ANALISIS SENTIMEN PUBLIK TERHADAP KEBIJAKAN TAPERA DI TWITTER: PENDEKATAN MACHINE LEARNING MENGGUNAKAN LOGISTIC REGRESSION

Dea Imelda Lika, Gunawan
Sistem Informasi, Universitas Muhammadiyah Bengkulu
Jalan Bali, Kampung Bali, Teluk Segara, Bengkulu 38119, Indonesia
deaemelda1@gmail.com

ABSTRAK

Kebijakan Tabungan Perumahan Rakyat (Tapera) menimbulkan beragam respons dari masyarakat, khususnya di media sosial X, yang mencerminkan persepsi publik terhadap kebijakan tersebut. Banyaknya opini yang berkembang menyebabkan sulitnya memahami kecenderungan sentimen masyarakat secara menyeluruh apabila dilakukan secara manual. Oleh karena itu, penelitian ini bertujuan untuk menganalisis sentimen publik terhadap kebijakan Tapera menggunakan pendekatan *machine learning*. Data penelitian diperoleh melalui teknik crawling pada media sosial X dengan kata kunci terkait Tapera, kemudian dilakukan proses pembersihan dan pra-pemrosesan teks. Pelabelan sentimen dilakukan secara otomatis menggunakan model bahasa IndoBERT dengan tiga kategori sentimen, yaitu positif, netral, dan negatif. Data selanjutnya dibobotkan menggunakan metode Term *Frequency-Inverse Document Frequency* (TF-IDF) dan diklasifikasikan menggunakan algoritma *Logistic Regression*. Hasil penelitian menunjukkan bahwa sentimen masyarakat terhadap kebijakan Tapera didominasi oleh sentimen netral dan negatif. Evaluasi model menghasilkan nilai akurasi sebesar 71,6% dengan performa yang relatif seimbang pada setiap kelas sentimen setelah dilakukan penyeimbangan data. Hasil ini menunjukkan bahwa pendekatan yang digunakan mampu memberikan gambaran sentimen publik terhadap kebijakan Tapera secara efektif.

Kata kunci: *Analisis Sentimen, Tapera, Media Sosial X, IndoBERT, Logistic Regression, TF-IDF*

1. PENDAHULUAN

Masifnya evolusi pada sektor perangkat digital telah merombak secara signifikan paradigma publik dalam mengakses, menyebarkan, dan menyampaikan informasi. Berbagai platform digital memungkinkan proses interaksi berlangsung dengan sangat cepat dan transparan, sehingga pendapat publik dapat terbentuk, berubah, dan berkembang dalam waktu singkat. Kondisi ini mendorong masyarakat menjadi lebih aktif berpartisipasi dalam diskusi publik melalui media digital. Salah satu dampak penting dari perkembangan tersebut adalah meningkatnya penggunaan media sosial sebagai sarana untuk berdiskusi, memberikan kritik, serta menanggapi berbagai isu yang muncul dalam kehidupan sosial [1].

Twitter (X) merupakan salah satu platform media sosial yang banyak digunakan masyarakat Indonesia untuk menyampaikan pendapat melalui unggahan teks singkat. Karakteristik platform ini yang bersifat *real-time* serta memiliki kemampuan penyebaran informasi yang cepat menjadikannya sumber data yang penting untuk memahami pola pandangan masyarakat. Dalam beberapa tahun terakhir, Twitter sering menjadi tempat berkembangnya diskusi publik mengenai kebijakan pemerintah, terutama kebijakan yang berpengaruh langsung terhadap kehidupan masyarakat secara luas. Oleh karena itu, platform ini semakin relevan sebagai bahan analisis terhadap opini masyarakat [2][3].

Salah satu kebijakan pemerintah yang menimbulkan perhatian dan diskusi luas di ruang publik digital adalah Tabungan Perumahan Rakyat (TAPERA). Kebijakan ini memunculkan beragam respons dari masyarakat, mulai dari dukungan hingga kritik. Kritik yang muncul umumnya berkaitan dengan isu pemotongan pendapatan pekerja, transparansi dalam pengelolaan dana, serta kejelasan manfaat jangka panjang yang akan diterima oleh peserta. Tingginya volume percakapan mengenai TAPERA di media sosial menunjukkan bahwa masyarakat memanfaatkan platform digital untuk menyampaikan pendapat dan menilai kebijakan pemerintah secara lebih terbuka dan kritis [4].

