REPOSITOR, Vol. 7, No. 4, November 2025, Pp. 513-520
ISSN : 2714-7975
E-ISSN : 2716-1382

Analisis Polaritas Terhadap Komentar Pada Platform Youtube Menggunakan Algoritma Naive Bayes dan XGBoost

Nazilullaily Nur Aisyah*[1], Setio Basuki[2]
[1,2]Universitas Muhammadiyah Malang
nazilullailynur@gmail.com*[1], setio_basuki@umm.ac.id[2]

Abstrak

Objek penelitian: penelitian ini bertujuan untuk menganalisis respons netizen dalam menanggapi permasalah politik di Indonesia menggunakan nada dan bahasa yang terdapat pada komentar pada konten youtube Pandji Pragiwaksono menggunakan kategori yang lebih detail terkait dengan yang digunakan pada komentar. Masalah Penelitian: kajian literatur mengenai topik ini menunjukkan penelitian sebelumnya memiliki fokus yang terbatas. Sehingga meninggalkan permasalah yang terabaikan dan perlu untuk dilakukan analisis lebih lanjut. Metode Penelitian: penelitian ini menerapkan metode klasifikasi pada beberapa polaritas sentimen politik yang terkait dengan nada dan bahasa. Terdapat 3 polaritas sentimen yang dijadikan label utama: Kontruktif (menyajikan kritik dan saran), Emosional (menyajikan ekspresi perasaan seseorang seperti amarah, kecewa, sedih, dll), Humoris (menyajikan pendekatan humor atau sarkasme), dan 1 label tambahan yaitu Others. Tahap kedua, dataset didapatkan menggunakan teknik crawling data komentar pada platform youtube Indonesia, dengan 3 kategori komentar dengan jumlah banyak (4000), komentar sedang (1000), dan komentar sedikit (150). Tahap ketiga, penelitian ini menggunakan pelabelan secara manual dan kemudian dilanjutkan dengan tahap preprocessing agar menyiapkan data untuk dapat diproses lebih lanjut. Tahap keempat, melakukan proses augmentasi pada data train untuk dapat meningkatkan performaa model. Tahap kelima, klasifikasi yang didukung oleh fitur teks klasik seperti Bow, TF-IDF, 2-Gram, 3-Gram, dan word embeddings contextual menggunakan IndoBert. Hasil penelitian: menunjukkan bahwa data yang sudah melalui proses augmentasi memiliki performa yang unggul dibanding dengan data original dengan akurasi mencapai 97% menggunakan fitur word embeddings. Secara keseluruhan penggunakan teknik augmentasi dapat meningkatkan performa pada setiap model yang digunakan.

Kata Kunci: *Analisis Polaritas, Fitur Klasik, Naive Bayes, Word Embeddings, XGBoost.*

Abstract

Research Objective: This study aims to analyze netizen responses in response to political problems in Indonesia using the tone and language contained in comments on Pandji Pragiwaksono's youtube content using more detailed categories related to those used in comments. Research Problem: a review of the literature on this topic shows that previous research has a limited focus. So that it leaves problems that are neglected and need to be analyzed further. Research Methods: this research applies classification methods to several political sentiment polarities related to tone and language. There are 3 sentiment polarities that are used as the main label: Constructive (presenting criticism and suggestions), Emotional (presenting an expression of one's feelings such as anger, disappointment, sadness, etc.), Humorous (presenting a humor or sarcasm approach), and 1 additional label namely Others. In the second stage, the dataset was obtained using the technique of crawling comment data on the Indonesian youtube platform, with 3 categories of comments with a large number (4000), medium comments (1000), and few comments (150). In the third stage, this research uses manual labeling and then continues with the preprocessing stage in order to prepare the data for further processing. In the fourth stage, the augmentation process is performed on the train data to improve the performance of the model. The fifth stage, classification supported by classic text features such as Bow, TF-IDF, 2-Gram, 3-Gram, and contextual word embeddings using IndoBert. The results showed that the data that has gone through the augmentation process has superior performance compared to the original data with an accuracy of 97% using the word

embeddings feature. Overall, the use of augmentation techniques can improve the performance of each model used.

