Jurnal Pustaka Cendekia Hukum dan Ilmu Sosial
Volume 4, No. 2, Juni 2026 – September 2026, pp. 2905-2919
E-ISSN: 2987-4475
https://pchukumsosial.org/index.php/pchs

Analisis Sentimen terhadap Isu Childfree pada Generasi Milenial Akhir di Indonesia dan Korea Selatan Menggunakan SVM di Platform X

Clarissa Graziela Tjengdra[1*], Dewi Setiowati[2]

[1,2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Esa Unggul, Jl. Arjuna Utara No.9, RT.1/RW.2, Duri Kepa, Kec. Kb. Jeruk, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta Indonesia
clarissagtj@student.esaunggul.ac.id

Abstract

The childfree phenomenon has become a global trend that has sparked widespread discussion, especially among the late millennial generation. This study aims to analyze public sentiment regarding the childfree issue among the late millennial generation in Indonesia and South Korea using the Support Vector Machine (SVM) algorithm on platform X. Indonesian and Korean tweet data were collected from platform X, each with 1000 tweets. Then they went through the process of cleansing, word normalization, tokenization (Sastrawi for Indonesian and for Korean), stemming, and stopword removal. Sentiments were categorized into positive, negative, and neutral through manual or semi-automatic labeling. Then divided into 80% training data and 20% test data. Implemented with an SVM model with a focus on testing the RBF kernel. Model performance evaluation used accuracy, precision, recall, and F1-score metrics. It is hoped that the results of this study can provide insight in analyzing late millennial sentiment, especially in bilingual sentiment, as well as serve as a guide for demographic stakeholders.

Keywords: Sentiment Analysis; Support Vector Machine (SVM); X Platform; Multilingual; Childfree

Abstrak

Fenomena childfree telah menjadi tren global yang memicu diskusi luas, terutama di kalangan generasi milenial akhir. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat mengenai isu childfree pada generasi milenial akhir di Indonesia dan Korea Selatan menggunakan algoritma Support Vector Machine (SVM) di platform X. Data tweet berbahasa Indonesia dan Korea dikumpulkan dari platform X, masing-masing sebanyak 1000 tweet. Kemudian melewati proses cleansing, normalisasi kata, tokenisasi (Sastrawi untuk Bahasa Indonesia dan untuk Bahasa Korea), stemming, dan stopword removal. Sentimen dikategorikan menjadi positif, negatif, dan netral melalui pelabelan manual atau semi-otomatis. Kemudian dibagi menjadi 80% data latih dan 20% data uji. Diimplementasikan dengan model SVM dengan fokus pada pengujian kernel RBF. Evaluasi performa model menggunakan metrik accuracy, precision, recall, dan F1-score. Diharapkan hasil penelitian ini dapat memberikan pemahaman dalam menganalisis sentimen milenial akhir, khususnya dalam sentimen dua bahasa, serta menjadi panduan bagi pemangku kepentingan demografis.

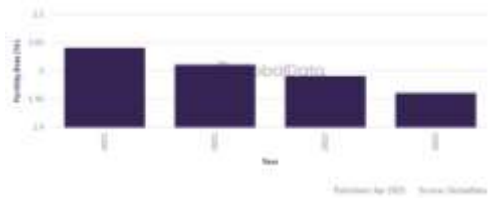
Kata Kunci: Analisis Sentimen; *Support Vector Machine* (SVM); Platform X; Multibahasa; *Childfree*.

✉Corresponding author: Clarissa Graziela Tjengdra
Email Address: clarissagtj@student.esaunggul.ac.id (Jl. Arjuna Utara No.9, RT.1/RW.2, Duri Kepa, Kec. Kb. Jeruk, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta Indonesia)
Received 20 July 2026, Accepted 26 July 2026, Published 01 August 2026

PENDAHULUAN

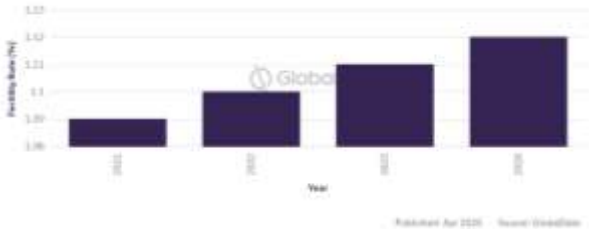
Fenomena *childfree* telah menjadi tren global dengan variasi penyebab di berbagai negara [7]. Di kalangan generasi milenial akhir (25-35 tahun), pilihan untuk tidak memiliki anak semakin banyak dibicarakan sebagai respons terhadap tantangan di zaman modern. Di Indonesia, isu *childfree* mulai mengemuka dan menjadi perbincangan publik setelah *influencer* Instagram Gita Savitri memutuskan untuk tidak memiliki anak melalui postingan Instagram Story-nya pada 14 Agustus 2021 [8]. Laporan BPS mengungkapkan 8,2% perempuan usia subur memilih *childfree*, dengan proporsi tertinggi pada kelompok urban berpendidikan tinggi [1]. Pilihan ini terutama dikaitkan dengan kesadaran otonomi tubuh, beban finansial, dan prioritas pengembangan karier [9]. Menariknya, diskusi di platform X

menunjukkan milenial akhir Indonesia cenderung memframing isu ini sebagai bagian dari hak reproduksi dan kebebasan personal [10]. Tingkat fertilitas total di Indonesia sendiri terus menurun, seperti yang bisa dilihat pada Gambar 1, dari 2,04 pada tahun 2021 menjadi 1,96 pada tahun 2024.



Gambar 1. Tingkat Fertilitas di Indonesia (2021 - 2024, %)
Sumber: GlobalData, 2025

Sementara itu, Korea Selatan menghadapi situasi yang lebih kompleks. Negara ini sempat menempati peringkat ke-198 dari 198 negara di seluruh dunia berdasarkan Dasbor Populasi Dunia Dana Penduduk Perserikatan Bangsa-Bangsa (UNFPA) pada tahun 2021, dengan tingkat fertilitas total Korea Selatan berada di angka 1,1. Menjadikannya negara dengan tingkat fertilitas terendah di dunia. Bagi milenial akhir Korea (25-35 tahun), keputusan ini merupakan respons terhadap hambatan struktural yang lebih sistemik: (1) harga properti mencapai 16 kali pendapatan tahunan rata-rata, (2) budaya kerja ekstrem (2.024 jam/tahun - tertinggi di OECD), dan (3) biaya pengasuhan anak yang mencapai Rp3,6 miliar per anak [5]. Ironisnya, meski pemerintah telah mengalokasikan Rp4.456 triliun untuk insentif *pronatalis*, survei Ipsos menemukan 62,2% wanita pekerja usia 25-45 tahun tetap mempertahankan pilihan *childfree* [2]. Namun, pada tahun 2024, tingkat kelahiran Korea Selatan mencatat kenaikan pertama kalinya dalam 9 tahun, menjadi 0,75 dari 0,72 pada tahun 2023 menurut data yang dirilis oleh kantor statistik nasional Korea Selatan [11]. Meskipun tingkat kesuburan tetap yang terendah di dunia, ini menandai peningkatan tahunan pertama sejak 2015.



