa di Turki yang belakangan ini ramai diperbincangkan. Dengan adanya data hasil analisis bencana gempa di Turki ini diharapkan dapat memperoleh wawasan mengenai pandangan atau opini yang diungkapkan di platform media sosial dengan metode ilmiah supaya pandangan atau opini yang tersedia menyajikan informasi yang mudah dipahami dan tidak menyebabkan kebingungan saat menerima dan memahami informasi tersebut.

Pemanfaatan data analisis sentimen untuk lembaga UNHCR (*United Nations High Commissioner for Refugees*) dapat memberikan manfaat besar dalam upaya penanganan pengungsi dan populasi terdampak bencana. Data analisis sentimen dapat membantu dalam mengidentifikasi kebutuhan mendesak dan prioritas masyarakat yang terdampak bencana. Melalui analisis sentimen, lembaga UNHCR dapat memahami aspirasi, kekhawatiran, dan masalah utama yang dihadapi oleh korban bencana gempa. Informasi ini memungkinkan untuk mengalokasikan sumber daya secara lebih efisien dan memberikan respon dengan tepat sasaran.

Lexicon based adalah metode analisis sentimen yang populer di media sosial karena kemudahan metode yang digunakan. Metode berbasis leksikon menggunakan kamus yang sudah berisi nilai polaritas positif, dan negatif guna memberikan pembobotan atau penilaian pada kata (Undap et al., 2021). Pendekatan berbasis leksikon adalah metode analisis sentimen yang menggunakan kamus yang menyertakan atau mencantumkan kata dengan polaritas kata yang disertakan di dalamnya (Malini et al., 2022). Menurut (Chiny et al., 2021) VADER adalah

leksikon berbasis aturan dan alat analisis sentimen yang secara khusus disesuaikan dengan sentimen yang diekspresikan di media sosial yang menggunakan kombinasi leksikon sentimen dan daftar fitur leksikal yang umumnya dilabeli menurut orientasi semantik sebagai positif atau negatif.

Naïve Bayes Classifier merupakan algoritma yang menggunakan konsep peluang, atau yang biasa dikenal dengan probabilitas, dan digunakan untuk klasifikasi dalam analisis sentiment (Amaliah & Dwi Nuryana, 2022). *Naïve Bayes Classifier* merupakan pengklasifikasi probabilitas sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data yang diberikan (Mustofa & Prasetyo, 2021).

Pada penelitian terdahulu yang dilakukan oleh (Aulia & Patriya, 2019) pendekatan dengan *Lexicon Based* pada data testing menghasilkan nilai akurasi 64,49% kemudian pada data *training* menggunakan labelisasi menghasilkan 94,2%. Dengan *Naïve Bayes Classifier* pada data *testing* menghasilkan nilai akurasi sebesar 86,5% dan pada data *training* sebesar 94,08%. Penelitian ini menunjukkan bahwa hasil dengan akurasi tertinggi adalah dengan *Naive Bayes Classifier*. Penelitian berikutnya oleh (Kardian & Gustiana, 2021) menghasilkan akurasi sebesar 71% dengan memanfaatkan pendekatan *Lexicon Based* dan algoritma klasifikasi *Naïve Bayes* dengan data tweet yang diolah sebanyak 1.261 data. Data *training* yang digunakan sebesar 90% dari 1.135 data dan data *testing* 10% dari 126 data.

Penelitian berikutnya dilakukan oleh (Undapet al., 2021) penelitian ini berusaha menganalisis situs pembajakan artikel dengan metode *Lexicon Based* dengan klasifikasi sentimen positif, negatif, dan netral. Kemudian performa dari *Lexicon Based* tidak diukur dengan metode lain menghasilkan skor pengujian akurasi sebesar 55,6% dengan menggunakan 3.108 data, dimana 1.730 data berhasil diprediksi

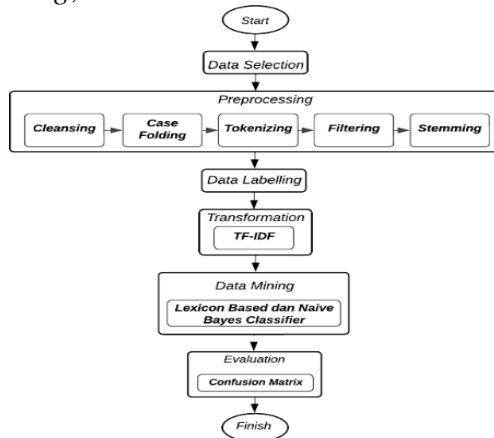
benar oleh sistem. Penelitian selanjutnya dilakukan oleh (Oktaviana et al., 2022) hasil penelitian ini menunjukkan penggunaan *Lexicon Based Features* mempengaruhi nilai akurasi objek dalam penelitian. Hasil evaluasi dalam identifikasi kelas sentimen pada data komentar pengguna Twitter meningkat setelah menggunakan *Lexicon Based Features*. Nilai akurasi sebelum menggunakan *Lexicon Based Features* 0.48, kemudian setelah menggunakan *Lexicon Based Features* sebesar 0.60. Nilai evaluasi meningkat 12% dibandingkan metode SVM tanpa menggunakan pendekatan *Lexicon Based Features*. Penelitian yang dilakukan ini menggunakan pendekatan *Lexicon Based* dan algoritma *Naive Bayes Classifier* untuk melakukan analisis sentimen pengguna Twitter tentang peristiwa bencana gempa bumi yang di Turki pada 6 Februari 2023 lalu