Keywords *Classic Features, Naive Bayes, Polarity Analysis, Word Embeddings, XGBoost.*

1. Pendahuluan

Media sosial berkembang pesat sejalan dengan pertumbuhan dan kemudahan akses informasi yang didukung oleh kekuatan teknologi komunikasi [1]. Indonesia merupakan salah satu negara teraktif di media sosial, yang memiliki pengguna aktif sebanyak 79 juta pengguna [1]. Media sosial merupakan saarana yang dapat digunakan oleh masyarakat untuk ikut berpartisipasi dalam proses politik [2]. Youtube adalah salah satu media sosial yang dapat digunakan untuk menyalurkan pendapat mengenai politik, sehingga kita dapat berkontribusi dan menyalurkan pendapat terhadap isu-isu nasional [3]. Pandji Pragiwaksono adalah salah satu *public figure* yang sangat kritis oleh politik yang ada di Indonesia sehingga beliau memanfaatkan media sosial youtube untuk menyalurkan pendapatnya [4]. Selain itu komentar di youtube juga dapat mencerminkan opini publik dan bahkan dapat mempengaruhi pembentukan opini politik pengguna lainnya atau dapat disebut sebagai polaritas.

Polaritas mendapat atau polaritas teks dapat merujuk pada kecenderungan atau terdapat beberapa sentimen yang umumnya dikategorikan sebagai sentimen positif, negatif, dan netral. Analisis polaritas atau sentimen bertujuan untuk mengkasifikasikan opini atau perasaan yang diekspresikan dalam bentuk teks. Akan tetapi penelitian ini mencoba melakukan klasifikasi dengan menggunakan kategori yang lebih detail sepert,kategori kontruktif yang menyajikan kritik dan saran serta argumentasi yang logis dan Solusi konkret [5]. Kemudian kategori emosional mencakup komentar yang menggambarkan perasaan marah, kecewa, sedih, kegembiraan yang subjektif [6]. Kemudian kategori humoris meliputi komentar yang menggunakan pendekatan humor atau sarkasme dalam menanggapi topik yang dibahas [7]. Dalam menanggapi situasi ini, penting untuk mencari solusi yang efektif, seperti menggunakan pendekatan berbasis teknologi, salah satunya menggunakan Machine Learning (ML) melalui proses analisis polaritas sentimen pada komentar-komentar youtube.

Penelitian pertama yang dilakukan Chely Aulia Misrun, et al (2023), berhasil melakukan klasifikasi komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naïve bayes classifier yang mendapatkan hasil dengan akurasi sebesar 78% menggunakan 2 kelas, yaitu positif dan negatif [8]. Penelitian kedua, yang dilakukan Mita Tri Leony (2024), berhasil melakukan analisis sentiment masyarakat terhadao politik dinasti di Indonesia menggunakan metode naïve bayes classifier yang mendapatkan hasil akurasi sebesar 78,26% menggunakan 2 kelas, negatif dan positif [9]. Penelitian ketiga yang dilakukan oleh M. Hudha, et al (2022), berhasil melakukan analisis mengenai sentiment pengguna youtube trhadal tayangan mata najwa dengan metode naïve bayes yang mendapatkan hasil akurasi sebesar 90,36% menggunakan 3 kelas, yaitu negatif, positif, dan netral [10].

Maka demikian, hasil literatur review menunjukkan bahwa penelitian sebelumnya hanya melakukan klasifikasi sentimen dengan kategori positif, negatif, dan netral. Belum terdapat penelitian yang mencoba untuk mengidentifikasikan kategori komentar secara nada dan Bahasa. Pada penelitian penelitian terdahulu juga hanya menggunakan fitur TF-IDF dan algoritma *naïve bayes classifier*, padahal masih banyak teknik lain yang bisa dimanfaatkan dan dapat lebih baik dalam menangani masalah. Penelitian yang saat ini penulis kerjakan memanfaatkan pendekatan algoritma ML klasik seperti Naive Bayes dengan representasi teks klasik seperti Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW), 2-gram, 3-gram, dan word-embedding yaitu IndoBert. Dari beberapa metode yang digunakan diharapkan dapat memberikan hasil yang efektif dalam melakukan analisis dan evaluasi terhadap permasalahan yang sedang dihadapi.