Gambar 2. Tingkat Fertilitas di Korea (2021 - 2024, %)
Sumber: GlobalData, 2025

Dalam memahami fenomena ini, analisis sentimen media sosial menjadi krusial karena tiga alasan utama: (1) platform X menjadi ruang utama ekspresi generasi milenial akhir, (2) diskusi digital mencerminkan perubahan nilai sosial yang sedang terjadi, dan (3) memberikan data real-time tentang persepsi publik. Namun, penelitian sebelumnya memiliki keterbatasan metodologis yang signifikan. Studi komparatif lintas budaya masih jarang, dengan mayoritas penelitian terfokus pada analisis monolingual.

Beberapa penelitian sebelumnya telah menganalisis sentimen terkait isu *childfree* di media sosial X. Penelitian oleh Putra, Permana, & Afdal (2024) serta Nurhidayati, Umaidah, & Enri (2024)

menggunakan algoritma *Naïve Bayes Classifier* (NBC) dan SVM untuk mengevaluasi persepsi masyarakat terhadap gerakan bebas anak di media sosial X [4] [6]. Margono & Aprillia (2024) juga menganalisis sentimen untuk komentar yang mendukung tidak memiliki anak di aplikasi X menggunakan *Naive Bayes*. Namun, penelitian-penelitian ini umumnya terfokus pada satu bahasa atau belum secara eksplisit membandingkan sentimen lintas budaya dengan mempertimbangkan nuansa linguistik yang mendalam. Selain itu, akurasi model yang dihasilkan masih bervariasi, seperti akurasi 66,20% pada penelitian Margono & Aprillia (2024) [3].

Penelitian ini mengusung pendekatan inovatif dengan mengadaptasi metode SVM (*Support Vector Machine*) untuk mengatasi tantangan utama: (1) keterbatasan data akibat perubahan kebijakan API platform X, dan (2) kompleksitas analisis multibahasa. SVM dipilih karena kemampuannya yang terbukti dalam menangani dataset terbatas sekaligus mengidentifikasi pola linguistik spesifik budaya, seperti perbedaan penggunaan istilah kunci ("tekanan sosial" vs "경제적 부담") atau pola emoji yang khas.

Melalui pendekatan kuantitatif-eksplanatori, penelitian ini dirancang untuk menjawab tiga pertanyaan kunci: (1) Bagaimana mengintegrasikan dan memproses data multibahasa dari media sosial untuk analisis sentimen terkait isu *childfree* menggunakan metode SVM? (2) Bagaimana cara pengumpulan dan pembersihan data baik bahasa Korea maupun Indonesia dari platform X untuk analisis sentimen di kalangan generasi milenial akhir? (3) Seberapa efektif performa model SVM dalam mengidentifikasi perbedaan pola sentimen di antara generasi milenial akhir di Indonesia dan Korea Selatan? Jawaban atas pertanyaan-pertanyaan ini tidak hanya akan mengisi celah literatur, tetapi juga memberikan panduan praktis bagi pemangku kepentingan di tengah tantangan demografis yang semakin mengkhawatirkan.

METODE

Scraper Data

Penelitian ini akan menggunakan data tweet dari platform X yang membahas masalah *childfree*. Pengumpulan data akan dilakukan dengan mengambil tweet yang berbahasa Indonesia dan Korea menggunakan tweet-harvest, masing-masing sebanyak 1000 tweet dan akan disimpan dalam format *Comma Separated Values* (CSV). Teknik pengumpulan data ini menggunakan bahasa pemrograman Python. Rentang waktu yang dipilih adalah dari tahun 2021 dan 2024, hal ini dikarenakan pada rentang waktu tersebut, mencakup periode awal peningkatan diskusi *childfree* di media sosial Indonesia (pada Agustus 2021) serta titik terendah tingkat kelahiran di Korea Selatan pada tahun 2021, yang kemudian dapat dibandingkan dengan kondisi terkini pada tahun 2024 di mana tingkat fertilitas Indonesia menunjukkan penurunan berkelanjutan menjadi 1,96 dari 2,04 pada tahun 2021, dan Korea Selatan mencatat kenaikan pertama dalam 9 tahun menjadi 0,75 dari 0,72 pada tahun 2023.

Pre-processing Data

Pada tahap ini, peneliti melakukan pengolahan data sehingga data mentah tadi dapat digunakan dalam pengujian data [12]. Proses *preprocessing* meliputi *cleaning*, normalisasi kata, tokenisasi, *stemming*, dan *stopword removal*. Tahapan-tahapan ini bertujuan untuk mereduksi kompleksitas bahasa alami, menghilangkan *noise*, dan menyeragamkan teks. Berikut detail tahapan yang dilakukan:

Cleansing

Tahap awal ini akan membersihkan data dari karakter yang tidak penting, seperti URL, *hashtag*, *mention username*, angka, simbol, dan *emoticon* yang tidak dibutuhkan dalam analisis sentimen utama.

Normalisasi kata

Tahap ini akan mengubah kata-kata tidak baku atau bahasa *slang* menjadi bentuk baku yang sesuai dengan kaidah bahasa masing-masing.

Tokenisasi

Proses ini akan membagi teks menjadi token atau kata-kata kecil [13]. Karena melibatkan dua bahasa berbeda:.

1. Untuk Bahasa Indonesia, akan digunakan pustaka Sastrawi. Sastrawi efektif dalam memisahkan kata berdasarkan spasi dan tanda baca, serta mendukung proses stemming.
2. Untuk Bahasa Korea, akan digunakan MeCab, sebuah alat tokenisasi berbasis analisis morfologis. MeCab efektif dalam menguraikan morfem-morfem kompleks pada bahasa Korea.