METODE

Penelitian ini mengimplementasikan metodologi KDD (*Knowledge Discovery in Database*). *Knowledge Discovery in Database* merupakan proses komputasi yang menggunakan algoritma matematika untuk mengekstraksi data dan menghitung probabilitas tindakan potensial di masa mendatang (Bhatia, 2019). *Knowledge Discovery in Database* sebuah proses untuk menemukan pola menarik yang tersembunyi dari sekumpulan data besar yang disimpan dalam basis data seperti *data warehouse* dan tempat penyimpanan data lainnya (Rahmah, 2021).

Metode yang disebut *Knowledge Discovery in Database* merupakan proses mencari pengetahuan baru tentang domain area aplikasi tertentu, model *Knowledge Discovery in Database* mencakup sejumlah proses yang dilakukan secara berurutan (Ghazal & Hammad, 2022). Penggunaan metode *Knowledge Discovery in Database* diperlukan karena data yang digunakan berisi data yang beragam dan tidak terstruktur. Metodologi *Knowledge*

Discovery in Database terdiri dari tahapan berikut yaitu *data selection*, *preprocessing*, *data labelling*, *transformation*, *data mining*, dan *evaluation*.



Gambar 1. Alur Tahap Penelitian

1. Data Selection

Proses scraping data atau pengumpulan data diambil menggunakan library *Snsrape* yang terdapat pada python, untuk menarik data berupa tweet pada Twitter. *Snsrape* adalah *scraper* yang dapat membantu mengambil informasi tanpa perlu menggunakan API dari Twitter. Informasi yang dapat diambil seperti dari profil pengguna, tagar, atau seperti postingan yang relevan dengan pencarian.

Setelah scraping selesai selanjutnya adalah melakukan proses *data selection* dengan melakukan pemilihan tweet dari dataset dengan tujuan mengumpulkan data yang berhubungan dengan objek penelitian yaitu tweet tentang bencana gempa di Turki yang didapatkan dari platform penyedia dataset Kaggle dan tweet sosial media Twitter yang mengandung peristiwa gempa di Turki pada Februari 2023. Pemilihan data dilakukan dengan menggunakan kata kunci “*turkey earyhquake*” dan kriteria pemilihan tweet yaitu tweet yang ditulis dengan bahasa Inggris. Setelah mendapatkan data dari proses tersebut, kemudian dilakukan proses pemilihan data. Dalam proses pemilihan data ini tweet akan diseleksi dan kemudian tweet yang sudah sesuai akan masuk ke proses *preprocessing* sesuai dengan kebutuhan penelitian ini.

2. Preprocessing

Preprocessing merupakan fase pengolahan data yang berfungsi menghasilkan dataset yang bersih dari *noise* sehingga siap digunakan pada tahapan pemodelan selanjutnya. *Preprocessing* adalah tahap pemrosesan data yang menciptakan dataset yang bebas *noise* dan data siap untuk digunakan pada tahap pemodelan berikutnya. Tahapan dalam preprocessing dalam penelitian ini sebagai berikut.

- Cleansing:** Melakukan pembersihan data dengan menghapus emotikon, tagar, url, nama pengguna, atau ikon yang tidak diinginkan untuk memudahkan proses penambangan data.
- Case Folding:** Mengubah kata-kata yang masih menggunakan huruf besar menjadi huruf kecil. Pada tahap case folding ini, suatu kata, teks atau kalimat yang mengandung huruf kapital, maka huruf kapital tersebut akan diubah menjadi huruf kecil (*lowercase*).
- Tokenizing:** Memecah kalimat dalam teks menjadi kata-kata tunggal untuk menentukan struktur setiap kata. Umumnya, kata diidentifikasi atau dipisahkan oleh spasi, tanda kutip tunggal ('), titik (.), titik koma (;), dan titik dua (:).
- Filtering:** Menyaring kata penghubung yang tidak diperlukan di dalam dokumen, misalnya “*if*”, “*when*”, “*and*”, “*or*”, “*however*” “*although*”, dan lain-lain. Pada tahap *filtering* ini kata umum yang tidak memiliki makna atau informasi akan dihilangkan agar volume kata yang disimpan pada sistem berkurang.
- Stemming:** Menghilangkan kata imbuhan yang terdapat pada suatu kata atau merubah suatu kata yang memiliki imbuhan menjadi kata dasar dengan tujuan menyiapkan data untuk diolah ke tahap selanjutnya.

