JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 23~28
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Eksplorasi Model Prediksi Sentimen Postingan Di Media Sosial

[1]Fitri Purwaningtias
[1]Universitas Bina Darma, Indonesia
[1]fitri.purwaningtias@binadarma.ac.id

Article Info

Article history:

Received, 2024-12-15
Revised, 2024-12-25
Accepted, 2024-12-26

Kata Kunci:

Analisis Sentimen
Media Sosial
SMOTE

ABSTRAK

Analisis sentiment menjadi teknik analisis teks yang bisa digunakan untuk memahami opini, perasaan atau sentiment dari suatu teks. Penelitian ini bertujuan mengeksplorasi dan membandingkan model prediksi sentiment pada data media sosial dengan tiga algoritma yaitu *GaussianNB*, *Logistic Regression* dan *Support Vector Machine* (SVM). Dataset yang digunakan diambil dari www.kaggle.com yang terdiri dari postingan media sosial dari *platform Twitter*, *Facebook* dan *Instagram* dengan kategori senteimen positif, negative dan netral. Proses analisis dengan *preprocessing* data teks, pelabelan data, ekstraksi fitur dengan *Bag of Words* (BoW) dan TF-IDF serta penanganan ketidakseimbangan data dengan SMOTE. Hasil penelitian menunjukkan model SVM dengan TF-IDF dan SMOTE performa terbaik dengan akurasi 93,25% pada data latih dan 92,50% data uji. Penelitian ini memberikan kontribusi dalam menentukan model terbaik untuk analisis sentimen data media sosial dan dapat menjadi acuan dalam pengembangan sistem prediksi sentiment lebih baik di masa depan.

Keywords:

Sentiment Analysis
Social Media
SMOTE

ABSTRACT

Sentiment analysis is a text analysis technique that can be used to understand the opinion, feeling, or sentiment of a text. This research aims to explore and compare sentiment prediction models on social media data with three algorithms, namely GaussianNB, Logistic Regression, and Support Vector Machine (SVM). The dataset used is taken from www.kaggle.com, which consists of social media posts from the Twitter, Facebook, and Instagram platforms with positive, negative, and neutral sentiment categories. The analysis process involves text data preprocessing, data labeling, feature extraction with Bag of Words (BoW) and TF-IDF, and handling data imbalance with SMOTE. The results showed that the SVM model with TF-IDF and SMOTE performed best, with 93.25% accuracy on training data and 92.50% on test data. This research contributes to determining the best model for sentiment analysis of social media data and can be a reference in developing better sentiment prediction systems in the future.

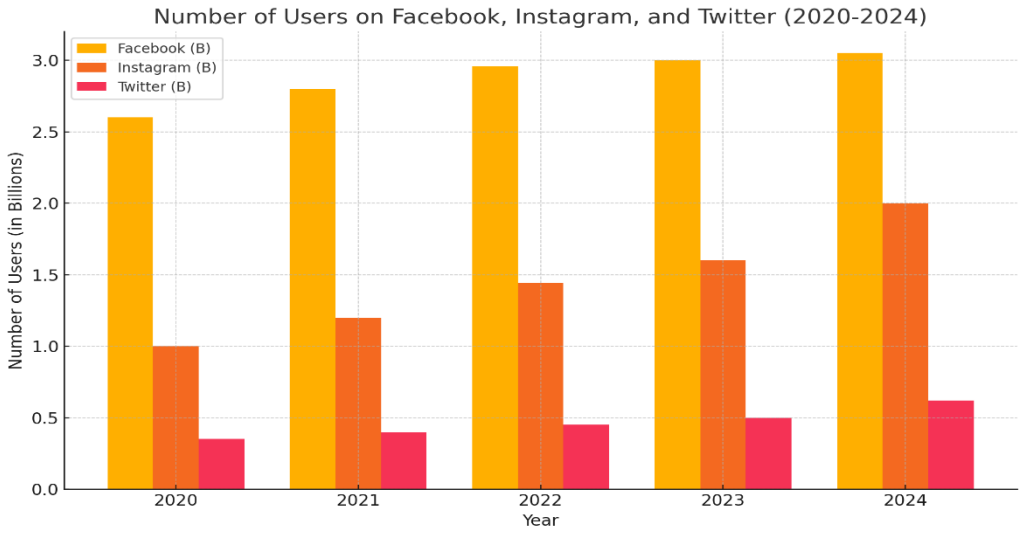
Penulis Korespondensi:

Fitri Purwaningtias,
Program Studi Sistem Informasi,
Universitas Bina Darma,
Email: fitri.purwaningtias@binadarma.ac.id

1. PENDAHULUAN

Pada saat ini media sosial menjadi bagian yang selalu berhubungan dengan kehidupan masyrakat. Platform media sosial tidak hanya menjadi tempat berbagi cerita, pengalaman dan opini tetapi juga mencerminkan emosi dan perasaan pengguna [1]. Melalui postingan sehari-hari pengguna media sosial sering mengungkapkan sentimen [2] yang beragam baik itu positif, negatif ataupun netral terhadap suatu peristiwa, produk atau pengalaman pribadi. Paltform media sosial yang marak digunakan masyarakat yaitu *Twitter*, *Facebook* dan *Instagram* dimana setiap bulannya berdasarkan laporan *We Are Social* [3] dan *Goodstats* [4] mengalami

kenaikan pengguna dari tahun 2020 sampai tahun 2024 . Berikut grafik jumlah pengguna dari *Twitter*, *Facebook* dan *Instagram* seperti Gambar 1:



Gambar 1 Grafik Jumlah Pengguna Aplikasi Media Sosial *Twitter*, *Facebook* dan *Instagram*

Gambar 1 menunjukkan jumlah pengguna aktif bulanan untuk ketiga platform media sosial dimana Twitter meskipun memiliki pengguna lebih sedikit tetapi pertumbuhannya stabil sebagai platform opini dan berita. Kemudian *Facebook* tetap *platform* dengan jumlah pengguna tertinggi tetapi pertumbuhannya lebih lambat di bandingkan *Instagram* karena Instagram menunjukkan pengguna cukup besar dibanding kedua *platform* sebelumnya.

Analisis sentiment menjadi teknik analisis teks yang bisa digunakan untuk memahami opini, perasaan atau sentiment dari suatu teks [5]. Dalam penelitian ini menunjukkan untuk ekspresi, opini dan perasaan pengguna dalam postingan mereka karena dengan banyaknya postingan menjadi sulit untuk menentukan sentiment positif, negatif ataupun netral secara manual [6]. Oleh karena itu penelitian ini bertujuan untuk menganalisis sentimen teks terhadap postingan pengguna di media sosial (*Twitter*, *Facebook* dan *Instagram*) dimana dataset yang diperoleh dari www.kaggle.com. Data yang peroleh dari tahun 2017-2023 sebanyak 732 data. Dalam penelitian ini pendekatan yang digunakan dengan *Natural Language Processing* yaitu bidang ilmu yang memanfaatkan kecerdasan buatan untuk dapat memberikan dukungan keputusan dalam menilai sentiment dengan mengkategorikan apakah snetimen itu positif, negatif atau netral [7] kemudian performa metode dievaluasi melalui analisis dan pengujian akurasi menggunakan algoritma *Gaussian Naïve Bayes, Support Vector Machine* dan *Random Forest*.