Banyaknya opini yang dihasilkan di media sosial menimbulkan permasalahan dalam proses analisis apabila dilakukan secara manual. Data yang tersedia umumnya bersifat tidak terstruktur, menggunakan bahasa informal, serta memiliki variasi ekspresi yang tinggi, sehingga menyulitkan proses pengelompokan dan penarikan kesimpulan secara objektif. Oleh karena itu, diperlukan pendekatan komputasional yang mampu mengolah data teks dalam jumlah besar secara sistematis agar kecenderungan sentimen masyarakat dapat dipahami secara akurat dan objektif.

Analisis sentimen merupakan metode yang umum digunakan untuk mengetahui sikap masyarakat berdasarkan informasi berbentuk teks dengan mengelompokkan opini ke dalam kategori sentimen positif, negatif, atau netral. Salah satu algoritma yang banyak digunakan dalam tugas klasifikasi teks adalah *Logistic Regression*. Algoritma ini dinilai efektif dan stabil, terutama bila digunakan bersama teknik representasi fitur seperti *Term Frequency-Inverse Document Frequency (TF-IDF)* yang mampu menonjolkan kata-kata penting dalam suatu dokumen [5][6].

Berdasarkan latar belakang dan permasalahan tersebut, tujuan penelitian ini adalah mengidentifikasi dan mengklasifikasikan sentimen masyarakat terhadap kebijakan TAPERA menggunakan data dari media sosial Twitter. Penelitian ini menerapkan algoritma *Logistic Regression* sebagai model klasifikasi utama [7].

Rencana penyelesaian masalah dilakukan melalui beberapa tahapan, yaitu pengumpulan data, pra-pemrosesan teks, pelabelan data, pembagian dataset, pembobotan kata menggunakan *TF-IDF*, penyeimbangan data, pembangunan model klasifikasi, serta evaluasi kinerja model. Seluruh tahapan tersebut dilakukan untuk memastikan bahwa hasil analisis mampu menggambarkan sentimen masyarakat dengan tingkat akurasi yang baik.

2. TINJAUAN PUSTAKA

2.1. Analisis sentiment

Analisis sentimen, atau opinion mining, merupakan proses identifikasi dan mengelompokkan opini dalam teks untuk menentukan sikap penulis, apakah positif, negatif, atau netral. Di media sosial, metode ini sangat penting untuk memahami persepsi publik secara otomatis terhadap berbagai isu. Secara umum, analisis sentimen fokus pada penerjemahan ekspresi subjektif seperti opini, sikap, dan emosi terhadap suatu objek, baik berupa produk, layanan, maupun kebijakan. Pendekatan ini mengubah data teks yang tidak terstruktur menjadi informasi yang lebih sistematis sehingga mempermudah analisis kecenderungan pandangan masyarakat [8][9].

2.2. Media sosial X (Twitter)

Media sosial *X* (sebelumnya dikenal sebagai Twitter) satu media komunikasi berani yang dipergunakan secara luas dalam berbagi informasi, menyampaikan pendapat, serta menanggapi berbagai isu sosial dengan cepat. Platform ini termasuk dalam jenis *mikroblog* , yaitu media yang memungkinkan penggunanya menulis dan membagikan pesan singkat atau tweet yang dapat dilihat oleh publik secara langsung Melalui fitur tersebut, pengguna dapat mengekspresikan pandangan, perasaan, dan reaksi mereka terhadap suatu peristiwa secara spontan dan terbuka. Kemampuan X dalam menyebarkan informasi secara *real-time* menjadikannya sumber data yang sangat potensial dalam penelitian yang sesuai dengan opini publik dan analisis sentimen di media sosial [10][11].

2.3. Machine Learning

Sebagai sub-bidang dalam ranah kecerdasan artifisial, pembelajaran mesin menitikberatkan pada pengembangan kemampuan perangkat digital untuk mengidentifikasi keteraturan atau skema tertentu di dalam kumpulan data secara otonom dan mempelajari data untuk menghasilkan prediksi atau keputusan secara otomatis. Melalui proses ini, sistem tidak diberikan aturan secara langsung, tetapi memperoleh kemampuannya berdasarkan pola yang ditemukan dari data pelatihan. Dengan kemampuan tersebut, machine learning sering digunakan dalam berbagai bidang, termasuk analisis data media sosial, karena dapat meningkatkan akurasi dalam mengklasifikasikan informasi serta memahami pola perilaku atau opini masyarakat [12].

2.4. Logistic Regression

Logistic Regression merupakan salah satu metode statistik yang digunakan untuk memprediksi kemungkinan suatu data termasuk ke dalam kategori tertentu berdasarkan variabel yang ada Metode ini bekerja dengan menghitung probabilitas melalui fungsi logistik, sehingga hasilnya dapat menunjukkan kecenderungan data terhadap kelas tertentu. Dalam analisis sentimen, algoritma ini sering digunakan untuk menentukan apakah suatu teks memiliki sentimen positif, negatif, atau netral. Selain mudah diterapkan, *Logistic Regression* juga memiliki tingkat efisiensi dan akurasi yang baik, sehingga banyak digunakan dalam penelitian berbasis pengolahan data teks[13].