Stemming

Pada tahap ini, kata-kata berimbuhan akan kembali ke bentuk aslinya (*stem*) atau bentuk leksikal dasarnya (*lemma*). Untuk Bahasa Indonesia, Sastrawi akan membantu proses *stemming*. Untuk Bahasa Korea, hasil tokenisasi MeCab yang berbasis morfem juga akan membantu dalam mendapatkan bentuk dasar kata. Tahap ini penting untuk mengurangi variasi kata dan meningkatkan efisiensi model.

Stopword Removal

Kata-kata yang umum dan tidak memiliki makna signifikan dalam konteks sentimen (misalnya, "yang", "dan", "di") akan dihapus dari teks. Proses ini akan menggunakan daftar *stopword* yang relevan untuk Bahasa Indonesia dan Bahasa Korea.

Labeling Sentimen

Setelah tahapan pra-pemrosesan, data teks akan melalui proses pelabelan sentimen. Tahap ini bertujuan untuk mengategorikan setiap tweet ke dalam kelas sentimen: Positif, Negatif, atau Netral. Pelabelan akan dilakukan secara manual oleh satu annotator yang memiliki pemahaman terhadap konteks isu childfree serta nuansa linguistik dari Bahasa Indonesia dan Bahasa Korea. Sentimen-sentimen tersebut akan dibagi menjadi:

1. Positif: Mengindikasikan opini yang mendukung atau setuju dengan pilihan childfree.
2. Negatif: Mengindikasikan opini yang menolak atau tidak setuju dengan pilihan childfree.
3. Netral: Mengindikasikan opini yang tidak menunjukkan kecenderungan positif atau negatif, atau bersifat informatif/faktual tanpa menyatakan pendapat.

Klasifikasi Sentimen Menggunakan SVM

Data yang telah direpresentasikan dalam bentuk vektor akan digunakan sebagai input dalam proses klasifikasi sentimen menggunakan algoritma *Support Vector Machine* (SVM). SVM adalah metode pembelajaran mesin yang efektif untuk klasifikasi, bekerja dengan mencari *hyperplane* terbaik yang dapat memisahkan kelas-kelas data secara optimal. Implementasi SVM dalam penelitian ini akan dilakukan menggunakan bahasa pemrograman Python, memanfaatkan pustaka *machine learning* seperti *scikit-learn*. Berikut detail implementasi SVM:

Pembagian Data

Dataset yang sudah bersih dan direpresentasikan akan dibagi menjadi data latih (*training data*) dan data uji (*testing data*). Pembagian dalam penelitian ini dibagi menjadi 80% data latih dan 20% data uji untuk memastikan bahwa model dapat belajar dari data yang sebelumnya dan menguji data yang baru.

Implementasi SVM

Model SVM akan dilatih menggunakan data latih dan kemudian digunakan untuk memprediksi sentimen pada data uji. Proses ini akan difasilitasi oleh modul SVM yang tersedia dalam pustaka *scikit-learn* di Python.

Kernel SVM

Mengingat kompleksitas data teks dari media sosial dan sifat multibahasa, fungsi kernel RBF (*Radial Basis Function*) akan digunakan karena kemampuannya dalam menangani hubungan non-linear antar data dan memetakan fitur ke ruang dimensi yang lebih besar. Python memungkinkan fleksibilitas dalam menguji berbagai *kernel* melalui parameter yang disediakan oleh pustaka *scikit-learn*. Persamaan yang digunakan adalah

$$K(x,y) = e^{-\gamma||x-y||^2} \quad (1)$$

Evaluasi Model

Setelah proses klasifikasi, performa model SVM akan dievaluasi untuk mengukur seberapa efektif model dalam mengklasifikasikan sentimen [14]. Metrik yang akan digunakan meliputi:

Accuracy

Mengukur persentase total prediksi yang benar dari seluruh data, dihitung dengan persamaan

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

dimana TP adalah true positive, TN adalah true negative, FP adalah false positive, dan FN adalah false negative.

Precision

Menghitung jumlah prediksi kelas positif yang benar-benar positif dengan menggunakan rumus yang ditemukan dalam persamaan

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall

Mengukur seberapa banyak instans kelas positif yang berhasil dideteksi oleh model, rumusnya

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-Score

Merupakan nilai harmonik rata-rata dari *precision* dan *recall*, sangat berguna ketika terdapat ketidakseimbangan data antar kelas, yang mungkin terjadi pada data sentimen, dengan rumus yang ditunjukkan pada persamaan

$$F1\text{-}Score = \frac{2xTP}{2xTP + FP + FN} \quad (5)$$

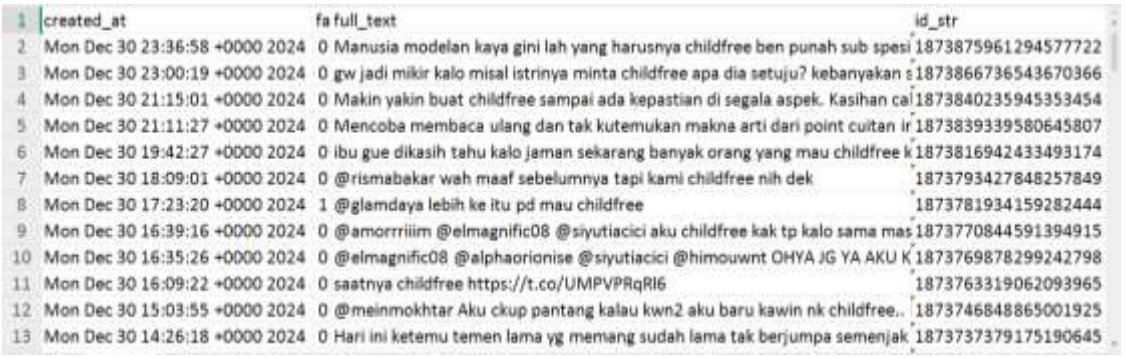
Untuk *Confusion matrix* sendiri, adalah tabel ringkasan yang digunakan untuk menilai kinerja model klasifikasi, khususnya dalam masalah klasifikasi dengan dua kelas (biner) maupun lebih seperti pada penelitian ini (positif, negatif, dan netral). Tabel ini membandingkan hasil prediksi model dengan nilai data yang diuji yang sebenarnya. Jadi, *Confusion matrix* untuk menampilkan nilai TP, TN, FP, dan FN dalam bentuk tabel visual.

