

Optimizing Complement Naive Bayes for Imbalanced Multiclass Election Tweet Classification Using a Balanced Class Approach with Optuna

Pieter Christy Yan Yudhistira*, Joshua Dwiputra Rendro Joelaskoro†,
Vanny Ade Gunawan‡, Christopher Robin Tanugroho§, Fitra Abdurrachman Bachtiar¶

Informatics Department, Faculty of Computer Science, Brawijaya University*†‡§¶

email: *pityudhistira28@student.ub.ac.id †joshuadwiputra@student.ub.ac.id ‡vannyade@student.ub.ac.id
§williamchen1506@student.ub.ac.id ¶fitra.bachtiar@ub.ac.id

Abstract—Pemilihan Umum 2024 di Indonesia merupakan salah satu peristiwa politik terbesar yang melibatkan sekitar 204,8 juta pemilih. Partisipasi publik, terutama di kalangan generasi muda, meningkat secara signifikan melalui platform seperti Twitter, yang menjadi sarana utama untuk berdiskusi tentang calon pemimpin, kebijakan, dan isu-isu nasional. Analisis cuitan ini untuk mengelompokkan topik berdasarkan kerangka ASTAGATRA—yang terdiri dari Trigatra (Geografi, Demografi, Sumber Daya Alam) dan Pancagatra (Ideologi, Politik, Ekonomi, Sosial Budaya, dan Pertahanan Keamanan)—sangat penting untuk memahami opini publik. Namun, proses ini menghadapi tantangan besar karena volume data yang sangat besar, keragaman konten, dan ketidakseimbangan jumlah data di setiap kategori.

Penelitian ini bertujuan untuk mengatasi tantangan tersebut dengan mengembangkan model klasifikasi menggunakan Complement Naive Bayes (CNB) dan Multinomial Naive Bayes (MNB), yang dioptimasi dengan hyperparameter menggunakan Optuna dan teknik resampling seperti SMOTE. Metode ini berhasil meningkatkan performa model dengan mengatasi ketidakseimbangan kelas dan mengoptimalkan parameter seperti smoothing. Model yang diusulkan mencapai balanced accuracy sebesar 60,19%, meningkat secara signifikan dibandingkan baseline cross-validation accuracy sebesar 50,44%. Hasil ini menunjukkan bahwa kombinasi teknik resampling dan optimasi hyperparameter efektif untuk klasifikasi teks multi-kelas.

Hasil penelitian menunjukkan bahwa model dapat melakukan prediksi yang lebih baik, baik untuk kelas dominan maupun kelas yang kurang terwakili seperti Demografi dan Geografi. Penelitian ke depan dapat mengintegrasikan teknik *preprocessing* lanjutan, seperti lemmatization, penanganan kosakata kompleks Twitter, dan penerapan POS tagging. Selain itu, eksplorasi model *deep learning* atau arsitektur pre-trained dapat lebih meningkatkan akurasi dan generalisasi klasifikasi. Pendekatan ini berkontribusi pada pengelolaan ketidakseimbangan kelas dalam klasifikasi teks cuitan dan memberikan wawasan terkait opini publik selama Pemilu 2024.

Index Terms—text classification, ASTAGATRA, naive bayes, optuna, balanced accuracy

I. INTRODUCTION

A. Latar Belakang

Pemilihan Umum 2024, atau yang lebih dikenal sebagai Pemilu 2024, merupakan salah satu peristiwa politik terbesar dalam sejarah Indonesia. Terlaksana di 14 Februari 2024, pemilu ini mencakup pemilihan legislatif (DPR, DPD, dan

DPRD) serta pemilihan calon presiden dan wakil presiden, akan diikuti kurang lebih 204,8 juta pemilih [1]. Hasil dari Pemilu 2024 ini akan menjadi penentu arah masa depan Indonesia dalam lima tahun mendatang. Tidak mengherankan jika masyarakat Indonesia sangat antusias untuk menyuarakan opini dan pandangan mereka tentang calon pemimpin negara [2].

Media sosial telah menjadi bagian penting dari sarana diskusi politik di Indonesia, terutama di kalangan generasi muda yang semakin aktif berpartisipasi dalam diskusi politik secara daring. Salah satu platform media sosial yang paling populer untuk tujuan ini adalah Twitter. Peran Twitter sebagai medium diskusi menjadi semakin signifikan menjelang Pemilu, sebagaimana dibuktikan oleh penggunaan hashtag #Pemilu2024 yang telah muncul dalam lebih dari seratus ribu cuitan [3] [4]. Banyak individu, kelompok, organisasi masyarakat, media massa, hingga calon pemimpin memanfaatkan Twitter untuk menyampaikan pendapat atau opini, berbagi berita atau informasi terbaru, serta berinteraksi mengenai isu-isu politik yang berkembang. Opini yang disuarakan oleh pengguna Twitter di Indonesia mencakup berbagai topik, mulai dari tanggapan terhadap calon pemimpin, respons terhadap debat antarcalon, komentar mengenai kebijakan dan janji kampanye, hingga keluhan dan harapan terkait masa depan Indonesia.