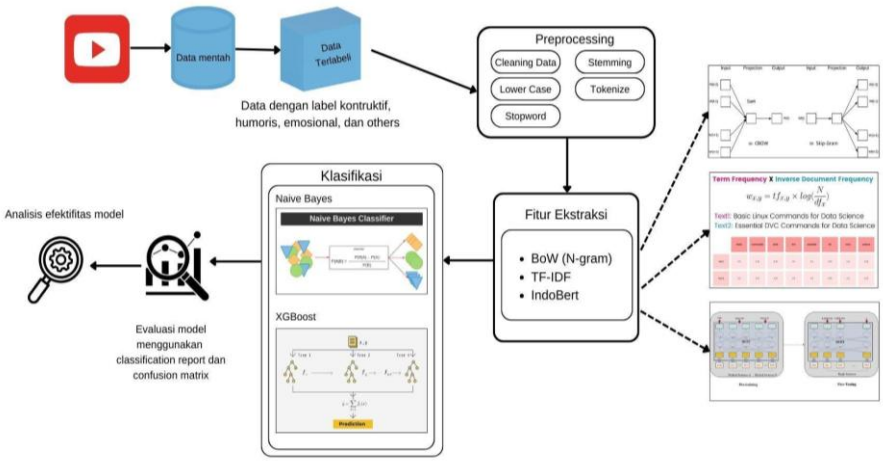
Penelitian ini bertujuan untuk menganalisis respons netizen dalam menanggapi permasalah politik di Indonesia menggunakan nada dan bahasa yang terdapat pada komentar pada konten youtube Pandji Pragiwaksono menggunakan kategori yang lebih detail terkait dengan yang digunakan pada komentar. Kontribusi penulis pada penelitian ini antara lain:

1. Membuat dataset dengan menggunakan teknik *crawling* data pada komentar youtube.
2. Melakukan pelabelan dengan menggunakan 4 kategori yaitu, kontruktif, emosional, humoris, dan others (lainnya).

3. Menggunakan Teknik analisis dengan berbagai fitur representasi teks menggunakan Teknik mulai dari klasik, contextual word-embedding. Dan membangun model menggunakan algoritma Naive Bayes Classiffier dan XGBoost.
4. Penelitian ini diharapkan dapat memberikan manfaat untuk para konten creator dalam mempertimbangkan konten-konten apa saja yang dapat menarik perhatian masyarakat.

2. Metode Penelitian

Berikut ini rancangan skema eksperimen yang akan diterapkan untuk mendapatkan hasil berupa model evaluasi yang dapat dilihat pada Gambar 1.



Gambar 1. Skema Penelitian

Skema penelitian ini menggambarkan alur analisis polaritas komentar terkait konten politik. Tahap pertama melakukan crawling data menggunakan API youtube, kemudian dilakukan pelabelan secara manual menjadi empat kategori yaitu, kontruktif, emosional, humoris, dan others (lainnya). Selanjutnya data melalui tahap *preprocessing* dan memastikan data siap untuk proses selanjutnya. Pada proses klasifikasi, terdapat 2 model algoritma yaitu *naïve bayes* dan *xgboost*. Selanjutnya dilakukan evaluasi untuk membanding performa masing-masing model.