HASIL DAN DISKUSI

Hasil

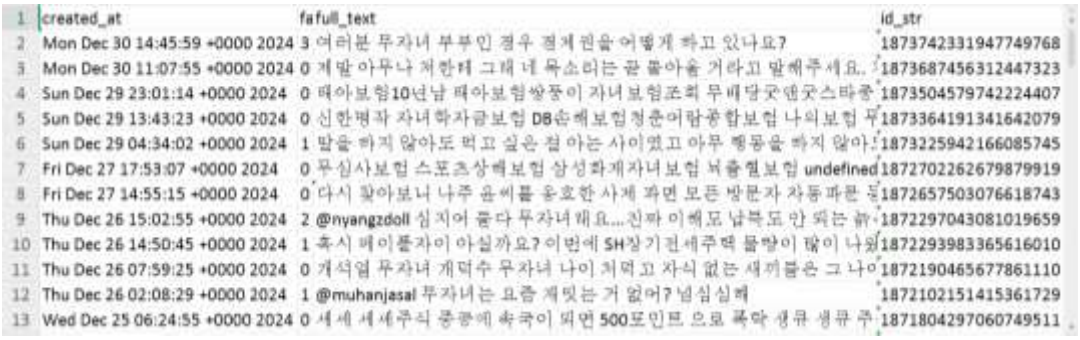
Crawling Data

Data yang berhasil dikumpulkan memperoleh tweet yang relevan mengenai isu childfree dari platform media sosial X. Proses ini diimplementasikan menggunakan bahasa pemrograman Python di lingkungan Google Colab, menggunakan pustaka tweet-harvest@2.6.1. Pengumpulan data dilakukan berdasarkan kata kunci spesifik: “childfree” untuk Bahasa Indonesia, dan “무자녀” atau “딩크족” untuk Bahasa Korea, mencakup dari tahun 2021 dan 2024. Dari proses *crawling* ini, berhasil dikumpulkan total 1502 *tweet* Bahasa Indonesia (gabungan dari tahun 2021 dan 2024) dan total 2063 *tweet* Bahasa Korea (gabungan dari tahun 2021 dan 2024). Dengan demikian, total keseluruhan data yang terkumpul dan akan diproses dalam penelitian ini adalah 3565 *tweet*. Seluruh data mentah yang diperoleh disimpan dalam format CSV (*Comma Separated Values*), dan contoh hasil *crawling* dapat dilihat pada Gambar 3 dan Gambar 4.



	created_at	fa	full_text	id_str
2	Mon Dec 30 23:36:58 +0000 2024	0	Manusia modelan kaya gini lah yang harusnya childfree ben punah sub spesi	1873875961294577722
3	Mon Dec 30 23:00:19 +0000 2024	0	gw jadi mikir kalo misal istrinya minta childfree apa dia setuju? kebanyakan s	1873866736543670366
4	Mon Dec 30 21:15:01 +0000 2024	0	Makin yakin buat childfree sampai ada kepastian di segala aspek. Kasihan cal	1873840235945353454
5	Mon Dec 30 21:11:27 +0000 2024	0	Mencoba membaca ulang dan tak kutemukan makna arti dari point cuitan ir	1873839339580645807
6	Mon Dec 30 19:42:27 +0000 2024	0	ibu gue dikasih tahu kalo jaman sekarang banyak orang yang mau childfree k	1873816942433493174
7	Mon Dec 30 18:09:01 +0000 2024	0	@rismabakar wah maaf sebelumnya tapi kami childfree nih dek	1873793427848257849
8	Mon Dec 30 17:23:20 +0000 2024	1	@glamdaya lebih ke itu pd mau childfree	1873781934159282444
9	Mon Dec 30 16:39:16 +0000 2024	0	@amorrriiim @elmagnific08 @siyutiacici aku childfree kak tp kalo sama mas	1873770844591394915
10	Mon Dec 30 16:35:26 +0000 2024	0	@elmagnific08 @alphaorionise @siyutiacici @himouwnt OHYA JG YA AKU K	1873769878299242798
11	Mon Dec 30 16:09:22 +0000 2024	0	saatnya childfree https://t.co/UMPVPRqRI6	1873763319062093965
12	Mon Dec 30 15:03:55 +0000 2024	0	@meinmokhtar Aku ckup pantang kalau kwn2 aku baru kawin nk childfree..	1873746848865001925
13	Mon Dec 30 14:26:18 +0000 2024	0	Hari ini ketemu temen lama yg memang sudah lama tak berjumpa semenjak	1873737379175190645

Gambar 3. Contoh Hasil *Crawling* Data Mentah (Bahasa Indonesia)



	created_at	fa	full_text	id_str
2	Mon Dec 30 14:45:59 +0000 2024	3	여러분 무자녀 부부인 경우 절세권을 어떻게 하고 있나요?	1873742331947749768
3	Mon Dec 30 11:07:55 +0000 2024	0	제발 아무나 저한테 그래 네 육소리는 곧 돌아올 거라고 말해주세요.	1873687456312447323
4	Sun Dec 29 23:01:14 +0000 2024	0	태아보험10년납 태아보험쌍둥이 자녀보험조회 무배당굿앤굿스타종	1873504579742224407
5	Sun Dec 29 13:43:23 +0000 2024	0	신한평생 자녀학자금보험 DB손해보험청춘어람종합보험 나의보험 무	1873364191341642079
6	Sun Dec 29 04:34:02 +0000 2024	1	말을 하지 않아도 먹고 싶은 점 아는 사이였고 아무 행동을 하지 않아	1873225942166085745
7	Fri Dec 27 17:53:07 +0000 2024	0	무심사보험 스포츠상해보험 삼성화재자녀보험 뇌출혈보험 undefined	1872702262679879919
8	Fri Dec 27 14:55:15 +0000 2024	0	다시 찾아보니 나주 윤씨를 용호한 사제 과면 보든 방문자 자동파문 등	1872657503076618743
9	Thu Dec 26 15:02:55 +0000 2024	2	@nyangzdoll 심지어 둘다 무자녀래요....진짜 이해도 남북도 안 되는 놈	1872297043081019659
10	Thu Dec 26 14:50:45 +0000 2024	1	혹시 페이퍼자이 아실까요? 이번에 SH장기전세주택 물량이 많이 나왔	1872293983365616010
11	Thu Dec 26 07:59:25 +0000 2024	0	개석열 무자녀 개딱수 무자녀 나이 처먹고 자식 없는 새끼들은 그 나이	1872190465677861110
12	Thu Dec 26 02:08:29 +0000 2024	1	@muhanjasal 무자녀는 요즘 제팃는 거 없어? 넘심심해	1872102151415361729
13	Wed Dec 25 06:24:55 +0000 2024	0	세세세세주식 종광에 속국이 되면 500포인트 으로 폭락 생큐 생큐 주	1871804297060749511