Beragamnya topik yang dibahas mendorong berbagai pihak, mulai dari pengamat politik hingga generasi muda, untuk mencari gambaran menyeluruh tentang opini masyarakat. Salah satu pendekatan yang diusulkan untuk mengelompokkan opini ini adalah menggunakan kerangka nilai ASTAGATRA. ASTAGATRA merupakan pendekatan yang mengacu pada aspek kehidupan nasional, yang terdiri dari nilai-nilai Trigatra alamiah (Geografi, Demografi, Sumber Daya Alam) dan Pancagatra sosial (Ideologi, Politik, Ekonomi, Sosial Budaya, dan Pertahanan Keamanan) [5]. Pendekatan ini tidak hanya mempermudah pengelompokan berdasarkan aspek nasional, tetapi juga membantu dalam mengidentifikasi dan memahami opini serta keluhan masyarakat di era digital.

Namun, dengan volume cuitan di Twitter yang sangat besar dan beragamnya jenis opini serta informasi yang disajikan,

muncul tantangan besar dalam pengelolaannya. Besarnya data ini tentu menyulitkan analisis dan klasifikasi cuitan terkait Pemilu 2024 ke dalam berbagai topik. Selain itu, interpretasi terhadap topik yang diangkat dalam setiap cuitan dapat berbeda tergantung pada pihak yang melakukan pengelompokan. Akibatnya, memperoleh inti informasi yang relevan terkait politik dan opini masyarakat menjelang Pemilu menjadi semakin kompleks dan menantang.

Selain tantangan volume, pengelompokan topik berdasarkan ASTAGATRA juga memiliki tantangan lain. Menghadapi permasalahan klasifikasi multi kelas, tidak jarang ada kelas tidak direpresentasikan dengan seimbang. Contohnya, setelah melakukan investigasi, topik seperti Demografi atau Geografi memiliki jumlah cuitan yang lebih sedikit jika dibandingkan topik-topik yang lebih populer seperti Politik. Sehingga hal tersebut menjadi tantangan klasifikasi permasalahan yang melibatkan multi kelas, tidak terkecuali permasalahan teks Twitter.

B. Masalah yang di Angkat

Berdasarkan permasalahan yang telah diidentifikasi, diperlukan sebuah model otomatisasi untuk menganalisis topik yang diangkat dalam cuitan Twitter selama Pemilu 2024. Model ini harus mampu mengklasifikasi cuitan dalam volume besar ke dalam kategori ASTAGATRA, yang meliputi topik Geografi, Demografi, Sumber Daya Alam, Ideologi, Politik, Ekonomi, Sosial Budaya, serta Pertahanan dan Keamanan. Selain itu, model ini juga harus dirancang untuk menghadapi tantangan persebaran topik yang tidak merata dalam cuitan Twitter dan tetap dapat mengidentifikasi topik-topik tersebut dengan akurat dan efektif.

C. Kajian Literatur

Berdasarkan kajian literatur sebelumnya, tantangan pertama yang perlu dijawab adalah bagaimana mengolah teks secara efektif. Penelitian [6] melakukan studi komparatif terhadap berbagai metode *preprocessing* teks, misalnya tokenisasi, stop-word removal, case folding, hingga stemming. Hasil penelitian tersebut menunjukkan bahwa penggunaan metode *preprocessing* dapat bervariasi tergantung pada kasus yang dihadapi dan tujuan analisis. Penyesuaian atau kustomisasi metode *preprocessing* juga disarankan untuk meningkatkan efektivitas pengolahan teks dalam pengembangan model.

Temuan ini diperkuat oleh penelitian [7], yang menganalisis dampak *preprocessing* pada klasifikasi sentimen teks Twitter terkait Covid-19. Hasilnya menunjukkan adanya peningkatan akurasi hingga 73% dibandingkan dengan analisis tanpa *preprocessing*, dengan rata-rata peningkatan sekitar 40%. Penelitian juga menunjukkan implementasi teknik pengolahan teks seperti *Annotation*, *polarity* dan *filter token* untuk meningkatkan kemampuan klasifikasi model.

Kami perlu mengidentifikasi teknik pengolahan yang mampu mengekstrak nuansa dari teks di Twitter. Penelitian [8] mengeksplorasi normalisasi teks untuk mengembangkan klasifikasi semantik hashtag. Hasil penelitian tersebut menunjukkan bahwa segmentasi hashtag dapat mendukung interpre-

tasi dan pemahaman teks Twitter yang sering kali dipenuhi berbagai hashtag. Selain itu, penelitian [9] mengeksplorasi tokenisasi pada teks *micro-blogging*, termasuk teks Twitter, untuk mengekstrak informasi dari singkatan dan emoji ASCII. Mereka melaporkan peningkatan kinerja hingga 96% dalam F-measure dan menekankan pentingnya tokenisasi yang lebih beragam dibandingkan dengan *tokenizer* standar.

Salah satu *library* yang kami identifikasi adalah Sastrawi, sebuah *library* yang dirancang untuk mendukung pengolahan teks dalam Bahasa Indonesia. Penelitian [10] menemukan bahwa pengolahan teks menggunakan Sastrawi menunjukkan performa yang lebih baik dalam klasifikasi dibandingkan dengan algoritma *Tala-Porter*, dengan hasil *exact match stemmer* mencapai 92% dan waktu proses yang lebih cepat. Selain itu, Sastrawi mampu menangani permasalahan *overstem* dan *understem* pada teks berbahasa Indonesia, termasuk pengolahan imbuhan khas bahasa Indonesia.

Topik lain yang sering dieksplorasi adalah implementasi *machine learning* dan *deep learning* dalam klasifikasi teks Twitter. Penelitian [11] membandingkan berbagai model *machine learning* seperti Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, dan Support Vector Machine (SVM). Hasil penelitian tersebut menunjukkan bahwa SVM memiliki performa tertinggi pada benchmark untuk kasus *binary classification*. Selain itu, penelitian [12] juga mengeksplorasi pengaruh teks terhadap arsitektur jaringan saraf menggunakan model klasifikasi teks terkini (*state-of-the-art*).