3. Data Labelling

Setelah melewati tahapan preprocessing, maka selanjutnya dilakukan

proses *data labelling*. *Data labelling* untuk pendekatan dengan *Lexicon Based* dilakukan secara otomatis, sistem akan menghitung skor suatu sentimen dengan bantuan kamus leksikon yaitu *VADER Sentiment*.

4. Transformation

Tahap transformasi merupakan tahap pengubahan data menjadi format yang dapat diproses selama tahap data mining. Transformasi data diperlukan karena pada proses selanjutnya pencarian data numerik akan diperlukan untuk memudahkan perhitungan selama penerapan kata-kata yang memiliki frekuensi yang lebih tinggi. Selain itu, metode ini juga membantu dalam mengidentifikasi kata yang sering muncul dalam dokumen.

$$TF - IDF = tf \times idf$$

$$tf = \frac{tf_{ij}}{\sum t_j}$$

$$idf = \log \frac{N}{df}$$

Keterangan:

tf : Term Frequency (frekuensi kata.

tf_{ij} : Jumlah kemunculan term i dalam dokumen j.

$\sum t_j$: Jumlah semua term dalam dokumen j.

Idf : Inverse Document Frequency.

N : Jumlah total dokumen dalam corpus N.

Df : Jumlah dokumen mengandung term sebuah kata.

5. Data Mining

a. Lexicon Based

Menurut (Malini et al., 2022) pendekatan berbasis leksikon adalah metode analisis sentimen yang menggunakan kamus yang menyertakan atau mencantumkan kata dengan polaritas kata yang disertakan di dalamnya. Dengan metode ini sistem akan memeriksa setiap kata, kemudian sebuah kata dikatakan positif jika kata tersebut masuk dalam kamus positif. Demikian juga, jika sebuah kata dimasukkan dalam kamus negatif, kata

tersebut adalah kata negatif. Data teks dapat dikategorikan oleh VADER menjadi tiga kategori sentimen berbeda yaitu positif, negatif, dan netral. Untuk melakukan proses klasifikasi sentimen tersebut dilakukan oleh VADER dengan menghitung skor negatif, positif, netral, dan compound score berdasarkan nilai polaritas.

$$CS = \frac{x}{\sqrt{x^2 + a}}$$

Keterangan:

CS : Compound Score.

X : Jumlah nilai valensi kata penyusun.

a : Konstanta normalisasi.

b. Naïve Bayes Classifier

Salah satu metode penambangan data yang paling populer untuk mengkategorikan sejumlah besar data adalah *Naive Bayes Classifier*, yang juga dapat digunakan untuk memperkirakan kemungkinan probabilitas kelas akan terdiri dari sekumpulan data tertentu. Algoritma *Multinomial Naive Bayes* dipilih untuk melakukan klasifikasi data ini karena algoritma ini dapat mengumpulkan bukti, evidensi dan atribut kelas, serta melatih model menggunakan sampel data untuk menentukan probabilitas jika diberikan kelas tertentu. Bentuk formula teorema Bayes adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H) P(H)}{P(X)}$$

Keterangan:

X : Kelas data yang tidak diketahui.

H : Hipotesis data X, kelas sesuai dengan kategori data tertentu.

P(H|X) : Kemungkinan hipotesis (probabilitas) H sesuai kondisi X.

P(H) : Kemungkinan hipotesis H (probabilitas hipotesis H).

6. Evaluation

Pada tahapan ini akan dilakukan pengujian dengan menggunakan *Confusion Matrix*. Hasilnya akan berupa nilai akurasi. Confusion matrix adalah metode pengujian yang dilakukan dalam evaluasi yang dirancang untuk menghitung kinerja pembelajaran mesin. Menurut (Kardian &

Gustiana, 2021) *confusion matrix* merupakan tabel yang menampilkan seberapa baik kinerja model klasifikasi pada set data uji dengan nilai yang diketahui. Matriks ini memiliki empat kombinasi berbeda dari hasil tes prediksi dan nilai aktual. Uji akurasi confusion matrix digunakan untuk menentukan akurasi model pembelajaran mesin dari *Naive Bayes Classifier* (Mustofa & Prasetyo, 2021). Dalam penelitian, digunakan multiclass confusion matrix 3x3, sehingga tabel confusion matrixnya seperti pada Tabel 1 berikut.