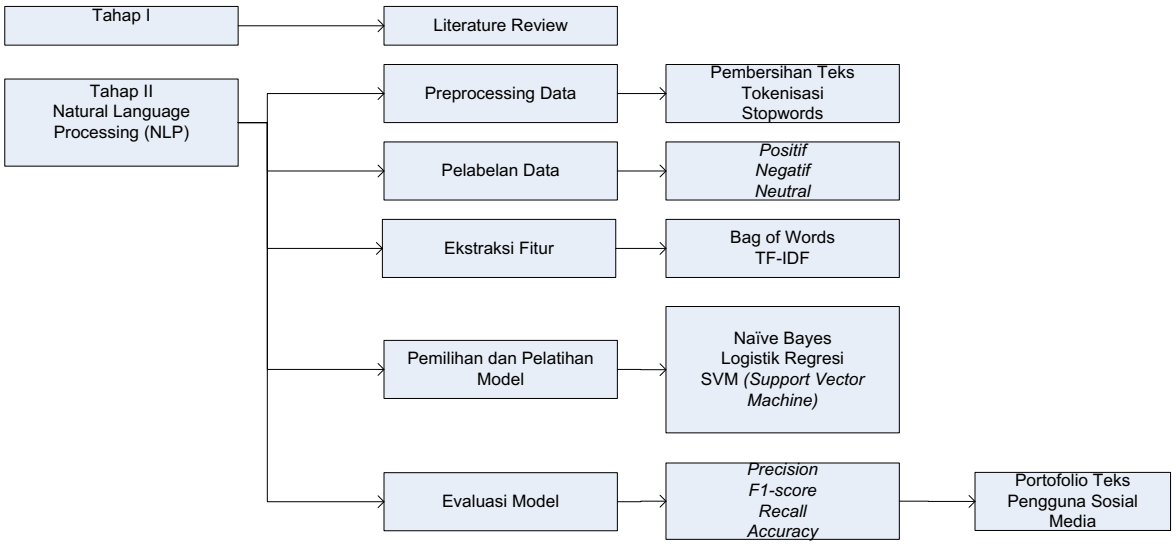
Beberapa penelitian sebelumnya yaitu mengenai analisis sentimen untuk ulasan platform media sosial menggunakan algoritma *Naïve Bayes* bahwa sebagian besar pengguna memberikan sentiment positif pada aplikasi *Twitte*, *Instagram* dan Tiktok [8]. Kemudian penelitian mengenai analisis sentiment wacana pemindahan ibu kota Indonesia menggunakan algoritma SVM (*Support Vector Machine*) menghasilkan bahwa proses pengujian yang dilakukan dari tweets sentiment sebanyak 1.236 tweets (404 positif dan 832 negatif) dengan akurasi SVM= 96,68% [9]. Selanjutnya untuk analisis sentiment pada ulasan di *Shopee* di *Google Play* Store menggunakan *Naïve Bayes* disimpulkan terdapat beragam sentiment baik positif, negatif dan netral [10]. Penelitian analisis sentiment masyarakat terhadap tindakan vaksinasi dalam upaya mengatasi pandemic covid-19 dengan metode *naïve bayes* dan SVM memperoleh hasil klasifikasi metode *naïve bayes* rata-rata akurasi 85,59 % dan SVM sebesar 84,41% [11].

Meskipun telah banyak penelitian yang dilakukan dalam analisis sentiment namun masih sedikit penelitian yang menganalisis pengguna terhadap platform media sosial *Twitter*, *Facebook* dan *Instagram* selain itu pada penelitian sebelumnya hanya menggunakan satu algoritma ataupun dua algoritma untuk perbandingan sedangkan penelitian ini menggunakan tiga algoritma berupa *Gaussian Naïve Bayes*, SVM *(Support Vector Machine*) dan *Logistic Regression*. Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis sentiment pada ketiga platform media sosial dan membandingkan model berdasarkan metric evaluasi akurasi untuk menentukan model terbaik. Dan manfaat yang diberikan dari peenlitian ini memberikan rekomendasi model terbaik untuk prediksi sentiment media sosial.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 23~28
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

2. METODE PENELITIAN

Alur proses menggambarkan tahapan dalam analisis sentiment yang dilakukan pada penelitian ini menggunakan *Natural Language Processing* (NLP) seperti gambar 2 dibawah ini:



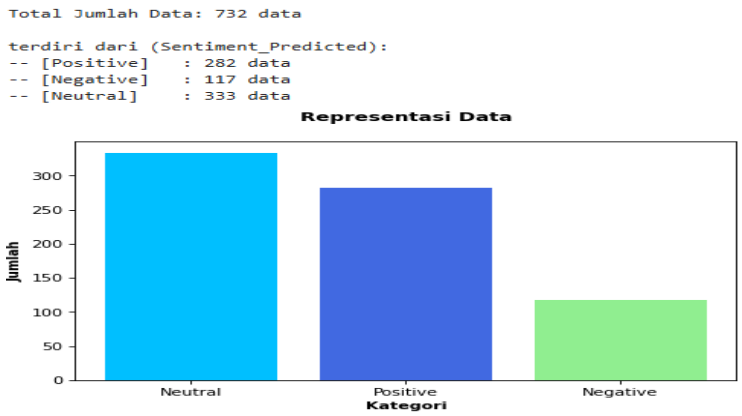
Gambar 2 Alur Penelitian

Dari gambar 2 menjelaskan tahapan dalam penelitian ini yang dimulai dengan literature review mengenai analisis sentiment, sosial media dan juga NLP. Kemudian dilakukan pengumpulan data melalui www.kaggle.com mengenai postingan teks pengguna dalam sosial media (*Twitter*, *Facebook* dan *Instagram*). Setelah data diperoleh maka tahap kedua proses NLP dengan melakukan *preprocessing* data yaitu membersihkan teks dari karakter khusus, mengkonversi teks menjadi huruf kecil dan lain-lain [12]. *Preprocessing* yang dilakukan dalam penelitian ini berupa pembersihan teks, menghilangkan kata-kata umum (*stopwords*), *tokenization* yaitu memecah teks menjadi unit-unit lebih kecil seperti kata atau frasa [13]. Selanjutnya tahapan kedua dalam NLP dengan melakukan pelabelan data yaitu memberikan sentiment dari postingan teks pengguna berupa *positive*, *negative* dan *neutral*. Setelah pelabelan data maka dilakukan ekstraksi fitur menggunakan *bag of words* (BoW) dan TF-IDF. Kemudian pemilihan model dan pelatihan model (*modelling*), melatih model atau menggunakan model yang sudah dilatih untuk tugas-tugas seperti klasifikasi, generasi teks atau pemrosesan bahasa alami lainnya [14]. Dimana model algortima yang digunakan yaitu *Gaussian Naïve Bayes, Logistic Regression* dan SVM (*Support Vector Machine*). Dan terakhir yaitu evaluasi model, mengevaluasi kinerja model menggunakan metrik-metrik seperti *accurate, precision, recall* [15]. Pada penelitian ini evaluasi model menggunakan akurasi, *F1-score* dan precision dari tiga model yang telah diuji dan di train pada dataset dengan tanpa menggunakan SMOTE dan dengan menggunakan SMOTE.