D. Metode Usulan

Kami mengusulkan optimasi klasifikasi topik cuitan Twitter pemilu 2024 dengan implementasi Complement Naive Bayes dan pengoptimalan parameter dengan Optuna. Kami juga mengeksplorasi target model yang mampu mengklasifikasikan teks dengan hasil yang seimbang dengan teknik resampling. Hal ini bertujuan untuk mampu mengenali topik secara keseluruhan dibandingkan sebatas akurasi model tertinggi. Dengan optimasi parameter, kami berhasil mengembangkan model dengan generalisasi yang baik dalam memprediksi topik dari dataset yang tidak seimbang.

E. Batasan Penelitian

Penelitian ini memiliki batasan-batasan sebagai berikut:

- 1) Penelitian berfokus pada dampak Optuna dalam optimasi dua model klasifikasi, yaitu Multinomial dan Complement Naive Bayes. Oleh karena itu, model lainnya, termasuk model berbasis *deep learning* atau *state-of-the-art* seperti INDOBERT, tidak menjadi bagian lingkup penelitian ini.
- 2) Penelitian ini juga terbatas pada pengembangan model yang telah dikembangkan sebelumnya dalam rangka mewakili Universitas Brawijaya dalam lomba Big Data Challenge Satria Data 2024. Evaluasi hasil prediksi selama lomba akan termasuk dalam penelitian ini.

II. BAGIAN METODOLOGI USULAN

Bagian ini menjelaskan gambaran umum alur penelitian yang ditunjukkan pada Gambar 1. Penelitian ini dimulai dari menyiapkan dataset cuitan pemilu Twitter dan membagi data latih dan validasi. *Preprocessing* dilakukan untuk membersihkan semua data yang digunakan mulai dari menghadapi *hashtag*, *tag*, *username*, *case folding* hingga *stemming*. Kemudian, pembobotan kata menggunakan *TF-IDF vectorization* dan teknik *resampling* RandomOverSampler dan SMOTE.

Pelatihan model dilakukan dengan optimasi parameter dengan Optuna dan mengimplementasikan Multinomial Naive Bayes dan Complement Naive Bayes. Terakhir, model dievaluasi menggunakan *classification report*, *confusion matrix* dan metrik *balanced accuracy* untuk menemukan model terbaik untuk prediksi data uji.

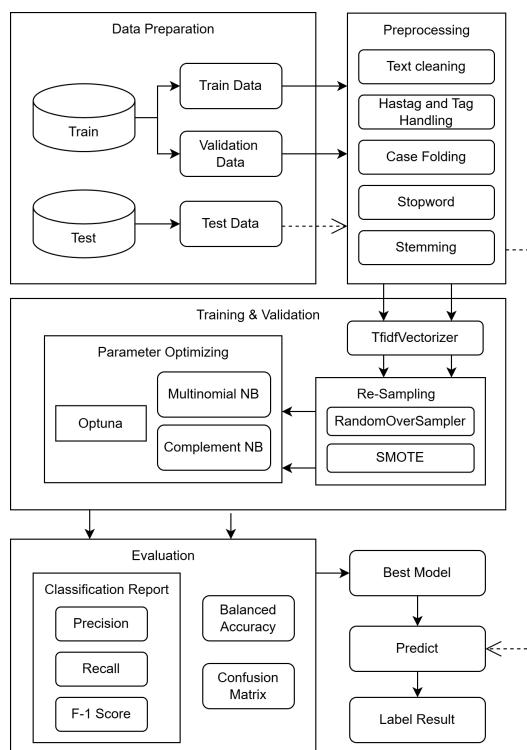


Fig. 1: Diagram Alur Eksperimen Penelitian

A. Dataset

Dataset yang kami gunakan untuk mengembangkan model klasifikasi kami merupakan dataset yang telah disediakan tim panitia Big Data Challenge Satria Data 2024. Dataset terdiri dari data yang sudah dipisahkan untuk latih dan uji kemampuan model. Data tersebut merupakan cuitan Twitter mengenai Pemilu Indonesia 2024 dan calon pasangan presiden dan wakil presiden Indonesia. Kami juga mengidentifikasi waktu pengambilan teks Twitter diambil pada masa kampanye Pemilu Indonesia 2024, yaitu sekitar Desember 2023 hingga Februari 2024. Data train sendiri juga telah dilabeli oleh tim panitia berdasarkan topik ASTAGATRA. Contohnya dapat dilihat dengan Gambar 2.

RT Alasan Anies pilih JIS ketimbang GBK untuk Kampanye Akbar terakhir :Ã¢Â€Âœlni karya anak bangsa yang dibangun tanpa tenaga asingÃ¢Â€Â Pemimpin yg baik yang pro rakyat dan bangga dgn hasil kerja anak bangsa dan itu ada di diri Abah Anies, Indonesia gak butuh pemimpin yg letoyÃ¢Â² dan cengeng https://t.co/U3zpARCjQ7 [RE AnKiiim_]	Politik
Luar biasa. Pasangan Capres Cawapres Ganjar Pranowo Mahfud MD perjuangkan pertumbuhan koprasir dan UMKM dengan kredit 35% . Mari bersama bangun ekonomi yang kuat .#GanjarMahfudRebound #GanjarPranowoPilihanUmat #JNK	Ekonomi
Adi menginformasikan bahwa isu utama di kampung mereka adalah sulitnya mendapatkan air bersih dan masalah naiknya air laut . #IndonesiaSentris #IndonesiaHijau #02Melanjutkan #AnakMudaIndonesiaEmas Prabowo Subianto	Sumber Daya Alam