2.5. Penelitian Terdahulu

Sejumlah penelitian telah memanfaatkan analisis sentimen berbasis Twitter untuk mengkaji opini publik terhadap berbagai isu. Husen et al. menerapkan algoritma *Support Vector Machine*, *Naïve Bayes*, dan *Logistic Regression* untuk menganalisis sentimen publik terhadap Bank BSI, dengan hasil bahwa *Logistic Regression* memiliki kinerja yang cukup baik dalam klasifikasi sentimen [14].

Penelitian oleh Sativa et al. menganalisis sentimen masyarakat terhadap kebijakan pemindahan Ibu Kota Negara Nusantara menggunakan *Naïve Bayes*, *Logistic Regression*, *dan K-Nearest Neighbors*, serta menunjukkan bahwa analisis sentimen mampu menggambarkan respon publik terhadap kebijakan nasional secara kuantitatif [15].

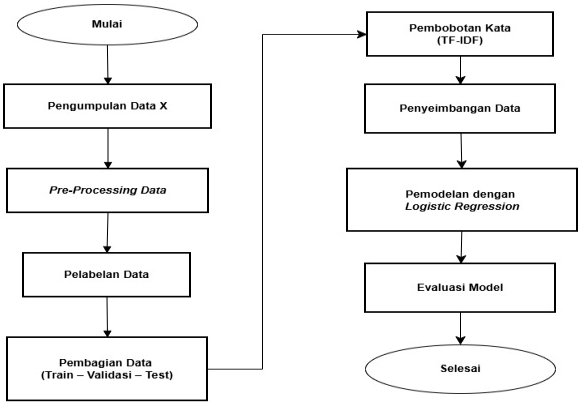
Selain itu, Anjani et al. mengkaji respons serta opini masyarakat terkait regulasi subsidi dan fluktuasi harga bahan bakar minyak (BBM) melalui Twitter dan menemukan dominasi sentimen negatif, yang menegaskan peran analisis sentimen sebagai alat evaluasi kebijakan publik [16].

Berdasarkan kajian tersebut, meskipun analisis sentimen terhadap kebijakan publik telah banyak dilakukan, penelitian yang secara khusus menggunakan *Logistic Regression* untuk identifikasi respon publik terhadap kebijakan TAPERA masih terbatas, sehingga penelitian ini memiliki relevansi untuk melengkapi kajian sebelumnya.

3. METODE PENELITIAN

Metode kuantitatif dimanfaatkan di riset ini dengan menerapkan metode machine learning untuk menganalisis sentimen publik terhadap kebijakan Tabungan Perumahan Rakyat (TAPERA) berdasarkan data dari media sosial X (Twitter). Algoritma yang

digunakan adalah *Logistic Regression*, yang diimplementasikan pada lingkungan Google Colab. Metode penelitian disusun secara berurutan mulai dari pengumpulan data hingga evaluasi model, sebagaimana ditunjukkan pada alur penelitian dalam Gambar 1.



Gambar 1. Alur penelitian

3.1. Pengumpulan Data X

Langkah awal riset dilakukan dengan penghimpunan data menggunakan teknik crawling terhadap media sosial X *(Twitter*). Proses crawling dilaksanakan dengan bantuan alat *tweet-harvest* berbasis *Node.js* yang proses melalui Google Colab. Pengambilan data dilakukan menggunakan Twitter Auth Token dengan kata kunci "tapera", "bp tapera", "kebijakan tapera", dan "iuran tapera".

3.2. Pra-Pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk eliminasi gangguan (*noise*) serta pengkondisian teks mentah, sebuah prosedur krusial agar informasi tersebut siap dikonsumsi dan dianalisis oleh algoritma machine learning secara optimal. Tahapan pra-pemrosesan yang diterapkan dalam penelitian ini meliputi:

a. Text Cleaning
   Merupakan tahap awal pra-pemrosesan yang bertujuan menghilangkan elemen non-linguistik, "tanda baca, simbol, tagar, URL, angka, dan karakter khusus". Tahap ini dilakukan agar teks yang dianalisis lebih bersih dan fokus pada informasi yang relevan untuk analisis sentimen.
b. Case Folding
   Tahap ini dilakukan dengan mengubah seluruh huruf dalam teks tweet menjadi huruf kecil untuk menyeragamkan representasi kata agar tidak terjadi perbedaan identifikasi antara variasi kapitalisasi kata dalam data set.
c. Tokenizing
   Tokenizing "merupakan pemrosesan pemecahan txt jadi unit kata atau token supaya setiap kata dapat dianalisis secara terpisah oleh model." Tahapan secara umum diterapkan dalam penelitian analisis sentimen untuk memecah kalimat menjadi elemen kata yang dapat diproses.
d. Normalization
   Normalization bertujuan untuk menyamakan bentuk kata tidak baku atau singkatan serta variasi bahasa informal menjadi bentuk standar sesuai kaidah tulisan bahasa Indonesia, sehingga meningkatkan konsistensi data teks untuk analisis lebih lanjut.
e. Stemming
   Stemming merupakan proses mengembalikan kata ke bentuk dasarnya dengan menghilangkan imbuhan. Proses ini membantu model untuk mengenali variasi kata yang memiliki akar kata sama sebagai satu representasi, sehingga dapat memperbaiki kualitas fitur yang digunakan dalam pemodelan.
f. Filtering
   Filtering dilakukan dengan menghapus kata-kata umum *(stopwords)* yang sering muncul tapi tidak berkontribusi terhadap penentuan sentimen seperti kata sambung atau kata penghubung. Penghilangan stopwords bertujuan menajamkan fitur kata yang relevan dengan sentimen.

3.3. Pelebelan Data

Segera setelah fase pra-pengolahan rampung, pelabelan otomatis diterapkan pada seluruh dataset menggunakan model IndoBERT, yang telah dioptimasi untuk mengenali karakteristik linguistik lokal. Implementasi model tersebut bertujuan untuk mengelompokkan data ke dalam tiga spektrum opini yang berbeda. Hasil klasifikasi akhir akan memisahkan teks ke dalam golongan sentimen positif, negatif, maupun kategori netral. Output nya menghasilkan dataset yang telah memiliki kelas sentimen yang jelas dan siap digunakan pada tahap pemodelan.

3.4. Pembagian Data

Dataset yang telah diberi label kemudian terbagi 2 klasifikasi, yaitu data latih dan data uji. Sebanyak 80% data digunakan sebagai data latih untuk membangun model, sedangkan 20% data digunakan sebagai data uji. Pembagian ini dilakukan untuk mengetahui kemampuan model dalam mengklasifikasikan sentimen pada data yang tidak terlibat dalam proses pelatihan.

3.5. Penyeimbangan Data

Tahap selanjutnya adalah penyeimbangan data, yang dilakukan karena total data perkelas sentimen tidak merata. Kondisi ini berpotensi menyebabkan model lebih cenderung mengenali kelas dengan jumlah data lebih besar. Oleh karena itu, diterapkan teknik penyeimbangan dengan menambah jumlah data pada kelas yang memiliki jumlah lebih sedikit, sehingga distribusi data antar kelas menjadi lebih proporsional.

3.6. Pembobotan Kata

Transformasi dataset tekstual ke dalam format vektor angka dilakukan melalui skema Term

Frequency–Inverse Document Frequency (TF-IDF). Teknik ini bertujuan untuk mengukur signifikansi setiap kata secara kuantitatif, sehingga setiap entri data memiliki bobot numerik berdasarkan tingkat kepentingan kata.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{t,d}} \quad (1)$$

$$\mathrm{IDF(t)} = \log\left(\frac{\mathrm{N}}{df_{\mathrm{t}}}\right) \quad (2)$$

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} \quad (3)$$

3.7. Pemodelan Dan Evaluasi

Tahap pemodelan dilakukan menggunakan algoritma Logistic Regression, yang dipilih karena memiliki kinerja yang stabil dan efektif dalam klasifikasi multikelas pada analisis sentimen. Model memprediksi probabilitas suatu kelas berdasarkan fungsi logistik sebagai berikut:

$$P(y \mid x) = \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}\right) \quad (4)$$

Verifikasi reliabilitas model dilakukan menggunakan confusion matrix dengan metrik accuracy, precision, recall, dan F1-score, yang dirumuskan sebagai berikut:

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

$$Precision = \frac{TP}{TP+FP} x\ 100\% \quad (6)$$

$$Recall = \frac{TP}{TP+FN} x\ 100\% \quad (7)$$

$$F1-Score = \frac{2 \times Precision \times Recall}{Precision + Recall} x\ 100\% \quad (8)$$

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Merujuk *output* pengumpulan data yang telah diimplementasikan, diperoleh sebanyak 6.439 tweet yang berkaitan dengan kebijakan Tabungan Perumahan Rakyat (Tapera) dari media sosial X (Twitter) pada periode September 2024 hingga Oktober 2025. Data yang diperoleh pada tahap ini masih berupa data mentah yang mengandung duplikasi, *retweet*, serta berbagai elemen teks seperti tagar, tautan, dan variasi bahasa. Untuk meningkatkan kualitas data, dilakukan proses pembersihan dengan menghapus tweet duplikat berdasarkan kesamaan isi teks. Setelah tahap ini, jumlah data berkurang menjadi 4.395 tweet unik yang selanjutnya digunakan sebagai dataset penelitian.