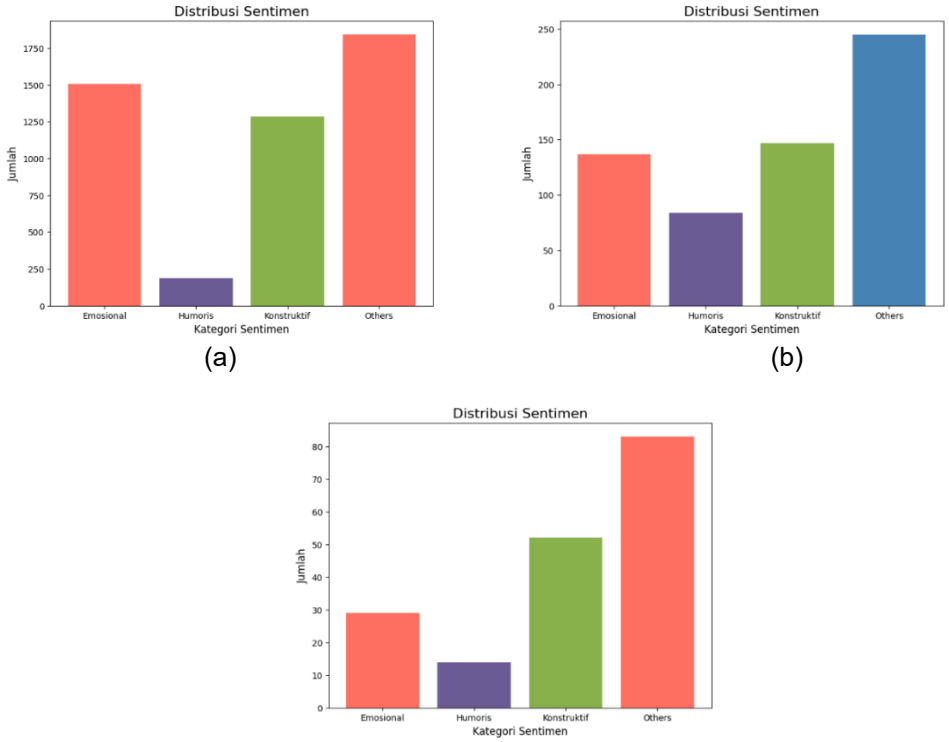
2.1 Dataset

Dalam penelitian ini menggunakan data teks yang diperoleh dari komentar media sosial youtube Indonesia dengan rentang waktu Januari-Agustus 2024 yang didapatkan dengan melakukan crawling data. Dari hasil crawling data komentar, diperoleh tiga dataset dengan data yang berjumlah banyak (4000 data komentar), data yang berjumlah sedang (1000 data komentar), dan data yang berjumlah sedikit (300 data komentar). kemudian dilakukan proses pelabelan pada masing-masing dataset secara manual dengan total 3 pembagian sentimen yaitu, konstruktif, emosional, humoris, dan others. Table 1 merupakan contoh dari masing-masing kategori sentimen.

Tabel 1. Sampel Data Komentar Pada Setiap Label Sentimen

No	Sampel Komentar	Label
1	Memang cuakks juga si daniel? pakek kalimat gitu ngapain? malah jadinya penghinaan	Humoris
2	ini pelajaran penting buat semua orang untuk kehidupan sehari-hari yang gk mungkin lepas dari masalah, meskipun kita benar kita gk bisa se enaknya ketika ada celah dikit aja itu bisa jadi bumerang untuk kita sendiri.	Konstruktif
3	Intinya mah dicari cari kesalahannya	Emosional
4	Terimakasih para member?	Others

Dari 3 dataset yang sudah dilakukan pelabelan sentiment, didapatkan jumlah sentiment dari masing-masing dataset sebagai berikut.



Gambar 2. Jumlah Distirbusi Sentiment Pada (a) Data 1 (4000), (b) Data 2 (1000), (c) Data 3 (150)

2.2 Preprocessing

Tahapan ini sangat penting yang mana tahapan preprocessing dilakukan untuk mempersiapkan data dan memperbaiki kualitas data dan dapat meningkatkan performa model. Ada beberapa tahan preprocessing seperti *cleaning data*: membersihkan data dari symbol, emoji, tanda baca, link url, dan angka. *Case folding*: dataset mengubah huruf kapital menjadi huruf kecil. *Tokenize*: kalimat komentar dipecah menjadi kalimat per kata. *Stopword removal*: Penghapusan kata hubung atau kata-kata yang tidak penting dalam kalimat, misal “di”, “ke”, “dari”, atau “yang”. Dan yang terakhir *stemming*: Pada proses ini dilakukan pengubahan kata-kata yang memiliki imbuhan seperti “menculik” menjadi “culik” [10]. Dalam penelitian ini data dibagi menjadi tiga bagian yaitu data latih (80%) dan data test (20%).