Fig. 2: Contoh pelabelan ASTAGATRA

Topik cuitan diambil berdasarkan tema yang diangkat dalam teks dan dicocokkan dengan topik yang ada dalam ASTAGATRA. Sebagai contoh:

- 1) Pada teks pertama, terdapat kata kunci seperti **kampanye** dan **pemimpin**, yang berhubungan dengan bahasan calon presiden Anies Baswedan yang memilih Stadion JIS sebagai lokasi kampanye pemilunya. Oleh karena itu, cuitan ini dapat dikategorikan ke dalam topik Politik.
- 2) Pada teks kedua, kata kunci seperti **UMKM**, **koperasi**, **kredit**, dan frasa **ekonomi yang kuat** menunjukkan adanya pembahasan mengenai ekonomi, khususnya mengenai kebijakan atau kondisi ekonomi yang berkaitan dengan usaha mikro, kecil, dan menengah (UMKM). Dengan demikian, teks ini dapat dikategorikan ke dalam topik Ekonomi.
- 3) Pada teks ketiga, kata kunci **air** dan pembahasan tentang **kesulitan mendapatkan air bersih** menunjukkan adanya perhatian terhadap isu sumber daya alam, terutama yang berkaitan dengan masalah akses terhadap air bersih. Oleh karena itu, cuitan ini termasuk dalam topik Sumber Daya Alam.

Selain itu, kami telah memvalidasi kebenaran dan keaslian cuitan dengan mencari teks asli di platform Twitter. Kami mengecek mulai dari data berlabel maupun yang tidak berlabel. Hasil pencarian menunjukkan bahwa sebagian besar cuitan yang kami analisis merupakan teks asli. Cuitannya juga masih dapat diakses di Twitter pada saat penelitian ini dilakukan, yang berlangsung hingga Desember 2024.

B. EDA (Exploratory Data Analysis)

Sebelum melakukan eksperimen, kami melakukan eksplorasi lebih lanjut terhadap dataset untuk memahami pola-pola yang ada dan menggali informasi lebih dalam. Pertama-tama, kami mengidentifikasi label ASTAGATRA pada data latih (train) dan menganalisis distribusi label multi-kelas. Persebaran label teks dapat dilihat pada Gambar 3.

Kami menemukan bahwa distribusi label sangat tidak seimbang. Dari total 5000 data pelatihan, hampir 3000 data di antaranya berlabel Politik. Sementara itu, label-label lainnya, seperti Demografi dan Geografi, hanya terwakili oleh sejumlah data yang sangat sedikit, dengan label lain memiliki distribusi yang jumlahnya sedikit dibandingkan dengan label Politik. Temuan ini menunjukkan adanya ketidakseimbangan dalam distribusi data yang perlu diperhatikan dalam proses klasifikasi.

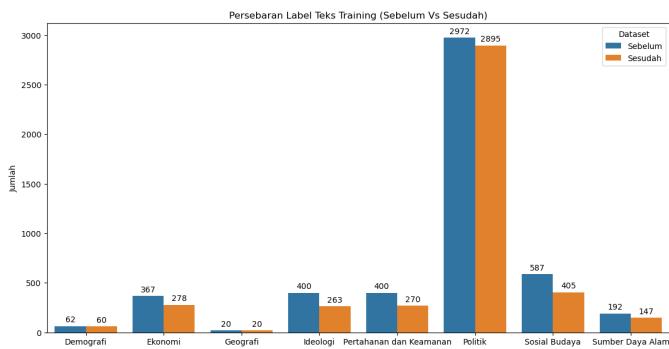


Fig. 3: Persebaran label teks training

Selanjutnya, kami melihat 10 kata atau token paling terbanyak dalam data latih. Sesuai dengan Gambar 4a, token yang paling banyak muncul adalah kata yang dapat diidentifikasi sebagai stopword (dan, yang, di, untuk, yg). Terdapat token RT dan [RE] yaitu token umum teks Twitter dan nama calon presiden (Anies, Prabowo dan Ganjar). Adanya kata “yg” menunjukkan adanya kata-kata yang harus ditangani sebelum implementasi model.

Seperti yang ditunjukkan pada Gambar 4b, token yang paling sering muncul setelah melalui tahap preprocessing adalah token yang berkaitan dengan calon presiden dan wakil presiden. Hal ini menunjukkan bahwa nama kandidat menjadi kata yang paling dominan dalam teks. Kami mempertahankan nama tersebut dalam analisis karena relevansi dan pengaruhnya terhadap konteks setiap dokumen. Selain itu, token lain yang sering muncul adalah ”pak,” ”jadi,” ”dukung,” dan ”Indonesia.”

Berdasarkan analisis terhadap 10 token terpopuler, ditemukan bahwa token-token ini merepresentasikan sejumlah besar kata dalam teks secara keseluruhan. Sebelum proses preprocessing, 10 token teratas muncul sebanyak 20.194 kata, atau setara dengan 12,3% dari total kata. Setelah preprocessing, jumlah kemunculan 10 token teratas menurun menjadi 10.963 kata, atau 11,6% dari total kata. Dengan demikian, dapat disimpulkan bahwa 10 token teratas memberikan kontribusi yang signifikan terhadap distribusi kata dalam teks, sehingga dapat mempengaruhi dalam pengembangan model ke depannya.