Tabel 1. Confusion Matrix

	Prediksi			
	Negatif		Netral	Positif
	Negatif	TNeg	FNegNet	FNegPos
	Netral	FNetNeg	TNet	FNetPos
Aktual	Positif	FPosNeg	FPosNet	TPos

Keterangan:

TNeg : Jumlah prediksi negatif dari data aktualnegatif.

FNegNet : Jumlah prediksi yang netral dari dataaktual yang negatif.

FNegPos : Jumlah prediksi yang positif dari dataaktual yang negatif.

FNetNeg : Jumlah prediksi yang negatif dari dataaktual yang netral.

TNet : Jumlah prediksi yang netral dari data aktual yang netral.

FNetPos : Jumlah prediksi yang positif dari dataaktual yang netral.

FPosNeg : Jumlah prediksi yang negatif dari dataaktual yang positif.

FPosNet : Jumlah prediksi yang netral dari dataaktual yang positif.

TPos : Jumlah prediksi yang positif dari dataaktual yang positif.

Akurasi adalah perbandingan jumlah proporsi prediksi yang benar di kotak positif negatif, dan netral. Untuk menghitung akurasi dengan tabel multiclass confusion matrix 3x3 dihitung menggunakan persamaan seperti yang digunakan oleh (Arini et al., 2020) ditentukan denganpersamaan berikut:

$$Akurasi = \frac{TPos + TNeg + TNet}{TPos + FPosNeg + FPosNet + FNegPos + TNeg + FNetNeg + FNetPos + FNetNeg + TNet}$$

HASIL DAN PEMBAHASAN

Dari hasil penelitian yang sudah dilaksanakan yaitu adalah melakukan

analisis sentimen dari datatweet mengenai isu bencana gempa yang terjadi di Turki pada 6 Februari 2023. Kemudian data tweet yang didapat diklasifikasikan ke dalam 3 kelas sentimen yaitu positif, negatif dan netral dengan menerapkan metode *Lexicon Based* menggunakan kamus *VADER Sentiment* dan pelalelan data manual oleh pakar untuk membuktikan hasil perbandingan akurasi pendekatan *Lexicon Based* dengan *Naive Bayes Classifier*, dan proses klasifikasi dengan algoritma *Naive Bayes Classifiertanpa* pendekatan *Lexicon Based*. Kemudian sistem dievaluasi dengan memanfaatkan *confusion matrix* untuk menentukan nilai akurasi model yang telah disiapkan sebelumnya.

1. Data Selection

Proses *scraping* data berupa tweet yang berasal dari sosial media Twitter dengan penggunaan kata kunci “*turkey earthquake*” dan tweet yang diambil dalam bahasa Inggris. Setelah proses *scraping* data dilakukan, proses selanjutnya yaitu melakukan penyeleksian data dari hasil *scraping* dengan tujuan menentukan apakah data tersebut sudah sesuai dengan kata kunci yang digunakan pada penelitian ini. Setelah melakukan seleksi data, data awal yang berjumlah 2.376 data tweet berkurang menjadi 1.028 data tweet.

2. Preprocessing

Dalam tahap preprocessing ini data yang sudah dipilih dan diseleksi akan dibersihkan dengan menghapus *noise* yang terdapat pada data. Dalam penelitian ini, tahap *preprocessing* terdiri dari 5 tahapan yaitu *cleaning*, *case folding*, *tokenizing*, *filtering*, dan *stemming*.

Tabel 1. Preprocessing

Tahapan	Hasil
Data Tweet	Alkhidmat Foundation Pakistan @AlkhidmatOrg in collaboration with Turkish partner organizations is responding on the ground... https://t.co/1gbCqJ9lwd
Cleaning	Alkhidmat Foundation Pakistan in collaboration with Turkish partner organizations is responding on the ground

<i>Case Folding</i>	alkhidmat foundation pakistan in collaboration with turkish partner organizations is responding on the ground
<i>Tokenizing</i>	[alkhidmat, foundation, pakistan, in, collaboration with, Turkish, partner, organizations, is, responding, on, the, ground]
<i>Filtering</i>	[alkhidmat, foundation, pakistan, collaboration, turkish, partner, organizations, responding, the ground]
<i>Stemming</i>	[alkhidmat, foundat, pakistan, collabor, turkish, partner, organ, respond, the ground]