Tabel 1. Data Mentah Hasil Crawling Twitter

NO	Kata Kunci	Tweet
1.	Tapera	Siapa yang nyuruh beli rumah bangsat lu tunjangan Segede gitu ngeluh. Lah gw gaji ga seberapa dipotong Tapera bangsat bang.#Tapera #Program #Viral
2.	BP Tapera	BP Tapera: Masyarakat MBR di Sumatra Utara Rasakan Manfaat Nyata Rumah Subsidi https://t.co/aBFMTryoWb
3.	Kebijakan Tapera	Kebijakan Tapera yang bersifat wajib terkesan kurang peka terhadap realitas ekonomi pekerja terutama mereka yang berpenghasilan pas-pasan atau baru memulai karier. Ini berpotensi memperparah tekanan finansial di tengah kenaikan biaya hidup.#Kebijakan Tapera #Tapera
4.	Iuran Tapera	Rakyat gaji 3 jt masih bayar iuran BPJS iuran tapera makan minum cicil rumah bbm pulsa dll. Sisa ditabung mentok 500rb. Wakil rakyat gaji 4,2 jt utuh karena macem2nya dibayari negara. Tunjangan2 matamu tak tunjang cangkemmu sisan. Bukan gaji tapi bisa masuk pundi2. #Iuran Tapera #tapera

Dari Tabel 1 dapat dilihat bahwa data yang ditampilkan masih berupa data mentah hasil crawling dari "media sosial X (Twitter)". Tweet yang terkumpul masih mengandung unsur seperti tagar, tautan, bahasa tidak baku, serta ekspresi emosional pengguna.

4.2. Pra-Pemrosesan Data

Pra-pemrosesan data dilakukan untuk membersihkan dan menyiapkan teks tweet sebelum dianalisis. Tahapan terdiri "text cleaning, case folding, tokenizing, normalisasi, stemming, dan filtering". Proses tersebut menghasilkan teks yang lebih konsisten dan relevan, sehingga siap digunakan pada tahap pelabelan dan pemodelan sentimen.

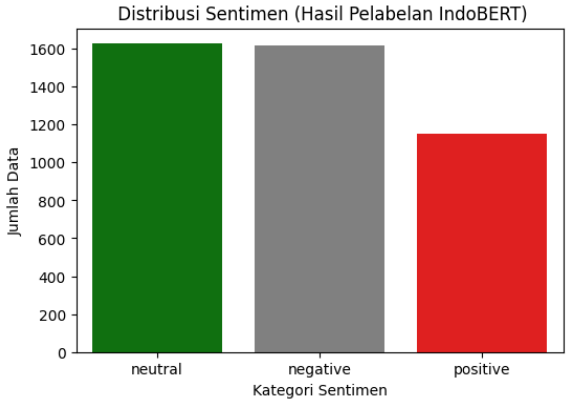
Tabel 2. Hasil Pra-Pemrosesan Data

Pra-Pemrosesan	Teks Sebelum	Teks Sesudah
Text Cleaning	"Siapa yang nyuruh beli rumah bangsat lu tunjangan Segede gitu ngeluh. Lah gw gaji ga seberapa dipotong Tapera bangsat bang.#Tapera #Program #Viral"	"Siapa yang nyuruh beli rumah bangsat lu tunjangan Segede gitu ngeluh Lah gw gaji ga seberapa dipotong Tapera bangsat bang"
Case Folding	"Siapa yang nyuruh beli rumah bangsat lu tunjangan Segede gitu ngeluh Lah gw gaji ga seberapa dipotong Tapera bangsat bang"	"siapa yang nyuruh beli rumah bangsat lu tunjangan segede gitu ngeluh lah gw gaji ga seberapa dipotong tapera bangsat bang"
Tokenizing	"siapa yang nyuruh beli rumah bangsat lu tunjangan segede gitu ngeluh lah gw gaji ga seberapa dipotong tapera bangsat bang"	['siapa', 'nyuruh', 'beli', 'rumah', 'bangsat', 'lu', 'tunjangan', 'segede', 'gitu', 'ngeluh', 'gaji', 'ga', 'seberapa', 'dipotong', 'tapera']