Gambar 4. Contoh Hasil *Crawling* Data Mentah (Bahasa Korea)

Pre-processing Data

Pada tahapan ini, dataset yang telah dilakukan crawling akan memasuki tahap pengolahan data. Tahapan ini bertujuan agar data mentah yang sudah didapat sebelumnya, menjadi format yang lebih bersih, terstruktur, dan sesuai untuk analisis lebih lanjut. Berikut detail tahapan yang dilakukan:

Cleansing

Tahap ini merupakan pembersihan awal yang menghilangkan elemen-elemen tidak relevan seperti URL, *mention*, *hashtag*, serta menyeragamkan penggunaan huruf.

1. Untuk Bahasa Indonesia: Angka akan dihapus dan simbol non-huruf/spasi diganti dengan spasi, lalu spasi berlebih dirapikan.
2. Untuk Bahasa Korea: Angka dan simbol (seperti tanda baca) tidak dihapus pada tahap ini, melainkan dipertahankan agar dapat ditangani secara spesifik oleh tokenizer MeCab. Hanya spasi berlebih yang dirapikan.
3. Seluruh teks kemudian diubah menjadi huruf kecil (case folding).

Pada Gambar 5 dan Gambar 6, hasil Cleansing & Case Folding menunjukkan teks asli dan hasil setelah tahap pembersihan dan penyeragaman huruf. Perhatikan bagaimana URL, @mention, #hashtag dihapus, serta angka/simbol untuk Bahasa Indonesia, sementara untuk Bahasa Korea angka dan simbol tetap dipertahankan.

Normalisasi kata

Setelah cleansing dan case folding, tahap normalisasi kata dilakukan untuk mengubah kata-kata slang atau tidak baku menjadi bentuk baku yang sesuai dengan aturan bahasa masing-masing. Proses

ini memanfaatkan kamus normalisasi kustom yang telah dikembangkan secara spesifik dari dataset penelitian ini.

Pada Gambar 5 dan Gambar 6, hasil Normalisasi Kata memperlihatkan perubahan kata-kata tidak baku atau slang (seperti "kayak" menjadi "seperti" dalam Bahasa Indonesia, atau "namchin" menjadi "namja chingu" dalam Bahasa Korea) menjadi bentuk standarnya.

Tokenisasi

Proses tokenisasi sendiri adalah pembagian teks menjadi bagian yang lebih kecil, seperti token atau kata-kata, yang merupakan langkah fundamental dalam NLP.

1. Untuk Bahasa Indonesia: Tokenisasi dilakukan dengan pemisahan sederhana berdasarkan spasi. Hasilnya adalah daftar kata (list of strings).
2. Untuk Bahasa Korea: Digunakan MeCab sebagai tokenizer morfologis. MeCab memecah kalimat menjadi morfem-morfem (unit makna terkecil) dan mengidentifikasi Part-of-Speech (POS) Tag untuk setiap morfem, meskipun untuk proses selanjutnya hanya morfemnya saja yang digunakan.

Pada Gambar 5 dan Gambar 6, hasil Tokenisasi menggambarkan bagaimana kalimat dipecah menjadi unit-unit kata. Perhatikan format list untuk Bahasa Indonesia, dan format morfem/POS tag untuk Bahasa Korea.

Stopword Removal

Tahap ini berfungsi menghilangkan kata-kata umum dan tidak signifikan dalam konteks sentimen dari teks. Ini membantu mengurangi dimensi data dan fokus pada kata-kata yang lebih informatif.

1. Untuk Bahasa Indonesia: Proses ini menggunakan stopword remover dari pustaka Sastrawi.
2. Untuk Bahasa Korea: Kata-kata yang ada dalam daftar stopword kustom (yang sudah didefinisikan) akan dihapus dari morfem-morfem yang dihasilkan MeCab.

Pada Gambar 5 dan Gambar 6, hasil Stopword Removal menampilkan teks setelah kata-kata umum dan tidak informatif dihilangkan. Perhatikan bagaimana beberapa kata dihilangkan, sementara angka dan simbol untuk Bahasa Korea tetap dipertahankan.

Stemming

Ini adalah tahap akhir pra-pemrosesan teks. Kata-kata berimbuhan akan dikembalikan ke bentuk kata dasarnya (stem) atau bentuk leksikal dasarnya (lemma), untuk mengurangi variasi kata dan meningkatkan efisiensi model.

1. Untuk Bahasa Indonesia: Proses stemming dilakukan menggunakan Sastrawi.
2. Untuk Bahasa Korea: Analisis morfologis MeCab yang sudah menghasilkan morfem dasar sangat membantu dalam mendapatkan bentuk dasar kata. Pada tahap ini, morfem-morfem dasar yang telah disaring akan digabungkan menjadi string final.

Pada Gambar 5 dan Gambar 6, hasil Stemming / Lemmatisasi menggambarkan bagaimana kata-kata berimbuhan dikembalikan ke bentuk dasarnya. Ini adalah format akhir teks yang akan digunakan sebagai input untuk model BERT. Berikut Gambar 5 dan Gambar 6 yang memperlihatkan hasil akhir setelah setiap tahap pre-processing dilakukan

Gambar 5. Hasil Tahap Pre-processing Data (Bahasa Indonesia)

Gambar 6. Hasil Tahap Pre-processing Data (Bahasa Korea)

Pengujian Data

Pada saat ini, kumpulan data yang telah melalui pra-pemrosesan lengkap dan representasi fitur akan dibagi menjadi data latih (training data) dan data uji (testing data). Pembagian ini esensial untuk memberikan pelatihan kepada model dan mengevaluasi bagaimana ia berfungsi dengan data yang belum pernah dia lihat sebelumnya.