3. Data Labelling

Langkah selanjutnya proses *data labelling*, pendekatan dengan *Lexicon Based* digunakan untuk memberi label setelah data dibersihkan selama tahapan *preprocessing* dengan memanfaatkan *VADER Sentiment*. Sentimen positif memiliki skor polaritas lebih besar dari nol (> 0), sedangkan nilai sentimen netral memiliki skor polaritas sama dengan nol ($= 0$), dan nilai sentimen negatif memiliki skor polaritas kurang dari nol (< 0) seperti penelitian yang dilakukan oleh (Akbari et al., 2023) dan penelitian (Syah et al., 2023).

tweet_final	Positif	Negatif	Netral	compound	Sentiment
earthquak leav 4 300 dead turkey syria peopl l...	0.000	0.452	0.548	-0.7650	negatif
god pleas help sad	0.539	0.348	0.112	0.1779	positif
alkhidmat foundat pakistan collabor turkish pa...	0.000	0.000	1.000	0.0000	netral
syria voic	0.000	0.000	1.000	0.0000	netral
may allah protect turkey syria	0.394	0.000	0.606	0.3818	positif
prayer turkey syria lebanon may allah help pro...	0.371	0.000	0.629	0.6486	positif

Gambar 1. Data Labelling dengan VADER Sentiment

Pada Gambar 1 diatas menunjukkan hasil pelabelan data menggunakan *VADER Sentiment*. Data tweet tersebut dikategorikan menurut hasil skor polaritas sentimen masing-masing yaitu positif, netral, dan negatif.

4. Transformation

Tahapan berikutnya adalah transformation, pada proses transformation ini tiap kosakata yang ada pada dokumen akan dilakukan pembobotan kata (perhitungan token pada setiap dokumen) dengan menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) memanfaatkan fitur *TfidfVectorizer* yang disediakan oleh library sklearn. Hasil ringkasan dataset yang telah dihitung tf-idf dapat dilihat pada Tabel 2.

Tabel 2. Hasil pembobotan TF-IDF

Term	Rank
Turkey	73.517630
Earthquake	66.338475
Syria	51.336146
Help	28.297637
People	28.004981

Tabel 2 menunjukkan bahwa Pada langkah ini, pembobotan dilakukan dengan menggunakan perhitungan TF-IDF untuk setiap kata dalam dataset, kemudian pengurutan dilakukan berdasarkan nilai rank terbesar. Dengan mengambil lima kata paling sering muncul yang didapat dari hasil visualisasi *wordcloud* untuk meringkas hasil transformasi tersebut.

5. Data Mining

Setelah dilakukan tahap transformasi pada data, selanjutnya akan dilakukan tahap *data mining*. Dalam tahapan *data mining* ini, dilakukan klasifikasi sentimen terhadap tweet opini pengguna Twitter terhadap bencana gempa Turki yang terjadi pada 6 Februari 2023 lalu menggunakan dua model yaitu *Lexicon Based* dengan *Naïve Bayes Classifier* (model 1), dan *Naïve Bayes Classifier* tanpa pendekatan *Lexicon Based* (model 2). Proses implementasi kedua model ini dilakukan dengan menggunakan tiga skenario pengujian yang telah dijelaskan sebelumnya pada tahapan *Transformation*. Dalam penelitian ini pengujian akan dilakukan dengan tiga skenario pembagian data. Berikut tiga skenario pembagian data yang dilakukan dan hasil akurasi yang didapatkan dari setiap skenario yang

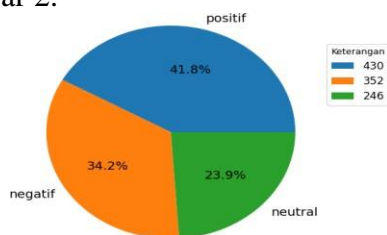
diringkas ke dalam Tabel 3 berikut.

Tabel 3. Pembagian Data Training dan Data Testing

Skenario Perbandingan	Data Trainig	Data Testing
90:10	925	103
80:20	822	206
70:30	719	309

a. Lexicon Based Features

Hasil analisis sentimen dengan menggunakan Lexicon Based Features sebagian besar didominasi dengan sentimen positif. Hasil klasifikasi menggunakan pendekatan *Lexicon Based* dengan bantuan *VADER Sentiment* ditunjukkan pada Gambar 2.

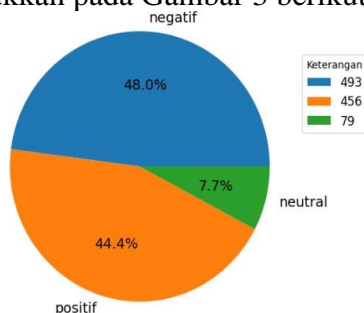


Gambar 2. Diagram Pie Pendekatan *Lexicon Based* dan *Naïve Bayes Classifier*

Gambar 2 memperlihatkan bahwa klasifikasi dengan pendekatan Lexicon Based menghasilkan persentase nilai polaritas sentimen positif sebesar 41.83%. Kemudian sentimen negatif menghasilkan nilai polaritas sebesar 34.93%, dan sentimen netral menghasilkan nilai polaritas sebesar 23.93% dengan 430 sentimen positif, 246 sentimen netral, dan 352 sentimen negatif.

b. Naïve Bayes Classifier

Tahap ini merupakan proses klasifikasi data yang dilakukan menggunakan algoritma *Multinomial Naïve Bayes*. Hasil klasifikasi menggunakan *Naïve Bayes Classifier* ditunjukkan pada Gambar 3 berikut.