Pra-Pemrosesan	Teks Sebelum	Teks Sesudah
Normalization	['siapa', 'nyuruh', 'beli', 'rumah', 'bangsat', 'lu', 'tunjangan', 'segede', 'gitu', 'ngeluh', 'gaji', 'ga', 'seberapa', 'dipotong', 'tapera']	['siapa', 'suruh', 'beli', 'rumah', 'bangsat', 'tunjangan', 'besar', 'keluh', 'gaji', 'tidak', 'seberapa', 'potong', 'tapera']
Stemming	['siapa', 'suruh', 'beli', 'rumah', 'bangsat', 'tunjangan', 'besar', 'keluh', 'gaji', 'tidak', 'seberapa', 'potong', 'tapera']	['suruh', 'beli', 'rumah', 'bangsat', 'tunjang', 'besar', 'keluh', 'gaji', 'potong', 'tapera']
Filtering	['suruh', 'beli', 'rumah', 'bangsat', 'tunjang', 'besar', 'keluh', 'gaji', 'potong', 'tapera']	"suruh beli rumah bangsat tunjang besar keluh gaji potong tapera"

4.3. Pelebelan Data

Hasil pelabelan sentimen menggunakan model bahasa IndoBERT mengklasifikasikan tweet dalam 3 klasifikasi, yaitu sentimen positif, netral, dan negatif. Distribusi jumlah tweet pada setiap kategori sentimen ditampilkan pada Gambar 2 sebagai gambaran awal kecenderungan opini publik terhadap kebijakan Tapera di media sosial X.



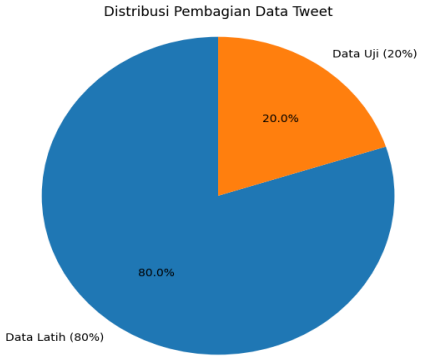
Gambar 2. Distribusi kategori sentimen

Pada Gambar 2 dapat dilihat bahwa hasil pelabelan sentimen terbagi “3 klasifikasi. Distribusi data memperlihatkan sentimen netral dan negatif secara nyata mendominasi dari positif. Secara rinci, jumlah tweet bersentimen netral sebanyak 1.626, sentimen negatif (1.615), dan positif (1.153) tweet.”

4.4. Pembagian Data

Hasil pembagian dataset sentimen selanjutnya divisualisasikan untuk menunjukkan proporsi data latih dan data uji yang digunakan dalam penelitian ini.

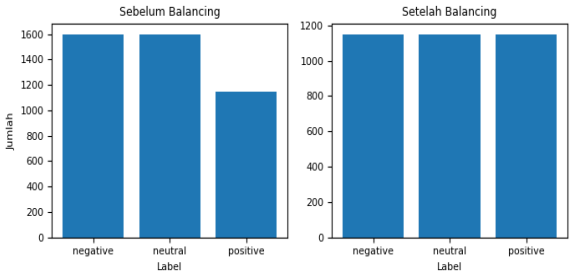
Visualisasi “pada Gambar 3 menyajikan distribusi dataset yang dipecah ke dalam dua fragmen utama dengan rasio 80% sebagai set pelatihan dan 20% sebagai set pengujian. Dominasi proporsi pada data latih bertujuan untuk memberikan ruang bagi sistem dalam menginternalisasi karakteristik polaritas teks secara mendalam. Sementara itu, alokasi data uji digunakan sebagai instrumen validasi guna mengukur sejauh mana model mampu menggeneralisasi informasi yang belum pernah ditemui selama fase pembelajaran.



Gambar 3. Visualisasi pembagian data

4.5. Penyeimbangan Data

Dilanjutkan, tahap penyeimbangan data untuk mengatasi ketidakseimbangan jumlah tweet pada setiap kelas sentimen.