2.3 Fitur Ekstraksi Teks

Penelitian ini menggunakan berbagai macam fitur ekstraksi, mulai dari fitur klasik, hingga contextual word-embedding. Secara lebih spesifik, fitur ekstraksi teks yang digunakan sebagai berikut: (a) fitur klasik Bag of Word (BoW), merupakan sebuah mesin yang menerima input berupa dokumen dan menghasilkan sebuah table yang berisi jumlah frekuensi kata yang tersedia untuk setiap dokumen [11]. (b) TF-IDF merupakan gabungan dari 2 metod yaitu TF, mengukur pentingnya kata yang dilihat dari seberapa sering kata tersebut muncul dalam dokumen. sedangkan IDF, adalah kemunculan kata terhadap keseluruhan dokumen dalam dataset. Dengan demikian, TF-IDF adalah ukuran yang dinormalisasi yang mempertimbangkan Panjang dokumen [12]. (c) IndoBert, merupakan model yang telah dilatih sebelumnya dengan menggunakan set data bahasa Indonesia yang ekstensif dan tersanitasi. IndoBert menjalani pelatihan dengan menggunakan korpus yang terdiri dari lebih 220 juta kata dalam bahasa Indonesia [13]. (d) 2-Gram menurut Manning dan Schütze (1999), 2-gram didefinisikan sebagai "urutan dua item yang berdekatan dari rangkaian yang diberikan". Dalam konteks NLP, item-item ini biasanya berupa kata-kata dalam teks [14]. (e) 3-Gram, seperti yang dijelaskan oleh Jurafsky dan Martin (2009),

adalah "urutan tiga item yang berdekatan". Dalam analisis teks, ini berarti tiga kata yang berurutan [15].

2.4 Augmentasi Data

Augmentasi data teks merupakan teknik oversampling bertujuan untuk menghasilkan data pelatihan sintesis tambahan data yang dilakukan pada data yang tidak mencukupi [29]. Augmentasi data teks dapat dilakukan dengan teknik *synonym replacement* menggunakan *library nlpaug*. Pustaka python ini menghasilkan data sintesis untuk meningkatkan kinerja model tanpa upaya manual [30]. Teknik ini memiliki skenario dengan mengganti beberapa kata dalam kalimat dengan kata sinomin tanpa merubah makna kalimat, kemudian hasil data augmentasi akan ditambahkan pada data asli untuk membentuk dataset baru. Berikut merupakan contoh kalimat augmentasi.

Tabel 2. Contoh Kalimat Sebelum dan Sesudah Augmentasi

No	Sebelum Augmentasi	Sesudah Augmentasi
1	bagus tepat sasaran jangan berfikir emosional lawan kita jahat	baik tepat sasaran jangan berfikir emosional musuh kita jahat
		keren tepat sasaran jangan berfikir emosional rival kita jahat

Proses augmentasi dilakukan dengan mengganti kata seperti kata sifat, kata kerja, maupun kata benda dalam kalimat menggunakan sinonim dari kata tersebut tanpa merubah susunan dan makna dari kalimat.

2.5 Skenario Pengolahan Data

2.5.1 Skenario 1

Melakukan klasifikasi menggunakan data original yang sebelumnya sudah melewati tahap preprocessing data.

2.5.2 Skenario 2

Melakukan klasifikasi menggunakan data augmentasi yang diimplementasikan pada data train.

2.6 Klasifikasi

Penelitian ini mencoba untuk menggunakan 2 model yaitu *naïve bayes* dan *xgboost*. *Naïve bayes* adalah algoritma pembelajaran sederhana yang memanfaatkan aturan Bayes, yang menganggap bahwa atributnya secara kondisional independen dari kelasnya[28]. NB mengadopsi pendekatan Bag-of-Words (BOW). Persamaan 1 dari teorema bayes [18]:

$$P(H|\mathrm{X}) = \frac{\mathrm{P(X\backslash H).\,P(H)}}{\mathrm{P(X)}} \quad (1)$$

Sedangkan XGBoost adalah metode pengajaran mesin yang berbasis Gradient Boosting Decision Tree untuk analisis regresi dan klasifikasi. XGBoost menemukan cara terbaik untuk menyeimbangkan penurunan fungsi dan menghindari overfitting dengan menggunakan ekspansi Taylor[19].