Gambar 3. Diagram Pie *Naïve Bayes Classifier*

Pada Gambar 3 klasifikasi dengan *Naïve Bayes* menghasilkan persentase nilai polaritas sentimen positif sebesar 44.4%. Kemudian sentimen negatif menghasilkan nilai polaritas sebesar 48%, dan sentimen netral menghasilkan nilai polaritas sebesar 7.7% dengan 456 sentimen positif, 493 sentimen negatif, dan 79 sentimen netral.

6. Evaluation

Tahap terakhir setelah klasifikasi adalah evaluasi. Pengujian dilakukan dengan menggunakan *confusion matrix*. Pengujian dilakukan dengan menampilkan hasil klasifikasi pendekatan *Lexicon Based* dan *Naïve Bayes Classifier*, dan hasil klasifikasi *Naïve Bayes* tanpa menggunakan pendekatan *Lexicon Based* dengan menghitung akurasi. Untuk tujuan penelitian ini, parameter yang digunakan adalah akurasi yang dihitung menggunakan *confusion matrix* dengan persamaan *confusion matrix* untuk 3 kelas sentimen yaitu positif, negatif, dan netral. Tabel 4, Tabel 5, dan Tabel 6 menampilkan hasil akurasi yang didapatkan oleh model dengan skenario pembagian data yang sudah ditentukan.

Tabel 4. Pengujian Skenario 90:10

ConfusionMatrix	<i>Lexicon Based Features</i> dengan <i>Naïve Bayes Classifier</i>	<i>Naïve Bayes Classifier</i>
Tneg	22	31
FNegNet	1	0
FNegPos	7	14
Confusion Matrix	<i>Lexicon Based Features</i> dengan <i>Naïve Bayes Classifier</i>	<i>Naïve Bayes Classifier</i>
FNetNeg	10	3
TNet	4	0
FNetPos	10	4
FPosNeg	8	16
FPosNet	0	0
TPos	41	35
Akurasi	65%	64%

Tabel 5. Pengujian Skenario 80:10

Confusion Matrix	<i>Lexicon Based Features</i> Dengan <i>Naïve Bayes Classifier</i>	<i>Naïve Bayes Classifier</i>
------------------	--	-------------------------------

<i>Bayes Classifier</i>		
Tneg	56	76
FNegNet	2	0
FNegPos	16	27
FNetNeg	14	6
TNet	4	0
FNetPos	10	0
FPosNeg	21	28
FPosNet	2	0
TPos	73	63
Akurasi	67%	67%

Tabel 6. Pengujian Skenario 90:10

Confusion Matrix	Lexicon Based Features dengan Naïve Bayes Classifier	Naïve Bayes Classifier
Tneg	83	105
FNegNet	4	0
FNegPos	21	43
FNetNeg	26	12
TNet	14	0
FNetPos	25	13
FPosNeg	19	34
FPosNet	2	0
TPos	115	102
Akurasi	69%	67%

Pada Tabel 4, Tabel 5, dan Tabel 6 dapat dilihat bahwa model *Lexicon Based* dengan *Naïve Bayes Classifier*, dan model *Naïve Bayes Classifier* tanpa pendekatan *Lexicon Based* memberikan hasil yang cukup baik dalam melakukan klasifikasi pada ketiga skenario yang sudah dibuat, mendapatkan nilai akurasi rata-rata yang cukup tinggi. Dengan nilai rata-rata akurasi untuk model dengan pendekatan *Lexicon Based* dan *Naïve Bayes Classifier* mendapatkan rata-rata akurasi sebesar 67%, sedangkan model *Naïve Bayes Classifier* tanpa menggunakan pendekatan *Lexicon Based* mendapatkan rata-rata akurasi sebesar 66%.