Dari total 2000 tweet relevan yang berhasil dikonsolidasi, data dibagi dengan rasio 80% untuk data latih dan 20% untuk data uji, serta dilakukan stratifikasi berdasarkan label sentimen untuk memastikan distribusi kelas yang proporsional di kedua set data. Output dari pembagian data adalah sebagai berikut:

A. Jumlah total data setelah seleksi final: 2000 tweet
B. Jumlah data latih (80%): 1600 tweet
C. Jumlah data uji (20%): 400 tweet

Distribusi sentimen pada data latih dan data uji menunjukkan proporsi yang sangat mirip, mengindikasikan bahwa proses stratifikasi berjalan dengan efektif:

1. Distribusi label di data latih (persentase):

A. Netral: 0.418125
B. Positif: 0.404375
C. Negatif: 0.177500

2. Distribusi label di data uji (persentase):
   A. Netral: 0.4175
   B. Positif: 0.4050
   C. Negatif: 0.1775

Penanganan Ketidakseimbangan Kelas

Meskipun pembagian data dilakukan secara proporsional, dataset ini memiliki ketidakseimbangan kelas (class imbalance) yang signifikan, di mana kelas negatif memiliki jumlah sampel yang lebih sedikit dibandingkan kelas positif dan netral. Untuk mengatasi masalah ini dan memastikan model dapat belajar secara efektif dari kelas minoritas, teknik oversampling SMOTE (Synthetic Minority Over-sampling Technique) diterapkan pada data latih (x_train, y_train).

Dengan bantuan SMOTE, sampel sintetis baru dibuat untuk kelas minoritas, sehingga menyeimbangkan distribusi jumlah sampel di antara semua kelas sentimen [15]. Proses ini membantu mengurangi bias model terhadap kelas mayoritas dan meningkatkan kemampuan model dalam mengidentifikasi serta memprediksi sentimen kelas minoritas secara lebih akurat. Berikut adalah distribusi kelas setelah aplikasi SMOTE:

1. Sebelum SMOTE:
   A. Kelas netral memiliki 669 sampel.
   B. Kelas positif memiliki 647 sampel.
   C. Kelas negatif adalah kelas minoritas yang paling sedikit, hanya memiliki 284 sampel.
2. Setelah SMOTE:
   A. SMOTE telah membuat sampel sintetis baru untuk kelas negatif dan positif.
   B. Jumlah sampel untuk ketiga kelas kini semuanya seimbang, yaitu masing-masing 669 sampel (positif: 669, netral: 669, negatif: 669).

Berdasarkan output yang diberikan, SMOTE berhasil menyeimbangkan kelas-kelas sentimen pada data latih dengan sangat baik. SMOTE telah berhasil meningkatkan jumlah representasi kelas minoritas dalam data latih, memastikan model memiliki jumlah contoh yang setara dari setiap kategori sentimen untuk dipelajari. Hal ini membantu mencegah model bias terhadap kelas mayoritas dan meningkatkan kemampuannya dalam mengidentifikasi sentimen negatif dan positif secara lebih efektif.

Diskusi

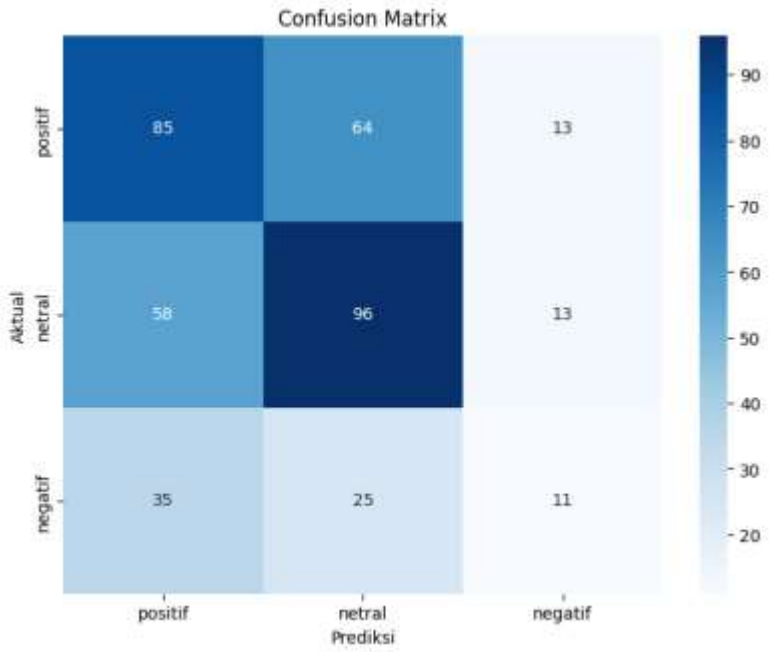
Analisis Hasil

Setelah tahap klasifikasi sentimen menggunakan Support Vector Machine (SVM) dan proses optimasi hyperparameter serta penanganan ketidakseimbangan kelas, model dievaluasi untuk mengukur performanya dalam mengklasifikasikan sentimen pada data uji yang belum pernah dilihat. Evaluasi ini

menggunakan metrik accuracy, precision, recall, dan F1-Score. Output dari evaluasi model adalah sebagai berikut:

```
--- SEL 12: Evaluasi Model ---

--- Hasil Evaluasi Keseluruhan Model ---
Accuracy: 0.4800
Precision: 0.4628
Recall: 0.4800
F1-Score: 0.4664

--- Laporan Klasifikasi Rinci per Kelas (Keseluruhan) ---
              precision    recall  f1-score   support

     negatif       0.30      0.15      0.20        71
      netral       0.52      0.57      0.55       167
     positif       0.48      0.52      0.50       162

    accuracy                           0.48       400
   macro avg       0.43      0.42      0.42       400
weighted avg       0.46      0.48      0.47       400
```

Gambar 7. Hasil Evaluasi Keseluruhan Model



Gambar 8. Confusion Matrix Keseluruhan Model

Model SVM yang dikembangkan menunjukkan performa dengan Akurasi keseluruhan sebesar 0.4800 dan F1-Score (weighted average) sebesar 0.4664. Meskipun angka akurasi keseluruhan ini mungkin belum setinggi yang diinginkan (seperti di atas 70%), hasil ini perlu diinterpretasikan dalam konteks kompleksitas dataset dan fenomena childfree di media sosial.