Tidak hanya keseluruhan token, kita bisa mengeksplor token khusus seperti hashtag, tautan, maupun username di Gambar 5 dan 6. Dalam gambar 5a, hashtag paling populer adalah hashtag yang mendukung calon presiden. Contohnya seperti

Common_words	count	Common_words	count		
0	dan	2845	0	anies	1976
1	RT	2795	1	ganjar	1776
2	[RE]	2795	2	prabowo	1622
3	yang	2554	3	pak	1050
4	di	2121	4	jadi	892
5	Ganjar	1766	5	pranowo	864
6	Anies	1672	6	mahfud	819
7	Prabowo	1304	7	dukung	743
8	untuk	1263	8	indonesia	660
9	yg	1079	9	jnk	607

(a) Original Data

(b) After Preprocessing

Fig. 4: Top 10 kata (token) yang paling banyak muncul

Common_words	count	Common_words	count		
0	#GanjarPranowoPilihanUmat	658	0	https://t.co/bxlvBhLvn2	42
1	#JNK	637	1	https://t.co/WmJR2OsbeZ	12
2	#GanjarMahfudRebound	558	2	https://t.co/S9S9Kanr0z	11
3	#GanjarMahfud2024	377	3	https://t.co/wK9XuBdLBr	11
4	#Coblos3	205	4	https://t.co/j5duokFbef	10
5	#AMINajaDulu	172	5	https://t.co/XpTnsPrinX	8
6	#02Melanjutkan	143	6	https://t.co/JkX9iUSAjR	8
7	#L3bihbaik	114	7	https://t.co/PbvdoVo2aly	8
8	#DuluJokowiSekarangGanjar	100	8	https://t.co/sV6ZKF9WZz	8
9	#IndonesiaSentris	91	9	https://t.co/K5uavFp1B1	7

(a) Hashtag Twitter

(b) Link

Fig. 5: Top 10 hashtag dan tautan yang paling banyak muncul

Common_words	count	Username	Count
0	@gQ+QGMYI209N7Py+h3gRyakQic4NLEkITolluALnZA=	Yurissa_Samosir	103
1	@ggAL2HicdWKRU2/VFUq3RTFxxtCsKyUXAKn9Soo=	ekowboy2	100
2	@ClqXqvGAT04tMt4OCATjo/q7vV/y8HeYal0gMfg8Y=	Mdy_Asmara1701	96
3	@L3R8XFbw3WGbxRPSj0/ohZTbqVGX7qtfwRg9zmhK7Q=	tempodotco	83
4	@0Zdeh9QcL+fs3hRaTcFuSLRH56REFyRLq4//dlc=	tvOneNews	60
5	@znOMP7ZMVU9dMuMNA/ciazC9q5+hwgVKrTsQNdQLKgTc=	kumparan	47
6	@hsjZceksZPM0NiWtknsbSbZ1XyaZTU4OHkGrcDk8=	BangPino_	41
7	@wEOWbjbQX93e2r1/iXQ1mV/rpGE9rnDgURpW0Y0=	OposisiCerdas	40
8	@vaBvSLyok3xdwCAiYpXkJarkzroalLp1Rdpv/z3CJE=	Paltiwest	38
9	@YcALzivlM9mnN2WtCz3omAmFrJU0a8m4seMBHJD2vc=	geloraco	34

(a) Username RT (Retweet)

(b) Username RE (Reply)

Fig. 6: Top 10 username yang paling banyak muncul

#AMINAjaDulu, #02Melanjutkan dan #GanjarPranowoPilihanUmat yang mewakili pasangan calon 01, 02 dan 03. Terdapat juga hashtag yang secara eksplisit tidak mendukung salah satu calon seperti #JNK dan #IndonesiaSentris. Setelah mengidentifikasi, #JNK sendiri merupakan organisasi dengan kepanjangan "Jaringan Nasional Keumatan Arsip" yang mendukung paslon 03 (Ganjar) [13]. Sedangkan #IndonesiaSentris merupakan sentimen terhadap tujuan kinerja mantan presiden Jokowi Dodo untuk mengubah fokus pembangunan keseluruh pelosok Indonesia, tidak hanya di pulau Jawa saja [14].

Berdasarkan top 10 tautan pada Gambar 5b, kami berhasil mengidentifikasi bahwa tautan tersebut menunjukkan postingan Twitter yang paling sering direply (balas) yang ada dalam data latih. Tautan "<https://t.co/txIvBhLvn2>" yang muncul sebanyak 42 kali ini merupakan tautan yang mendukung calon pasangan 01 (Anies) dan dibuktikan dengan adanya hashtag #AMINAjaDulu [15]. Username dari tautan tersebut, @ekowboy2 juga muncul sebagai username Reply terbanyak kedua pada Gambar 6b.