3. HASIL DAN ANALISIS

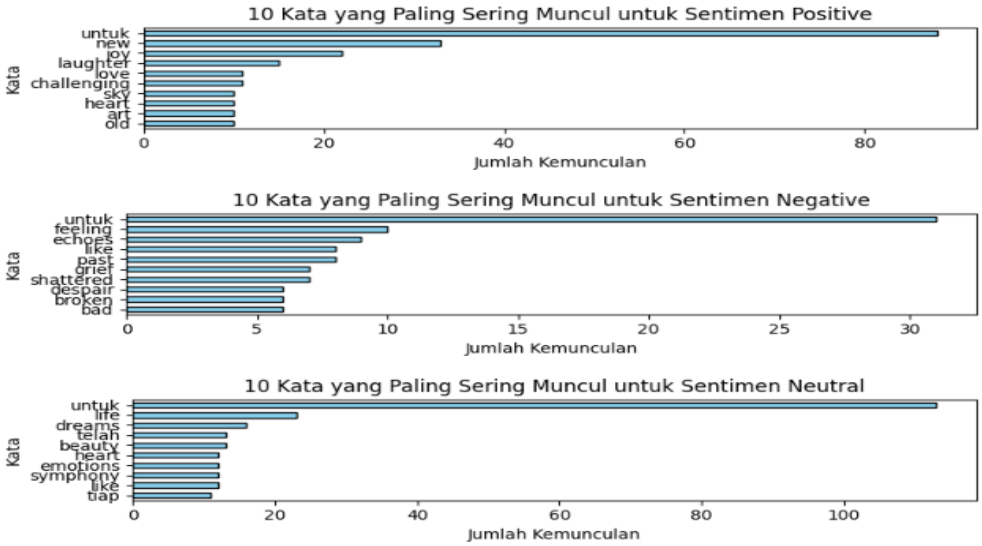
Penelitian ini menggunakan dataset dalam analisis klasifikasi teks berisi informasi berisi *Text*, *Platform* dan *Sentiment_Predicted*. Pada kolom Text itu berisi postingan pengguna selanjutnya kolom *Platform* sosial media (Twitter, Facebook dan Instagram) Kemudian untuk kolom Sentiment_*Predicted* berisi sentiment yang digunakan sebagai label target yang menunjukkan hasil klasifikasi bernilai 0 (*Positive*), 1 *(Negative*) dan 2 (Neutral). Dataset dilakukan fitur ekstraksi menggunakan *Count Vectorizer* yaitu BoW (*Bag of Word*) dan TF-IDF. Dataset yang digunakan berjumlah 732 data dimana untuk "*Positive*" berjumlah 282 data, kemudian "*Negative*" berjumlah 117 data dan "*Neutral*" ada 333. Representasi dataset yang digunakan seperti gambar 3 dibawah ini:

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 23~28
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085



Gambar 3 Representasi Data

Selanjutnya dataset diatas telah dilakukan *preprocessing* data berupa menjadikan teks jadi huruf kecil semua, penghapusan karakter khusus dan kata-kata umum (*stopwords*) dan tokenisasi. Kemudian diberikan pelabelan sentimen *positive*, *negative* dan *neutral*. Setelah diberi pelabelan sentimen dilakukan ekstraksi fitur dengan menggunakan *Count Vectorizer* berupa *Bag of Word (BoW)* dan TF-IDF. Untuk hasil dari 10 kata paling sering muncul dari ekstraksi BoW seperti gambar 4 dibawah ini:



Gambar 4 Frekuensi Kemunculan Data

Selanjutnya pemilihan model dengan perbandingan algoritma *Gaussian Naïve Bayes, Logistic Regression* dan *Support Vector Machine (*SVM) dengan data yang digunakan sebagai data latih 80% dan 20% data uji seperti gambar 5:

```
Total data dalam dataset: 732
Jumlah data latih: 585
Jumlah data uji: 147
```

Gambar 5 Total Data Latih dan Data Uji

Pada dataset yang ada memiliki ketidakseimbangan kelas terutama pada kelas *Negative* yang lebih sedikit dibanding kelas lainnya. Sehingga penelitian menggunakan SMOTE untuk ketidakseimbangan kelas dalam dataset untuk meningkatkan performa klasifikasi data dimana satu kelas memiliki contoh jauh lebih sedikit dibandingkan kelas lainnya [16]. Gambar 6 mengenai data train dan data uji sebelum SMOTE dan sesudah SMOTE:

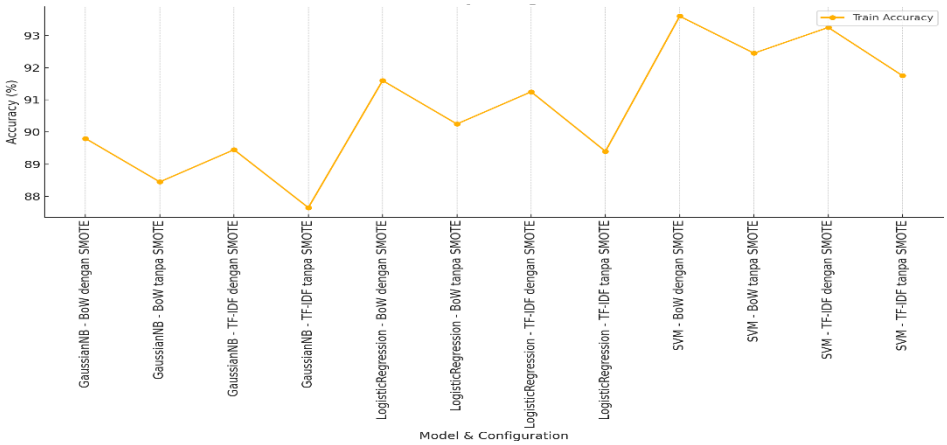
```
Jumlah data latih dan data uji sebelum SMOTE:
Data latih: 585
Data uji: 147

Jumlah data pada data pelatihan sebelum SMOTE:
Sentiment_Predicted
Neutral     274
Positive    222
Negative     89
Name: count, dtype: int64

Jumlah data setelah SMOTE:
Sentiment_Predicted
Neutral     274
Positive    274
Negative    274
```