Gambar 4. Visualisasi penyeimbangan data

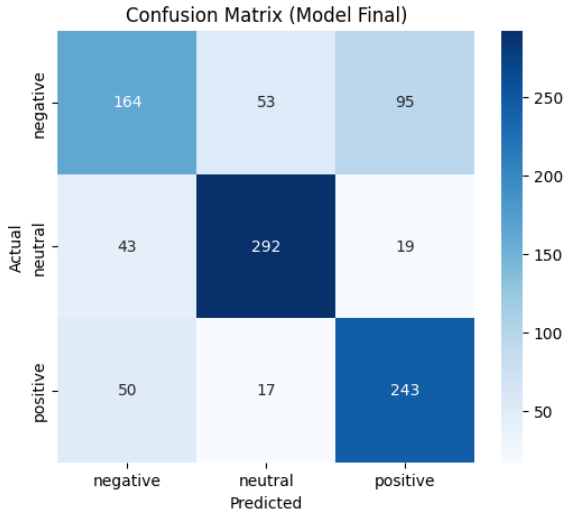
Dari Gambar 4 terlihat bahwa sebelum penyeimbangan, distribusi data sentimen belum merata dengan dominasi sentimen netral dan negatif. Untuk menghindari bias model terhadap kelas tertentu, dilakukan penyeimbangan data menggunakan teknik oversampling sehingga jumlah data pada setiap kelas menjadi lebih seimbang dan mendukung kinerja klasifikasi yang lebih stabil.

4.6. Pembobotan kata

Data text yang melalui proses penyeimbangan selanjutnya diubah ke dalam bentuk numerik memanfaatkan TF-IDF. Sistem ini bekerja dengan mengevaluasi relevansi tiap kata melalui penghitungan seberapa sering istilah tersebut muncul, yang kemudian dikonversi menjadi nilai bobot statistik dalam korpus data dan keseluruhan dataset, sehingga fitur yang dihasilkan dapat merepresentasikan informasi sentimen secara lebih efektif dalam proses klasifikasi.

4.7. Pemodelan dan Evaluasi

Pada tahap pemodelan, data yang telah melalui proses pembobotan TF-IDF digunakan sebagai masukan untuk algoritma *Logistic Regression.* Model ini diterapkan pada data latih dan selanjutnya dievaluasi menggunakan data uji guna menilai kemampuan klasifikasi sentimen terhadap kebijakan Tapera secara objektif.



Gambar 5. Visualisasi confusion matrix

Dari confusion matrix pada Gambar 5 dapat dilihat bahwa model mampu mengklasifikasikan sebagian besar data uji dengan benar pada masing-masing kelas sentimen. Prediksi yang tepat ditunjukkan oleh nilai pada diagonal utama, sementara kesalahan klasifikasi masih ditemukan, terutama antara kelas sentimen negatif dan positif. Kondisi ini menunjukkan adanya kemiripan pola kata pada kedua kelas tersebut yang memengaruhi hasil prediksi model.

Pengujian terhadap kapabilitas model selanjutnya berpijak pada nilai akurasi, presisi, recall, dan f1-score. Penggunaan ragam metrik evaluasi ini merupakan langkah krusial untuk mengekstraksi representasi kinerja model yang lebih objektif dan menyeluruh dalam memproses data.

```
==Logistic Regression Result==
Accuracy: 0.7161885245901639
              precision    recall  f1-score   support

    negative       0.64      0.53      0.58       312
     neutral       0.81      0.82      0.82       354
    positive       0.68      0.78      0.73       310

    accuracy                           0.72       976
   macro avg       0.71      0.71      0.71       976
weighted avg       0.71      0.72      0.71       976
```

Gambar 6. Hasil akurasi model

Berdasarkan Gambar 6, model Logistic Regression menghasilkan nilai akurasi sebesar 71,6%, yang menunjukkan bahwa sebagian besar data uji dapat diklasifikasikan dengan tepat. Nilai presisi, recall, dan f1-score per kelas sentimen menunjukkan kinerja yang cukup seimbang setelah proses penyeimbangan data dilakukan. Output ini mengindikasikan model yang dibangun memadai untuk digunakan pada analisis sentimen kebijakan Tapera pada media sosial X.

5. KESIMPULAN DAN SARAN

Studi ini berfokus pada eksplorasi persepsi serta opini publik mengenai regulasi Tabungan Perumahan Rakyat (Tapera) dengan mengekstraksi data percakapan yang berkembang di platform media sosial X menggunakan pendekatan *machine learning*. Proses penelitian meliputi pengumpulan data, pra-pemrosesan teks, pelabelan sentimen menggunakan model IndoBERT, pembobotan kata dengan TF-IDF, serta klasifikasi menggunakan algoritma *Logistic Regression*. Hasil analisis menunjukkan bahwa sentimen publik didominasi oleh sentimen netral dan negatif, yang mencerminkan respons masyarakat yang cenderung kritis terhadap kebijakan Tapera. Evaluasi model menghasilkan tingkat akurasi sebesar 71,6% dengan "nilai presisi, recall, dan f1-score" yang relatif seimbang pada setiap kelas sentimen setelah dilakukan penyeimbangan data, sehingga model dinilai mampu mengklasifikasikan sentimen dengan cukup baik. Untuk penelitian selanjutnya, disarankan memperluas jumlah dan variasi data, membandingkan beberapa algoritma klasifikasi, serta mengembangkan analisis sentimen berbasis aspek guna memperoleh hasil yang lebih mendalam dan komprehensif.