2.7 Evaluasi

Penelitian ini menggunakan *classification report* untuk mengukur kinerja model klasifikasi. Laporan klasifikasi memberikan informasi rinci tentang kinerja model klasifikasi pada data uji mengenai *accuracy*, *precision*, *recall*, dan *f1-score*. [20]. Dari beberapa model dan fitur ekstraksi yang sudah diimplementasikan, akan dilakukan perbandingan untuk mendapatkan hasil performa terbaik dari model dalam eksperimen ini.

3. Hasil Penelitian dan Pembahasan

Pada bagian ini, menyajikan hasil eksperimen dari penelitian tentang nada dan bahasa dalam komentar Youtube dengan topik politik. Proses analisis hasil klasifikasi dilakukan dengan membandingkan setiap metode berdasarkan *classification report* yang berupa *recall*, *pressicion*, *f1-score*, dan *accuracy*. Penelitian ini menggunakan pembagian 80% train, 10% test, dan 10% valid pada eksperimen klasifikasi.

Berikut adalah inisialisasi dataset:

- Data 1 = Dataset dengan jumlah komentar sedikit (150 komentar)
- Data 2 = Dataset dengan jumlah komentar sedang (1000 komentar)
- Data 3 = Dataset dengan jumlah komentar banyak (4000 komentar)

Berikut adalah Tabel 3 perbandingan dari model yang sudah digunakan untuk mengklasifikasikan data.

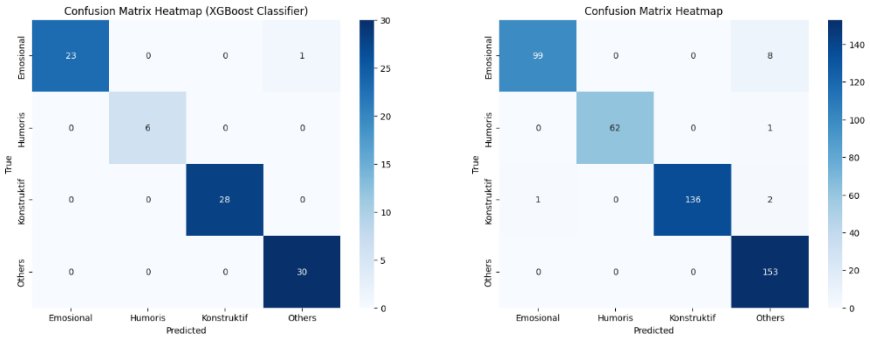
Tabel 3. Perbandingan Antar Motode

Fitur Representasi Teks + Model	Data 1		Data 2		Data 3	
	Ori	**Aug**	**Ori**	**Aug**	**Ori**	**Aug**
BoW + Naïve Bayes	66%	94%	50%	86%	60%	73%
FT-IDF + Naïve Bayes	**70%**	84%	48%	92%	**70%**	73%
2-Grams + Naïve Bayes	**70%**	91%	43%	**97%**	43%	88%
3-Grams + Naïve Bayes	37%	93%	39%	95%	39%	86%
IndoBert Embedding +Naïve Bayes	59%	80%	48%	68%	52%	55%
BoW + XGBoost	57%	**97%**	53%	96%	60%	74%
FT-IDF + XGBoost	64%	89%	57%	92%	52%	74%
2-Grams + XGBoost	44%	51%	41%	62%	48%	57%
3-Grams +XGBoost	34%	44%	39%	38%	39%	42%
IndoBert Embedding + XGBoost	**64%**	**99%**	**57%**	**93%**	**70%**	**92%**