C. Data Preprocessing and Data Extraction

Sebelum mengimplementasikan model dan perancangan model, teks Twitter perlu melakukan proses *preprocessing*. Dampak pengolahan teks yang kami implementasikan juga berdasarkan konsiderasi untuk meningkatkan performa model [7] [6] [16] [17]. Berikut merupakan uraian teknik *preprocessing* yang diimplementasikan dalam model:

- 1) *Case Folding*: Semua teks diubah menjadi huruf kecil untuk memastikan bahwa kata-kata seperti "politik" dan "Politik" diperlakukan sama. Hal ini untuk menghindari kedua kata tersebut memiliki makna berbeda dan menyederhanakan pemahaman model.
- 2) Tokenisasi: Memecah teks yang merupakan kumpulan kata dalam bentuk kalimat menjadi token (kata atau frasa) yang terpisah. Contohnya kalimat *RT Anies Bakal Bangun 11 Stadion Sepak Bola Seperti JIS: Agar Iklim Kompetisi Sehat!* <https://t.co/VI5smUTpWH> [RE tvOne-News] menjadi "RT", "Anies", "Bakal" hingga "tvOne-News".
- 3) Penghapusan Stopwords: Stopwords merupakan token yang tidak membawa informasi penting dan berpotensi mempengaruhi performa model, seperti konjungsi (yang, jika, bila, atau, dan) atau preposisi (di, ke, dari). Token-token tersebut akan dihapus untuk menyederhanakan teks.
- 4) Penghapusan Tautan, Hashtag dan Username: Menghapus URL, simbol hashtag (#) dan mention username, karena informasi tersebut tidak relevan dalam konteks topik politik. Untuk hashtag sendiri, kami hanya menghapus simbol "#" dikarenakan kami mengidentifikasi adanya relevansi terhadap topik politik [8].
- 5) Stemming: Mengubah kata-kata ke bentuk dasarnya. Misalnya, kata "berkampanye" akan distem menjadi "kampanye". Kami mengimplementasikan stemming library Sastrawi karena kemampuannya dalam melakuakan pengolahan teks bahasa Indonesia [18].

- 6) Penggunaan TF-IDF (Term Frequency-Inverse Document Frequency): Teknik menghitung representasi numerik dari kata-kata dalam dokumen. Teknik ini memberikan bobot yang lebih besar pada kata-kata yang jarang muncul di seluruh data latih, namun sering muncul dalam satu dokumen.
- 7) Imbalanced Resampling: Menggunakan teknik seperti RandomOverSampler dan SMOTE (Synthetic Minority Over-sampling Technique) untuk menangani masalah ketidakseimbangan kelas. Tujuannya agar setiap kelas dapat direpresentasikan dan model mampu mengidentifikasi topik lebih baik.

D. Implementasi Model

Untuk mengklasifikasikan cuitan Twitter berdasarkan topik ASTAGATRA, penelitian ini mengimplementasikan dua model klasifikasi berbasis Naive Bayes, yaitu **Multinomial Naive Bayes** (MNB) dan **Complement Naive Bayes** (CNB). Model Multinomial Naive Bayes digunakan sebagai baseline, namun karena dataset yang digunakan memiliki ketidakseimbangan kelas, Complement Naive Bayes dipilih sebagai alternatif yang lebih cocok karena kemampuannya menangani masalah ketidakseimbangan kelas secara lebih efektif.

Pertama, dilakukan optimasi hyperparameter menggunakan framework Optuna, yang memanfaatkan teknik pencarian parameter terbaik, seperti smoothing parameter (alpha), untuk menghindari kemungkinan probabilitas nol serta parameter lain yang relevan untuk meningkatkan performa model. Selanjutnya, setelah parameter model dioptimalkan, pelatihan dilakukan dengan dataset pelatihan yang telah melalui tahap preprocessing. Akhirnya, model diuji menggunakan dataset uji, dan performanya dievaluasi melalui metrik-metrik tertentu untuk mengukur keberhasilan klasifikasi. Pendekatan ini bertujuan untuk memastikan prediksi yang lebih akurat dan seimbang, khususnya dalam menangani distribusi data yang tidak merata.

E. Evaluation

Untuk mengevaluasi performa model klasifikasi teks, beberapa metrik berikut digunakan:

- 1) Confusion matrix: Matriks yang menunjukkan jumlah prediksi yang benar dan salah per kelas.
- 2) Classification Report: Menyediakan informasi tentang precision, recall, dan F1-score untuk setiap kelas.
- 3) Balanced accuracy: Menghitung rata-rata akurasi per kelas, yang lebih cocok untuk dataset yang tidak seimbang. Formula balanced accuracy sebagai berikut:

$$\text{Balanced Accuracy} = \frac{\sum_{n=8}^n \text{Recall}}{8} \quad (1)$$

- 4) Precision, Recall, dan F1-Score: Precision mengukur ketepatan model dalam memprediksi positif, sedangkan recall mengukur seberapa banyak kasus positif yang berhasil ditemukan oleh model. F1 score merupakan rata-rata harmonis antara precision dan recall, yang

memberikan gambaran lebih baik mengenai keseimbangan keduanya.

$$\text{Recall} = \frac{tp}{tp + fn} \quad (2)$$

$$\text{Precision} = \frac{tp}{tp + fp} \quad (3)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

III. RESULTS AND DISCUSSION

Berikut ini adalah hasil eksperimen pengembangan model:

TABLE I: Komparasi Eksperimen Model

Tabel	Akurasi	Precision	Recall	F1 Score	Balanced Acc
Base MNB	0.6671	0.4450	0.6671	0.5338	0.1250
Base CNB	0.7097	0.6768	0.7097	0.6469	0.2468
SMOTE MNB	0.5403	0.6795	0.5403	0.5739	0.4586
SMOTE CNB	0.4355	0.6862	0.4355	0.4781	0.5407
Optuna SMOTE MNB	0.4171	0.6976	0.4171	0.4530	0.5583
Optuna SMOTE CNB	0.3456	0.7025	0.3456	0.3712	0.5550
Optuna RO MNB	0.3871	0.6850	0.3871	0.4179	0.5350
Optuna RO CNB	0.3710	0.6923	0.3710	0.4043	0.5330
Best Model (SMOTE CNB)	0.3677	0.7052	0.3677	0.4026	0.4932

Setelah melatih model dengan dataset pelatihan yang telah diproses, kami memperoleh hasil yang menunjukkan bahwa Complement Naive Bayes yang dioptimasi dengan Optuna memberikan hasil yang lebih baik dibandingkan dengan Multinomial Naive Bayes. Dengan teknik resampling dan optimasi parameter, model kami berhasil mencapai balanced accuracy yang signifikan lebih tinggi dibandingkan model tanpa penanganan ketidakseimbangan kelas.