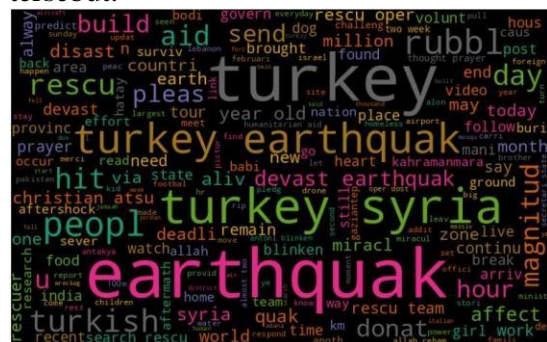
Beberapa kata sering muncul dalam pendapat yang bersifat positif, negatif, dan netral ditampilkan dalam wordcloud. Berdasarkan hasil dari visualisasi wordcloud bahwa frekuensi kata “*Turkey Earthquake*”, “*Syria*”, “*Earthquake*”, dan lain-lain banyak ditemukan dalam setiap sentimen. Hasil dari visualisasi *wordcloud* dikarenakan setiap tweet yang berisi istilah tersebut menggunakan sejumlah kata tertentu, dan semakin banyak kata yang

digunakan, ukuran teks di dalam gambar bertambah besar.



Gambar 4. *Wordcloud* Sentimen Positif

Pada Gambar 4 yang merupakan *wordcloud* yang termasuk dalam kelas sentimen positif banyak membicarakan mengenai “*donation*”, “*support*”, “*please*”, “*help*”, “*aid*”, “*thank*” dan lain-lain. Tweet tersebut berupa kata-kata yang diharapkan masyarakat dengan cara memberikan bantuan agar bencana yang terjadi tersebut cepat teratasi dan meringankan penderitaan korban bencana di daerah yang terdampak tersebut.



Gambar 5. Wordcloud Sentimen Netral

Pada Gambar 5 *wordcloud* sentimen netral, kata “*people*”, “*state*”, “*remain*”, dan kata lainnya yang ada pada *wordcloud* sentimen netral banyak terdeteksi di tweet dengan sentimen netral. Tweet tersebut lebih fokus pada pembahasan fakta dan berita daripada mengungkapkan sentimen emosional atau sikap yang kuat.



Gambar 6. Wordcloud Sentimen Negatif

Pada Gambar 6 *wordcloud* sentimen negatif yaitu “death”, “kill”, “victim”, dan lainnya. Tweet tersebut mengandung frasa-frasa yang negatif yang cenderung memicu provokasi, kemarahan, dan kritikan.

Hasil temuan penelitian ini menunjukkan bahwa pelabelan dengan *VADER Sentiment* menghitung nilai polaritas kata perkata, sedangkan pelabelan oleh ahli bahasa dilakukan dengan menentukan suatu kalimat tersebut masuk ke dalam kelompok sentimen yang memiliki kecenderungan yang lebih mendominasi. Kekurangan dalam penggunaan *VADER Sentiment* tanpa proses normalisasi selama tahap preprocessing membuat sentimen yang dilabeli oleh mesin dengan bantuan *VADER Sentiment* memberikan hasil yang kurang maksimal dikarenakan jumlah sentimen yang masuk ke kelas positif dan negatif berbeda dengan pelabelan yang sudah dilakukan oleh pakar bahasa

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat diambil kesimpulan yaitu:

1. Dalam analisis sentimen terhadap Gempa Turki menggunakan pendekatan *Lexicon Based* dan *Naïve Bayes Classifier* berhasil diimplementasikan dengan baik dalam mengklasifikasi sentimen positif 41.83%, sentimen negatif 34.93%, dan sentimen netral 23.93%. Untuk model *Naïve Bayes Classifier* tanpa pendekatan *Lexicon Based* dengan pelabelan pakar mendapatkan hasil klasifikasi sentimen dengan 44% positif, 48% negatif, dan 7.7% netral. Hasil klasifikasi ini memiliki perbedaan dikarenakan pelabelan *VADER Sentiment* menghitung nilai polaritas kata perkata, sedangkan pelabelan oleh pakar dengan menentukan kalimat tersebut cenderung masuk ke dalam sentimen mana yang lebih dominan.
2. Berdasarkan hasil pengujian pendekatan *Lexicon Based* dan *Naïve Bayes Classifier* dengan tiga skenario

pengujian yang sudah dibuat, didapatkan nilai rata-rata akurasi sebesar 67% untuk model yang dibuat. Sedangkan hasil pengujian *Naïve Bayes Classifier* tanpa menggunakan pendekatan *Lexicon Based*, memperoleh nilai rata-rata akurasi dengan nilai 66% untuk model *Naïve Bayes Classifier* yang dibuat tanpa pendekatan *Lexicon Based* pada sentimen analisis terhadap Gempa Turki.