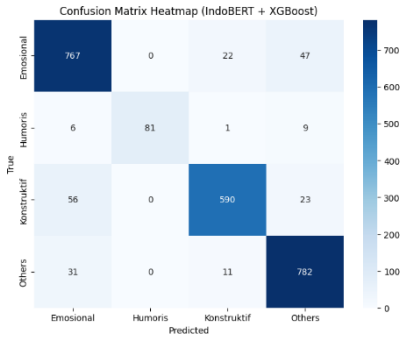
Pada penerapan skenario 1, algoritma naïve bayes dan fitur ekstraksi TF-IDF dan 2-Gram memberikan performa terbaik yang memperoleh hasil 70% pada data 1. Sedangkan pada algoritma XGboost, fitur wod embedding menggunakan indobert yang memberikan hasil terbaik yaitu 64%. Selanjutnya eksperimen pada data 2 dan data 3, fitur indobert menggunakan algoritma xgboost mencapai akurasi tertinggi dengan hasil 57% pada data 2 dan 70% pada data 3. Eksperimen ini belum dikataan berhasil dikarenakan performa model yang sangat buruk, sehingga perlu dilakukan eksperimen lanjutan dengan mencoba menerapkan teknik augmentasi pada masing-masing data.

Sedangkan pada skenario 2, menunjukkan bahwa teknik augmentasi dapat meningkatkan performa model secara signifikan. Terlebihnya pada fitur word embedding dan model xgboost memiliki hasil yang terbaik pada data 1 dan data 3 dengan nilai sebesar 99% pada data 1 dan 92% pada data 2. Sebaliknya, pada data 2 performa tertinggi didapatkan oleh fitur ekstraksi teks 2-gram dengan akurasi 97%. Jadi secara keseluruhan penerapan skenario 1 dan skenario 2, dapat dibandingkan bahwa model pada skenario 2 menggunakan augmentasi dapat meningkatkan performa model secara signifikan. Terutama pada teknik *word embeddings* menggunakan *indobert* terjadi peningkatan pada masing masing data.

Gambar 3 berikut adalah *heatmap* yang menunjukkan model dengan performa terbaik dari data 1, data 2, dan data 3.



(a) Data 1 (b) Data 2



(c) Data 3

Gambar 3. Heatmap yang Menunjukkan Performa Terbaik

Pada data 1, perpaduan antara fitur *word embedding indobert* dan *xgboost* memiliki performa tertinggi mencapai 99% Dimana hamper seluruh data dapat diklasifikasikan dengan tepat. Pada data 2, fitur *2-gram* dan *naïve bayes* memiliki performa terbaik dalam melakukan klasifikasi. Dimana fitur klasik *2-gram* lebih mampu untuk menangkap frasa yang bermakna sehingga memiliki keunggulan karena model bisa lebih efisien dan minim resiko *overfitting*. Sedangkan pada data 3, fitur *word embedding indobert* dan *xgboost* memiliki performa terbaik mencapai 92%.

4. Kesimpulan

Berdasarkan penelitian ini, teknik *word embeddings contextual* merupakan salah satu teknik ekstraksi teks yang dikembangkan untuk melakukan klasifikasi teks dan dapat menunjukkan performa yang baik dibandingkan dengan teknik ekstraksi teks dengan fitur klasik. Kami juga menemukan bahwa kombinasi penggunaan data augmentasi mampu memberikan performa terbaik dalam melakukan klasifikasi dibanding dengan menggunakan data original. Akan tetapi Melalui penelitian ini, diharapkan dapat membantu satu sama lain dalam menentukkan konten politik apa yang mungkin akan mendapatkan respon baik maupun buruk oleh masyarakat. Penelitian ini belum bisa dikatakan sempurna dikarenakan terdapat beberapa kekurangan yang terdapat didalamnya.

Referensi

[1] E. H. Susanto, “Media Sosial Sebagai Pendukung Jaringan Komunikasi Politik”.
[2] Geograf, “Pengertian Sistem Politik Indonesia: Definisi dan Penjelasan Lengkap Menurut Ahli,” Geograf. Accessed: Mar. 02, 2025. [Online]. Available: https://geograf.id/jelaskan/pengertian-sistem-politik-indonesia/
[3] R. K. Anwar, A. Rusmana, and M. T. Rahman, “The Politics Of Information On Traditional Medical Practices In Bandung Barat,” *Mimb. J. Sos. Dan Pembang.*, vol. 34, no. 1, pp. 158–165, Jun. 2018, doi: 10.29313/mimbar.v34i1.3256.
[4] “Biodata Pandji Pragiwaksono, Pendidikan, Perjalanan Karier, dan Prestasinya,” kumparan. Accessed: Mar. 02, 2025. [Online]. Availa ble: https://kumparan.com/profil-tokoh/biodata-pandji-pragiwaksono-pendidikan-perjalanan-karier-dan-prestasinya-22ux6z9UWM1