DAFTAR PUSTAKA

[1] T. Ramadhanu Hidayat and D. Prastyo, "ANALISIS SENTIMEN KOMENTAR PADA KONTEN KUALIFIKASI PIALA DUNIA TIMNAS INDONESIA DI MEDIA SOSIAL TIKTOK MENGGUNAKAN ALGORITMA DECISION TREE DAN METODE N-GRAM," vol. 9, no. 6, pp. 9657–9663, 2025.

[2] D. Setiyawati and N. Cahyono, "Analisa Sentimen Pengguna Sosial Media Twitter Terhadap Perokok di Indonesia," vol. 12, no. 1, pp. 262–272, 2023.

[3] A. Y. Ma'rufudin, "Analisis Sentimen Petani Milenial Pada Media Sosial X Menggunakan Algortitma Support Vector Machine (SVM) Fakultas Teknik dan Ilmu Komputer , Universitas Teknokrat Indonesia , Indonesia Sentiment Analysis of Millennial Farmers on Social Media X Using th," vol. 5, no. 3, pp. 845–857, 2025.

[4] H. Leidiyana, T. Misriati, and R. Aryanti, "Klasifikasi Sentimen Terhadap Kebijakan Tapera Menggunakan Komparasi Machine Learning dan SMOTE," *J. Komtika (Komputasi dan Inform.*, vol. 8, no. 2, pp. 125–135, 2024, doi: 10.31603/komtika.v8i2.12595.

[5] A. Steven, D. Prasetya, and A. S. Aji, "Public Sentiment Analysis Towards Mayor Teddy Indra Wijaya with Logistic Regression Approach Analisis Sentimen Publik Terhadap Mayor

Teddy Indra Wijaya dengan Pendekatan Logistic Regression," vol. 5, no. January, pp. 222–231, 2025.

[6] N. Hadi and D. Sugiarto, "Analisis Sentimen Pembangunan IKN pada Media Sosial X Menggunakan Algoritma SVM , Logistic Regression dan Naïve Bayes," vol. 10, no. 1, pp. 37–49, 2025, doi: 10.30591/jpit.v10i1.7106.

[7] ASH SHIDDICKY, "ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN VAKSINASI COVID-19 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE LOGISTIC REGRESSION," 2022.

[8] T. Hidayat, T. Informatika, U. I. Syekh-yusuf, K. Tangerang, and I. Gelap, "ANALISIS SENTIMEN TWEET MASYARAKAT PADA ISU INDONESIA GELAP DI MEDIA SOSIAL X MENGGUNAKAN ALGORITMA," vol. 9, no. 6, pp. 9664–9670, 2025.

[9] I. Muhandhis and A. S. Ritonga, "Public Sentiment Analysis on TikTok about Tapera Policy using Random Forest Classifier," *Sistemasi*, vol. 14, no. 1, p. 354, 2025, doi: 10.32520/stmsi.v14i1.4878.

[10] D. Syamdova *et al.*, "Prediksi interaksi pengguna media sosial x dengan algoritma machine learning," vol. 9, no. 2, pp. 211–222, 2025.

[11] Y. Wang, J. Guo, C. Yuan, and B. Li, "applied sciences Sentiment Analysis of Twitter Data," pp. 1–14, 2022.

[12] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decis. Anal. J.*, vol. 11, no. April, p. 100473, 2024, doi: 10.1016/j.dajour.2024.100473.

[13] A. J. Saputra, S. Achmadi, and K. Auliasari, "Penggunaan Metode Logistic Regression Untuk Analisis Sentimen Pembangunan Kota Nusantara Pada Media Sosial," 2025.

[14] R. A. Husen, R. Astuti, L. Marlia, and L. Efrizoni, "Sentiment Analysis of Public Opinion on Twitter Toward BSI Bank Using Machine Learning Algorithms Analisis Sentimen Opini Publik pada Twitter Terhadap Bank BSI Menggunakan Algoritma Machine Learning," vol. 3, no. October, pp. 211–218, 2023.

[15] A. N. Sativa, A. Rizky, I. Putri, and J. A. Putri, "Analisis Sentimen Twitter Ibu Kota Negara Nusantara Menggunakan Algoritma Naive Bayes , Logistic Regression dan K-Nearest Neighbors," vol. 3, no. 2, pp. 34–40, 2024.

[16] R. S. Anjani, S. H. Nabilah, and H. M. Winata, "Analisis sentimen publik terhadap kebijakan subsidi dan kenaikan harga bbm di indonesia melalui media sosial x (twitter)," vol. 11, no. 8, 2025.