Distribusi kelas yang lebih seimbang melalui SMOTE dan RandomOverSampler terbukti efektif dalam meningkatkan performa model, terutama pada kelas-kelas yang lebih sedikit terwakili seperti Demografi dan Geografi.

IV. KESIMPULAN DAN SARAN

Didasarkan hasil uji akurasi klasifikasi topik ASTAGATRA, implementasi usulan berhasil mencapai balanced accuracy 60,19 % dengan mengimplementasikan resampling SMOTE dan optimasi parameter Optuna dengan balanced class approach. Model menunjukkan generalisasi prediksi yang cukup baik, dikarenakan hasil uji berhasil meningkat dibandingkan hasil cross validasi model sebelumnya dengan balanced accuracy 50,44 %. Hasil percobaan ini menunjukkan bahwa model yang dikembangkan mampu menangani permasalahan ketidakseimbangan kelas dalam klasifikasi cuitan Twitter Pemilu 2024. Dengan demikian, pendekatan ini dapat menjadi solusi yang efektif untuk meningkatkan performa klasifikasi pada dataset yang tidak seimbang.

Ke depannya, pengembangan lebih lanjut dapat mencakup eksplorasi teknik ekstraksi dan preprocessing teks yang lebih mendalam. Contohnya, penggunaan lemmatization, penanganan kosakata kompleks khas pada Twitter, serta penerapan POS tagging untuk memperkaya representasi data teks. Selain itu, penelitian dapat diperluas dengan mengeksplorasi teknik *deep learning* atau pre-trained models yang lebih mutakhir. Implementasi model yang lebih kompleks diharapkan mampu

meningkatkan akurasi prediksi dan memberikan kemampuan generalisasi yang lebih baik pada klasifikasi teks.

ACKNOWLEDGMENT

Kami mengucapkan terima kasih yang sebesar-besarnya kepada tim panitia Big Data Challenge Satria Data 2024 yang telah menyediakan dataset yang kami gunakan dalam penelitian ini. Kami juga ingin menyampaikan apresiasi kepada tim BCC FullSenyum, khususnya kepada rekan Pieter Christy Yan Yudhistira, Akwila Febryan Santoso dan Richard, yang mewakili Universitas Brawijaya dalam ajang Satria Data 2024, serta yang telah mengizinkan kami untuk mengembangkan source code yang mereka buat dalam penelitian ini.

Selanjutnya, kami juga mengucapkan terima kasih kepada rekan-rekan kami, Dzaki Rafif Malik dan Nazura Wirayuda Tama, yang turut mewakili Universitas Brawijaya di Satria Data 2024. Terima kasih atas arahan, saran, dan dukungan mereka yang sangat berharga dalam kelancaran dan penyempurnaan penelitian ini.

REFERENCES

- [1] K. P. Umum, “Dpt pemilu 2024 dalam negeri dan luar negeri, 204,8 juta pemilih,” 2024. [Online]. Available: <https://www.kpu.go.id/berita/baca/11702/dpt-pemilu-2024-nasional-2048-juta-pemilih>
- [2] Oktavia, “Antusiasme menjelang pemilu 2024 tinggi: Tantangan pendekatan politik di kalangan anak muda,” *Kawula17*, 2023. [Online]. Available: <https://kawula17.id/artikel/antusiasme-menjelang-pemilu-2024-tinggi-tantangan-pendekatan-politik-di-kalangan-anak-muda>
- [3] R. D. Rachmanta. (2024) Hari pencoblosan pemilu 2024 tiba, deretan hashtag ini trending di x. Accessed: 2024-12-21. [Online]. Available: <https://www.suara.com/teknologi/2024/02/14/103112/hari-pencoblosan-pemilu-2024-tiba-deretan-hashtag-ini-trending-di-x.html>
- [4] “Indonesia — x (twitter) trending topics and hashtags today,” 2024, accessed: 2024-12-21. [Online]. Available: <https://trends24.in/indonesia/>
- [5] D. A. Octavian, “Implementasi tax amnesty dapat memperkokoh ketahanan ekonomi,” 2017, accessed: 2024-12-21. [Online]. Available: https://www.lemhannas.go.id/images/Publikasi_Humas/Jurnal/Jurnal_Edisi_30_Juni_2017.pdf
- [6] C. P. Chai, “Comparison of text preprocessing methods,” *Natural Language Engineering*, vol. 29, no. 3, p. 509–553, 2023.
- [7] P. Sridevi and T. Velmurugan, “Impact of preprocessing on twitter based covid-19 vaccination text data by classification techniques,” in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAIC)*, 2022, pp. 1126–1132.
- [8] T. Declerck and P. Lendvai, “Processing and normalizing hashtags,” in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, R. Mitkov, G. Angelova, and K. Bontcheva, Eds. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sep. 2015, pp. 104–109. [Online]. Available: <https://aclanthology.org/R15-1015>
- [9] G. Laboreiro, L. Sarmento, J. Teixeira, and E. Oliveira, “Tokenizing micro-blogging messages using a text classification approach,” 10 2010, pp. 81–88.
- [10] M. A. Rosid, A. S. Fitriani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, “Improving text preprocessing for student complaint document classification using sastrawi,” *IOP Conference Series: Materials Science and Engineering*, vol. 874, no. 1, p. 012017, jun 2020. [Online]. Available: <https://dx.doi.org/10.1088/1757-899X/874/1/012017>
- [11] P. A. Telnoni, R. Budiawan, and M. Qana'a, “Comparison of machine learning classification method on text-based case in twitter,” in *2019 International Conference on ICT for Smart Society (ICISS)*, vol. 7, 2019, pp. 1–5.
- [12] J. Camacho-Collados and M. T. Pilehvar, “On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis,” 2018. [Online]. Available: <https://arxiv.org/abs/1707.01780>
- [13] @KAMRANK41723936, “The future of ai is here. #artificialintelligence #technology,” X (formerly Twitter) post, Dec. 2024, <https://x.com/KAMRANK41723936/status/1756932273109418122>.