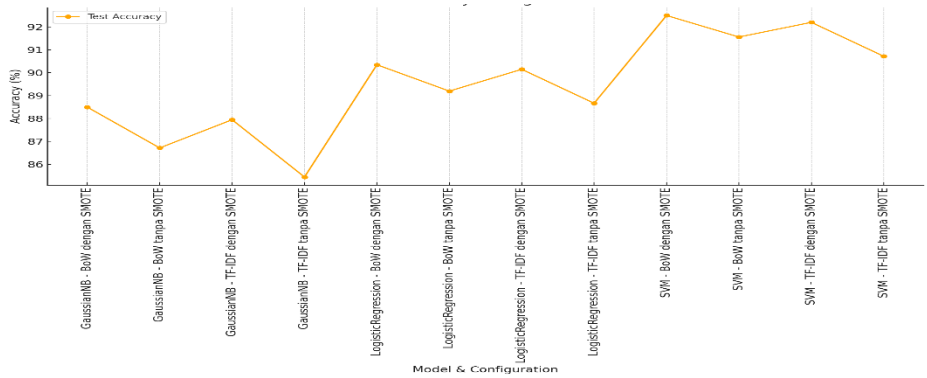
Gambar 6 Data Latih dan Data Uji Sebelum SMOTE dan Setelah SMOTE

Dari penelitian ini menghasilkan perbandingan dari data uji dan data train model algoritma *Gaussian Naïve Bayes, Logistic Regression*, SVM dengan BoW dan TF-IDF tanpa menggunakan SMOTE dan dengan menggunakan SMOTE seperti gambar 7 dan gambar 8 di bawah ini:



Gambar 7 Data Train

Dari gambar 7 dan gambar 8 memperlihatkan bahwa model SVM dengan TF-IDF dan SMOTE memiliki akurasi tertinggi sekitar 93% pada data train dan 92% pada data uji. Sedangkan *GaussianNB* akurasinya lebih rendah sehingga pada data train ini SVM dengan SMOTE memberikan performa terbaik pada data latih dengan TF-IDF yang lebih unggul dari BoW.



Gambar 8 Data Test

4. KESIMPULAN

Penelitian ini bertujuan mengeksplorasi dan membandingkan model prediksi sentiment berdasarkan akurasi dengan studi kasus media sosial. Dengan penggunaan tiga model yang dievaluasi *GaussianNB*, *Logistic Regression* dan SVM (*Support Vector Machine*) dengan konfigurasi *Bag of Words*, TF-IDF serta penggunaan SMOTE. Hasil analisis menunjukkan bahwa SVM dengan TF-IDF dan SMOTE kombiansi terbaik memprediksi sentiment pada data media sosial dengan akurasi tertinggi 93, 25% pada data train dan 92,50% pada data uji. Selain itu, model *Logistic Regression* menunjukkan performa cukup baik sedangkan GaussianNB memiliki akurasi lebih rendah yang menunjukkan bahwa *GaussianNB model* sederhana kurang cocok untuk

datset kompleks. Dengan hasil ini bisa memberikan rekomendasi bagi pengembangan model prediksi sentiment pada *platform* media sosial lainnya.