[5] Eriyanto, *Analisis Isi: Pengantar Metodologi untuk Penelitian Ilmu Komunikasi dan Ilmu-ilmu Sosial Lainnya*. Prenada Media, 2015.
[6] "(PDF) Komunikasi Dan Media Sosial," ResearchGate. Accessed: Mar. 02, 2025. [Online]. Available: https://www.researchgate.net/publication/329998890_KOMUNIKASI_DAN_MEDIA_SOSIAL
[7] S. A. RIZKI, "Analisis Sentimen Pada Kolom Komentar Media Sosial Youtube Terhadap Fenomena Childfree di Indonesia (Studi Kasus Akun Youtube Analisa Channel dan Adi Hidayat Official)," other, IAIN SALATIGA, 2023. Accessed: Mar. 02, 2025. [Online]. Available: http://e-repository.perpus.uinsalatiga.ac.id/18741/
[8] Chely Aulia Misrun, E. Haerani, M. Fikry, and E. Budianita, "Analisis sentimen komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier," *J. CoSciTech Comput. Sci. Inf. Technol.*, vol. 4, no. 1, pp. 207–215, Apr. 2023, doi: 10.37859/coscitech.v4i1.4790.
[9] "Leony, Mita Tri. Analisis Sentimen Masyarakat Terhadap Politik Dinasti Di Indonesia Menggunakan Metode Naive Bayes Classifier. Diss. UIN SUSKA RIAU, 202 - Yahoo Hasil Pencarian." Accessed: Mar. 02, 2025. [Online].
[10] Universitas Muria Kudus, M. Hudha, E. Supriyati, and T. Listyorini, "Analisis Sentimen Pengguna Youtube Terhadap Tayangan #Matanajwamenantiterawan Dengan Metode Naïve Bayes Classifier," *JIKO J. Inform. Dan Komput.*, vol. 5, no. 1, pp. 1–6, Apr. 2022, doi: 10.33387/jiko.v5i1.3376.
[11] F. Alzami, E. D. Udayanti, D. P. Prabowo, and R. A. Megantara, "Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis," in *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Aug. 2020, pp. 235–242. doi: 10.22219/kinetik.v5i3.1066.
[12] "Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)." Accessed: Mar. 04, 2025. [Online]. Available: https://www.jurnal.iaii.or.id/index.php/RESTI/
[13] P. Sayarizki, Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Indones. J. Comput. Indo-JC*, vol. 9, no. 2, pp. 61–72, Aug. 2024, doi: 10.34818/INDOJC.2024.9.2.934.
[14] "Foundations of Statistical Natural Language Processing." Accessed: Feb. 21, 2025. [Online]. Available: https://nlp.stanford.edu/fsnlp/
[15] "GloVe: Global Vectors for Word Representation - ACL Anthology." Accessed: Mar. 04, 2025. [Online]. Available: https://aclanthology.org/D14-1162/
[16] B. Li, Y. Hou, and W. Che, "Data Augmentation Approaches in Natural Language Processing: A Survey," *AI Open*, vol. 3, pp. 71–90, 2022, doi: 10.1016/j.aiopen.2022.03.001.
[17] *nlpaug: Natural language processing augmentation library for deep neural networks*.
[18] S. Thomas, Y. Yuliana, and N. P, "Study Analisis Metode Analisis Sentimen pada YouTube," *J. Inf. Technol.*, vol. 1, no. 1, pp. 1–7, 2021, doi: 10.46229/jifotech.v1i1.201.
[19] "(PDF) Perbandingan Algoritma XGBoost dan SVM Dalam Analisis Opini Publik Pemilihan Presiden 2024," *ResearchGate*, Feb. 2025, doi: 10.33022/ijcs.v13i3.4041.
[20] "(PDF) Klasifikasi Opini Publik di Twitter Terhadap Bakal Calon Presiden Indonesia Tahun 2024 Menggunakan LSTM Secara Realtime Berbasis Website," *ResearchGate*, Oct. 2024, doi: 10.35970/infotekmesin.v14i2.1908.