- [14] L. A. Andini, "Indonesia sentris, wujud persatuan indonesia," Press release on the official website of the Provincial Government of Bangka Belitung, May 2024, accessed: 2024-12-27. [Online]. Available: https://babelprov.go.id/siaran_pers/indonesia-sentris-wujud-persatuan-indonesia
- [15] @ekowboy2, "Exploring the latest advancements in ai technology. #ai #innovation," X (formerly Twitter) post, Dec. 2024, <https://x.com/ekowboy2/status/1741796805996912649>.
- [16] S. K. Narayanasamy, Y.-C. Hu, S. M. Qaisar, and K. Srinivasan, "Effective preprocessing and normalization techniques for covid-19 twitter streams with pos tagging via lightweight hidden markov model," *Journal of Sensors*, vol. 2022, no. 1, p. 1222692, 2022. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/1222692>
- [17] M. A. Palomino and F. Aider, "Evaluating the effectiveness of text pre-processing in sentiment analysis," *Applied Sciences*, vol. 12, no. 17, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/17/8765>
- [18] S. Contributors, "Sastrawi: High quality stemmer library for indonesian language," GitHub repository, 2024, accessed: 2024-12-27. [Online]. Available: <https://github.com/sastrawi/sastrawi>



Vanny Ade Gunawan adalah seorang mahasiswa Teknik Informatika di Universitas Brawijaya yang lahir pada 4 Agustus 2006 di Malang. Ia memiliki minat yang mendalam pada dunia informatika dan terus mengexplorasi berbagai aspek di bidang tersebut. Di luar kesibukannya, Vanny sangat menyukai anjing dan menikmati waktu luangnya dengan menonton film serta mendengarkan musik, menunjukkan sisi kreatif dan santainya. Dalam proyek ini mengerjakan bagian analisis dan eksplorasi analisis data (EDA), menambahkan insight dari hasil analisis video presentasi final.

AUTHORS



Pieter Christy yan Yudhistira meraih penghargaan sebagai lulusan terbaik angkatan kelulusan tahun 2023 di SMAK Nasional Anglo. Saat ini, ia menempuh program studi Teknik Informatika Universitas Brawijaya dan berminat terjun ke sains data dan kecerdasaan artifisial. Pieter juga menunjukkan kemampuan akademik dan non akademik yang memuaskan sehingga berhasil menjadi penerima Beasiswa Unggulan Kemendikbud 2024.

Pieter bergabung sebagai anggota di komunitas Basic Computing Community departemen sains data dan sebagai media partner di divisi creative media. Selain studinya, Pieter tergabung di laboratorium Sistem Cerdas Fakultas Ilmu Komputer dan saat ini sedang melakukan riset mengenai Graph Neural Network. Dalam proyek ini mengerjakan bagian pengembangan dan pengevaluasian model, penataan notebook .ipynb, penyusunan isi paper dan mengarahkan jalannya video presentasi.



Christopher Robin Tanugroho adalah seorang mahasiswa Teknik Informatika Universitas Brawijaya. Ia lahir pada tanggal 10 November 2004 dengan semangat dan ambisi yang tinggi untuk menggapai cita-cita. Ia mengikuti komunitas yang dapat membantu Ia untuk menggapai cita-citanya selama di Fakultas Ilmu Komputer. Di luar kesibukannya, Christopher menikmati waktu luangnya dengan bermain game, membaca komik, mendengarkan musik, dan mendalami bidang yang ia minati. Dalam proyek ini mengerjakan bagian data understanding, menyampaikan pengantar dataset hingga preprocessing di presentasi final dan menyiapkan script video presentasi final yang kami semua gunakan.



Joshua Dwiputra Rendro Joelaskoro adalah seorang mahasiswa Teknik Informatika di Universitas Brawijaya. Lahir pada 17 Februari 2005, Joshua memiliki ambisi besar yang ia wujudkan melalui berbagai kesempatan yang datang kepadanya. Ia aktif berpartisipasi dalam berbagai bidang, khususnya teknologi dan bisnis, sebagai bagian dari upayanya untuk meraih cita-cita. Selain itu, Joshua juga terjun ke dunia consulting yang berfokus pada mendukung perkembangan bisnis yang lebih baik, mulai dari lingkungan universitas hingga tingkat nasional.

Dalam proyek ini mengerjakan bagian urgensi masalah, identifikasi masalah yang diangkat, menyunting penulisan kata dan kalimat dalam paper dan penyusunan PPT final yang ditayangkan di video presentasi.