Peningkatan signifikan pada performa model, terutama pada F1-Score (macro average) selama cross-validation hyperparameter tuning (mencapai 0.6845), menunjukkan bahwa strategi penanganan ketidakseimbangan kelas (dengan SMOTE dan class_weight='balanced' pada SVM) serta optimasi hyperparameter terbukti efektif dalam membantu model belajar secara lebih seimbang dari semua kelas sentimen. Secara rinci, performa model per kelas adalah sebagai berikut:

1. Kelas Positif: Menunjukkan peningkatan recall yang drastis dibandingkan hasil awal, mencapai 0.52. Ini menunjukkan bahwa model sekarang lebih mampu menemukan tweet dengan sentimen positif.
2. Kelas Netral: Menunjukkan performa yang relatif baik, dengan F1-Score sebesar 0.55. Kelas ini cenderung lebih mudah diidentifikasi karena sifat informatifnya yang lebih lugas.
3. Kelas Negatif: Meskipun telah diupayakan dengan penanganan ketidakseimbangan kelas, kelas negatif masih menjadi tantangan terbesar bagi model, dengan recall sebesar 0.15 dan F1-Score sebesar 0.20. Hal ini mungkin disebabkan oleh beberapa faktor, seperti jumlah sampel asli kelas negatif yang paling sedikit, ambiguitas intrinsik dalam tweet yang berekspresi negatif, atau keragaman ekspresi negatif yang membuat polanya sulit ditangkap model.

Analisis Sentimen Komparatif (Per Negara & Tahun)

Analisis sentimen komparatif per negara dan tahun memberikan wawasan lebih lanjut tentang pola diskusi isu childfree di Platform X. Perbandingan distribusi sentimen antara Bahasa Indonesia dan Bahasa Korea, serta antara tahun 2021 dan 2024, disajikan di bawah ini

```
===== Tahun 2021 =====

--- ID (500 tweet relevan) ---
sentiment
negatif    27.20%
netral     39.20%
positif    33.60%
Name: proportion, dtype: object

--- KO (500 tweet relevan) ---
sentiment
negatif     8.20%
netral     48.20%
positif    43.60%
Name: proportion, dtype: object
```

```
===== Tahun 2024 =====

--- ID (500 tweet relevan) ---
sentiment
negatif    25.80%
netral     23.00%
positif    51.20%
Name: proportion, dtype: object

--- KO (500 tweet relevan) ---
sentiment
negatif     9.80%
netral     56.80%
positif    33.40%
Name: proportion, dtype: object
```

Gambar 9. Hasil Komparatif Distribusi Sentimen per Negara dan Tahun

Berdasarkan distribusi sentimen di atas, dapat ditarik beberapa perbandingan dan tren yang signifikan mengenai isu childfree di Indonesia dan Korea Selatan pada rentang waktu 2021 dan 2024.

1. Perbandingan Tren di Indonesia: Sentimen positif di Indonesia cenderung meningkat dari 33.60% pada tahun 2021 menjadi 51.20% pada tahun 2024. Seiring dengan itu, sentimen negatif menunjukkan sedikit penurunan dari 27.20% menjadi 25.80%, dan sentimen netral juga menurun signifikan dari 39.20% menjadi 23.00%.
2. Perbandingan Tren di Korea Selatan: Berbeda dengan Indonesia, sentimen positif di Korea Selatan cenderung menurun dari 43.60% pada tahun 2021 menjadi 33.40% pada tahun 2024. Di sisi lain, sentimen netral justru meningkat signifikan dari 48.20% di 2021 menjadi 56.80% di 2024, sementara sentimen negatif juga cenderung sedikit meningkat dari 8.20% menjadi 9.80%.

3. Perbandingan Sentimen Positif Antar Negara (Tahun 2021 dan 2024): Di Indonesia, sentimen positif mengalami peningkatan signifikan dari 33.60% pada tahun 2021 menjadi 51.20% pada tahun 2024. Peningkatan ini mengindikasikan semakin besarnya dukungan dan penerimaan publik terhadap pilihan childfree, selaras dengan diskusi di Platform X yang cenderung membingkai isu ini sebagai bagian dari hak reproduksi dan kebebasan personal di kalangan milenial akhir. Sebaliknya, di Korea Selatan, sentimen positif menunjukkan penurunan dari 43.60% pada tahun 2021 menjadi 33.40% pada tahun 2024. Penurunan ini dapat merefleksikan bahwa meskipun pilihan childfree semakin banyak, hal tersebut mungkin lebih merupakan respons pragmatis terhadap hambatan struktural yang sistemik, seperti tingginya biaya properti, budaya kerja ekstrem, dan biaya pengasuhan anak yang mahal, daripada pilihan gaya hidup yang sepenuhnya positif. Dengan demikian, terlihat adanya perbedaan tren sentimen positif yang jelas antara kedua negara dalam periode yang sama, mencerminkan latar belakang budaya dan tekanan demografis yang berbeda.

KESIMPULAN

Penelitian ini berhasil menggunakan pipeline analisis sentimen terhadap isu childfree pada generasi milenial akhir di Indonesia dan Korea Selatan menggunakan algoritma Support Vector Machine (SVM). Tahapan pra-pemrosesan teks multibahasa telah berhasil diterapkan secara spesifik untuk Bahasa Indonesia (menggunakan text.split() untuk tokenisasi, Sastrawi untuk stopword removal dan stemming) dan Bahasa Korea (menggunakan MeCab untuk tokenisasi morfologis, serta filter stopword dan lemmatisasi berbasis MeCab). Representasi fitur teks dilakukan menggunakan embedding kontekstual dari model BERT (bert-base-multilingual-cased).

Model SVM yang dilatih dengan data yang telah diseimbangkan oleh SMOTE dan parameter class_weight='balanced' menunjukkan Akurasi keseluruhan sebesar 0.4800 dan F1-Score (weighted average) sebesar 0.4664. Proses Hyperparameter Tuning menggunakan GridSearchCV berhasil mengoptimalkan kinerja model. Performa model menunjukkan kemampuan yang lebih baik dalam mengklasifikasikan sentimen pada kelas positif dan netral dibandingkan kelas negatif yang masih menjadi tantangan.

Analisis sentimen komparatif memberikan wawasan tentang pola sentimen yang berbeda antara Indonesia dan Korea Selatan terkait isu childfree pada tahun 2021 dan 2024. Hasil analisis sentimen positif antara Indonesia dan Korea Selatan pada periode ini menunjukkan tren yang berbanding terbalik. Didapatkan analisis sentimen positif di Indonesia dari tahun 2021 (33.60%) dan 2024 (51.20%) meningkat sebanyak 17.60%. Sementara itu, di Korea Selatan, sentimen positif terhadap childfree mengalami penurunan dari 2021 (43.60%) dan 2024 (33.40%) sebanyak 10.20%. Ini mengindikasikan perbedaan fundamental dalam penerimaan dan perspektif terhadap fenomena childfree di kedua negara seiring berjalannya waktu.