DAFTAR PUSTAKA

- Akbari, W. A., Tukino, T., Huda, B., & Muslih, M. (2023). Sentiment Analysis of Twitter User Opinions Related to Metaverse Technology Using Lexicon Based Method. *Sinkron*, 8(1), 195–201. <https://doi.org/10.33395/sinkron.v8i1.11992>
- Amaliah, F., & Dwi Nuryana, I. K. (2022). Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter. *Journal of Informatics and Computer Science (JINACS)*, 3(03), 384–393. <https://doi.org/10.26740/jinacs.v3n03.p384-393>
- Arini, A., Wardhani, L. K., & Octaviano, D. (2020). Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character Menggunakan Algoritma Naïve Bayes Classifier (Nbc) Pada Tweet Hashtag #2019gantipresiden. *KILLAT*, 9(1), 103–114. <https://doi.org/10.33322/kilat.v9i1.878>
- Aulia, G. N., & Patriya, E. (2019). Implementasi Lexicon Based Dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019. *Jurnal Ilmiah Informatika Komputer*, 24(2), 140–153. <https://doi.org/10.35760/ik.2019.v24i2.2369>

- Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. In *Studies in Computational Intelligence* (Vol. 639, hal. 313–340). Springer Verlag. https://doi.org/10.1007/978-3-319-30319-2_13
- Bhatia, P. (2019). *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge University Press. <https://doi.org/https://doi.org/10.1017/9781108635592>
- Chiny, M., Chihab, M., Bencharef, O., & Chihab, Y. (2021). LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model. *International Journal of Advanced Computer Science and Applications*, 12(7), 265–275. <https://doi.org/10.14569/IJACSA.2021.0120730>
- Fan, X., Scaringi, G., Korup, O., West, A. J., van Westen, C. J., Tanyas, H., Hovius, N., Hales, T. C., Jibson, R. W., Allstadt, K. E., Zhang, L., Evans, S. G., Xu, C., Li, G., Pei, X., Xu, Q., & Huang, R. (2019). Earthquake-Induced Chains of Geologic Hazards: Patterns, Mechanisms, and Impacts. *Reviews of Geophysics*, 57(2), 421–503. <https://doi.org/10.1029/2018RG000626>
- Ghazal, M. M., & Hammad, A. (2022). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. *International Journal of Construction Management*, 22(9), 1632–1646. <https://doi.org/10.1080/15623599.2020.1738205>
- Kardian, A. R., & Gustiana, D. (2021). Analisis Sentimen Berdasarkan Opini Pengguna pada Medi Twitter Terhadap BPJS Menggunakan Metode Lexicon Based dan Naïve Bayes Classifier. *Jurnal Ilmiah Komputasi*, 20, 39–52. <https://doi.org/http://dx.doi.org/10.32409/jiks.tik.20.1.401>
- Malini, Z., Riskiyanti, C., Gede, I. K., Putra, D., Kt, A. A., & Cahyawan, A. (2022). Analisis Sentimen Pola Pikir Masyarakat Indonesia Terkait Virus Covid-19 Dalam Media Sosial Twitter Menggunakan Metode Rule Based Leksikon. *Jurnal Ilmiah Teknologi dan Komputer*, 3(1).
- Mustofa, R. L., & Prasetyo, B. (2021). Sentiment analysis using lexicon-based method with naïve bayes classifier algorithm on #newnormal hashtag in twitter. *Journal of Physics: Conference Series*, 1918(4). <https://doi.org/10.1088/1742-6596/1918/4/042155>
- Oktaviana, N. E., Sari, Y. A., & Indriati. (2022). Analisis Sentimen terhadap Kebijakan Kuliah Daring Selama Pandemi Menggunakan Pendekatan Lexicon Based Features dan Support Vector Machine. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 9(2), 357. <https://doi.org/10.25126/jtiik.2022925625>
- Rahmah, S. A. (2021). Klasterisasi Pola Penjualan Pestisida Menggunakan Metode K-Means Clustering (Studi Kasus Di Toko Juanda Tani Kecamatan Hutabayu Raja). *Djtechno: Jurnal Teknologi Informasi*, 1(1), 1–5. <https://doi.org/10.46576/djtechno.v1i1.964>
- Syah, R. I., Hoiriyah, & Walid, M. (2023). Analisis Sentimen Pengguna Media Sosial Terhadap Aplikasi M-Health Peduli Lindungi Dengan Metode Lexicon Based Dan Naïve Bayes. *Indonesian Journal of Business Intelligence Volume*, 6(1), 43–54.
- Undap, M. G., Rantung, V. P., & Rompas, P. T. D. (2021). Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode Lexicon-Based. *Jointer - Journal of Informatics Engineering*, 2(02), 39–

46.

<https://doi.org/10.53682/jointer.v2i02.44>

World Health Organization. (2023). *Earthquake response in Türkiye and Whole of Syria* (Nomor February) <https://www.who.int/publications/m/item/who-flash-appeal--earthquake-response-in-turkiye-and-whole-of-syria>.