REFERENSI

[1] K. Adib, M. R. Handayani, W. D. Yuniarti, and K. Umam, "Opini Publik Pasca-Pemilihan Presiden: Eksplorasi Analisis Sentimen Media Sosial X Menggunakan SVM," *SINTECH (Science Inf. Technol. J.*, vol. 7, no. 2, pp. 80–91, 2024, doi: 10.31598/sintechjournal.v7i2.1581.

[2] G. R. Putri, M. A. Maulana, and S. Bahri, "Perbandingan Algoritma Naïve Bayes dan TextBlob Untuk Mendapatkan Analisis Sentimen Masyarakat Pada Sosial Media," *Teknika*, vol. 13, no. 2, pp. 213–218, 2024, doi: 10.34148/teknika.v13i2.815.

[3] We Are Social, "10 Media Sosial dengan Pengguna Terbanyak 2024," 2024. .

[4] Goodstats, "Statistik Pengguna Twitter 2020-2024," 2024. https://data.goodstats.id/statistic/10-media-sosial-dengan-pengguna-terbanyak-2024-CaJT1.

[5] A. Sentimen *et al.*, "ANALISIS SENTIMEN MENGGUNAKAN PENDEKATAN MANUSIA-KOMPUTER (IMK) DENGAN TEKNIK DATA MINING PADA MEDIA SOSIAL (STUDI KASUS : FAKULTAS TEKNIK UNIVERSITAS JABAL GHAFUR)," pp. 96–104.

[6] R. Azhar, A. Surahman, and C. Juliane, "Analisis Sentimen Terhadap Cryptocurrency Berbasis Python TextBlob Menggunakan Algoritma Naïve Bayes," *J-SAKTI (Jurnal Sains ...*, 2022, [Online]. Available: http://tunasbangsa.ac.id/ejurnal/index.php/jsakti/article/view/443.

[7] F. Rumaisa, Y. Puspitarani, A. Rosita, A. Zakiah, and S. Violina, "Penerapan Natural Language Processing (NLP) di bidang pendidikan," *J. Inov. Masy.*, vol. 1, no. 3, pp. 232–235, 2021, doi: 10.33197/jim.vol1.iss3.2021.799.

[8] S. Saepudin, S. Widiastuti, and C. Irawan, "Sentiment Analysis of Social Media Platform Reviews Using the Naïve Bayes Classifier Algorithm," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 2, pp. 236–243, 2023, doi: 10.32736/sisfokom.v12i2.1650.

[9] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[10] M. R. Firdaus, N. Rahaningsih, and R. D. Dana, "Analisis Sentimen Aplikasi Shopee di Goole Play Store Menggunakan Klasifikasi Algoritma Naïve Bayes," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 228–237, 2024.

[11] B. Laurensz and Eko Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 2, pp. 118–123, 2021, doi: 10.22146/jnteti.v10i2.1421.

[12] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[13] U. V. Ucak, I. Ashyrmamatov, and J. Lee, "Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization," *J. Cheminform.*, vol. 15, no. 1, pp. 1–13, 2023, doi: 10.1186/s13321-023-00725-9.

[14] I. Ul Haq, M. Pifarré, and E. Fraca, "Novelty Evaluation using Sentence Embedding Models in Open-ended Cocreative Problem-solving," *Int. J. Artif. Intell. Educ.*, no. 0123456789, 2024, doi: 10.1007/s40593-024-00392-3.

[15] H. Schuff, L. Vanderlyn, H. Adel, and N. T. Vu, "How to do human evaluation: A brief introduction to user studies in NLP," *Nat. Lang. Eng.*, vol. 29, no. 5, pp. 1199–1222, 2023, doi: 10.1017/S1351324922000535.

[16] N. MUSTAFA, J.-P. LI, R. A., and M. Z., "A Classification Model for Imbalanced Medical Data based on PCA and Farther Distance based Synthetic Minority Oversampling Technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 1, pp. 61–67, 2017, doi: 10.14569/ijacsa.2017.080109.