Berdasarkan temuan dan batasan penelitian ini, berikut adalah rekomendasi untuk pengembangan lebih lanjut:

1. Perluasan Dataset: Untuk meningkatkan performa model secara signifikan, disarankan untuk mengumpulkan lebih banyak data berlabel, terutama untuk kelas minoritas, yang bisa memberikan lebih banyak contoh bagi model untuk belajar.
2. Peningkatan Kualitas Anotasi: Mengingat pelabelan dilakukan oleh satu annotator, penelitian selanjutnya disarankan dapat melibatkan beberapa annotator dan mengukur Inter-Annotator Agreement (IAA) untuk meningkatkan konsistensi dan objektivitas label sentimen.
3. Pengembangan Pra-pemrosesan Teks: Kamus normalisasi kata (baik Bahasa Indonesia maupun Korea) dapat terus diperluas dengan lebih banyak slang, singkatan, dan typos yang ditemukan dalam dataset media sosial untuk meningkatkan kualitas fitur teks.
4. Eksplorasi model dan teknik lanjutan dapat mencakup beberapa pendekatan, di antaranya adalah mencoba teknik oversampling lain seperti ADASYN atau metode ensemble untuk imbalanced learning (misalnya EasyEnsemble, BalancedBagging). Selain itu, penting juga untuk mengeksplorasi model deep learning yang lebih canggih, seperti melakukan fine-tuning pada model BERT secara langsung, yang mungkin dapat menangkap nuansa sentimen yang lebih kompleks. Penambahan fitur-fitur non-teks, seperti metadata tweet atau fitur berbasis emoticon khusus, juga dapat melengkapi word embedding untuk meningkatkan akurasi klasifikasi.

REFERENSI

[1] E. Fitriana, “Hubungan Antara Persepsi dan Sikap Mahasiswa dengan Fenomena Perilaku Childfree dalam Perspektif Agama Islam (Studi Pada Mahasiswa S1 FIAI UII Angkatan 2019-2023),” Disertasi Doktoral, Universitas Islam Indonesia, Yogyakarta, Indonesia, 2024.

[2] A. Rismawati, “Fenomena Childfree Sebagai Prinsip Hidup Perspektif Hukum Islam (Studi Kasus Wanita Karir Permodalan Nasional Madani Jakarta),” Disertasi Doktoral, UNUSIA, Jakarta, Indonesia, 2023.

[3] H. Margono and F. N. Aprillia, “Analisis Sentimen Komentar Childfree di Aplikasi X Menggunakan Naïve Bayes,” The Indonesian Journal of Computer Science, vol. 13, no. 3, 2024.

[4] L. Nurhidayati, Y. Umaidah, and U. Enri, “Analisis Sentimen Isu Childfree Di Media Sosial Twitter Menggunakan Algoritma Support Vector Machine,” Jurnal Ilmiah Wahana Pendidikan, vol. 10, no. 4, pp. 422-430, 2024.

[5] O. M. Prasetya, “Analisis Peran Budaya Jeong dalam Menerapkan Kepemimpinan di PT. Korea Tomorrow dan Global Indonesia,” Disertasi Doktoral, Universitas Kristen Indonesia, Jakarta, Indonesia, 2024.

[6] M. A. S. Putra, I. Permana, and M. Afdal, “Analisis Sentimen Masyarakat Mengenai Gerakan Childfree di Media Sosial X Menggunakan Algoritma NBC dan SVM: Sentiment Analysis of Childfree Campaign on X Social Media Using NBC and SVM Algorithms,” MALCOM:

Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 4, pp. 1189-1198, 2024.

[7] A. Zuhdiantito, "Fenomena Childfree di Kalangan Pasangan Suami Istri Perspektif Maqashid Syariah dan Hak Reproduksi Perempuan (Studi Kasus Pada Generasi Milenial dan Generasi Z Kabupaten Sleman)," Disertasi Doktoral, Universitas Islam Indonesia, Sleman, Indonesia, 2023.

[8] H. Sadiah and M. Amelia, "Konstrusi Sosial Fenomena Child Free di Sosial Media Instagram (Studi pada Kasus Gita Savitri)," Kompetensi, vol. 17, no. 1, pp. 1-10, 2024.

[9] A. C. Hasibuan, "KEPUTUSAN MENIKAH DENGAN PILIHAN TANPA MEMILIKI ANAK (CHILDFREE) DALAM PERSPEKTIF HUKUM ISLAM," Disertasi Doktoral, Universitas Malikussaleh, [Kota Universitas Malikussaleh, jika diketahui], Indonesia, 2024.

[10] A. Pebriansah, "Childfree Dalam Konteks Hak Asasi Manusia: Tantangan Dan Perlindungan Serta Pencapaian Hak-Hak Individu," Jurnal Darussalam: Jurnal Pendidikan, Komunikasi Dan Pemikiran Hukum Islam, vol. 16, no. 1, pp. 194-218, 2024.

[11] L. Yonghwa and F. Ananda, "Faktor Sosial di Balik Rendahnya Angka Kelahiran Di Korea Selatan," in *SEMINAR NASIONAL LPPM UMMAT*, vol. 3, pp. 1060-1069, 2024.

[12] N. N. E. Perimawati, R. R. Huizen, and D. P. Hostiadi, "Analisa Pengaruh Pre-Procesing Data Untuk Model Deteksi Akun Palsu Pada Media Sosial," in *Seminar Hasil Penelitian Informatika dan Komputer (SPINTER)*, vol. 2, no. 1, pp. 1117-1122, 2025.

[13] Jamaluddin, T., Bijaksana, M. A., & Asror, I. (2020). Perbandingan Algoritma Sentencepiece BPE dan Unigram Pada Tokenisasi Artikel Bahasa Indonesia. eProceedings of Engineering, 7(2).

[14] B. A. Maulana, M. J. Fahmi, A. M. Imran, and N. Hidayati, "Analisis Sentimen Terhadap Aplikasi Pluang Menggunakan Algoritma Naive Bayes dan Support Vector Machine (SVM): Sentiment Analysis of Pluang Applications With Naive Bayes and Support Vector Machine (SVM) Algorithm," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 2, pp. 375-384, 2024.

[15] R. Syahwaluddin and D. Alita, "Penerapan Oversampling Pada Klasifikasi Ujaran Kebencian Menggunakan Bidirectional Encoder Representations from Transformers," The Indonesian Journal of Computer Science, vol. 13, no. 4, 2024.