Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3464~3478
e-ISSN: 2548-6861

Comparative Performance of SVM and BERT-Base Using Hybrid Preprocessing for Fast Fashion Sentiment Analysis

Restu Lestari Mulianingrum [1]*, Erwin Yudi Hidayat [2]*
* Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia
111202214668@mhs.dinus.ac.id [1], erwin@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2025-09-29
Revised 2025-11-17
Accepted 2025-11-22

Keyword:

BERT,
Fast Fashion,
Sentiment Analysis,
SVM,
TikTok

ABSTRACT

Fast fashion poses major environmental and social challenges, yet public awareness in Indonesia remains insufficiently understood. This study compares Support Vector Machine and BERT-Base for sentiment analysis of 3,513 TikTok comments on fast fashion sustainability using a hybrid preprocessing pipeline that incorporates a 404-entry slang dictionary and IndoNLP utilities to address informal language, code-mixing, and character elongation. Sentiment labels generated using VADER were validated against 1,747 manually annotated samples, achieving Cohen's Kappa of 0.7155, indicating substantial agreement. BERT-Base achieves 92.7% accuracy with F1-scores of 0.86, 0.94, and 0.93 for negative, neutral, and positive classes, while SVM attains competitive 90.4% accuracy with F1-scores of 0.84, 0.93, and 0.91. BERT demonstrates superior negative sentiment detection with recall of 0.87 compared to SVM at 0.82, critical for identifying sustainability concerns. Computational analysis reveals significant trade-offs as BERT requires 230.2 seconds of GPU training and 3.449 seconds of inference, whereas SVM operates efficiently on CPU with 25.9 seconds of training and 0.051 seconds of inference, representing 8.9× and 67.6× efficiency advantages. The sentiment distribution comprising 46.9% neutral, 34.5% positive, and 18.6% negative comments indicates limited critical awareness among Indonesian users. These findings demonstrate that systematic preprocessing bridges the performance gap between classical and transformer models while enabling deployment decisions based on resource constraints, providing methodological insights for low-resource informal text analysis and practical guidance for scalable social listening, greenwashing detection, and evidence-based sustainability communication strategies.

I. INTRODUCTION

The global fashion sector has experienced a significant transformation following the emergence of fast fashion models, characterized by rapid manufacturing cycles, continuous trend shifts, and low prices. [1], [2]. This business model promotes excessive clothing consumption, leading to significant environmental and social impacts. The United Nations (UN) reports that the fashion industry ranks as the world's second-largest polluting sector, contributing 8–10% of global carbon emissions and concerning the circumstances for 20% of water waste, with these figures expected to rise by 60% before 2030 [1], [3]. Additionally, the United Nations Environment Programme (UNEP) reports that synthetic fibers account for up to 62% of the global textile market, making them a primary source of microplastics. Furthermore, 85–87% of clothing ends up in landfills, and only 1% is successfully recycled [4], [5]. Beyond environmental issues, fast fashion creates social impacts, particularly poor working conditions that often result from low-cost production in developing countries [2], [3], [6], [7]. Indonesia is among the affected countries, where relatively low labor costs and limited regulatory enforcement create conditions for worker exploitation, leading to a dilemma between industrial growth and the protection of worker welfare [8].

Despite widespread criticism of its negative impacts, the international fast fashion market continues to experience robust growth due to consumer demand for affordable clothing. In 2025, the market is estimated to reach USD 163.21 billion and is projected to reach USD 214.24 billion by 2029 [9]. In Indonesia, brands such as Zara, Shein, H&M, and Uniqlo dominate primarily among Gen Z through social media and influencer strategies [10]. For this generation, fashion trends represent more than basic clothing needs, serving as forms of self-expression and social identity [11]. This phenomenon has sparked more public discussions on social media. TikTok constitutes a main platform for education and awareness, where videos about sustainable clothing influence consumer behavior and foster diverse opinions on the fast fashion industry. [11], [12], [13]. Content spanning from sustainability campaigns and environmental education to criticism of labor exploitation and textile waste reflects users' active engagement in this phenomenon. The resulting comments reveal diverse public opinions, making them relevant for systematic study.

Sentiment analysis provides a relevant approach for understanding public opinion on social media platforms, including TikTok. TikTok user comments contain positive, negative, and neutral expressions that can be classified as indicators of public perception toward specific phenomena [14], [15], [16]. However, TikTok content frequently deviates from standard Indonesian language conventions, incorporating abbreviations, slang, and informal sentence structures. These characteristics pose significant challenges for text processing and sentiment analysis.

Support Vector Machine (SVM) was selected as the representative classical model in this study due to its strong theoretical foundation in optimization theory and its use of the structural risk minimization principle. SVM identifies an optimal hyperplane that maximizes the margin between classes, making it robust to overfitting and effective in high-dimensional feature spaces commonly found in text-based sentiment analysis [17]. Compared with probabilistic models (e.g., Naïve Bayes), which rely on strong feature independence assumptions, or tree-based models (e.g., Random Forest), which are often susceptible to class imbalance, SVM provides more stable performance on noisy and imbalanced sentiment data, typically present in social media text [17], [18], [19].

Multiple studies have reported strong SVM performance in sentiment analysis applications. Sitepu *et al.* [20] reported an average accuracy of 97.3% on Shopee product reviews applying a linear SVM with TF-IDF. Rahardi *et al.* [21] achieved 92% accuracy with perfect recall for negative sentiment on Indonesian COVID-19 vaccination tweets. Similarly, Khan *et al.* [18] reported that SVM obtained 80.4% accuracy on IBM search panel data, slightly outperforming Random Forest. This margin-based classification mechanism, when combined with TF–IDF feature representation, enables SVM to maintain interpretability while requiring relatively modest computational resources [22], [23], [24].

With the evolution of deep learning methods, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), have introduced a new paradigm for comprehensively understanding linguistic context. In this study, BERT is employed as a representative modern model as a result of its ability to capture sentence-level context. BERT-Base has been proven effective in sentiment analysis, consistently getting high accuracy across various datasets. Chiorrini *et al.* [25] reported an accuracy of 92% for sentiment analysis on Twitter data. Khan *et al.* [26] also achieved 92% accuracy on tweet sentiment classification. Geetha and Renuka [27] demonstrated that BERT-Base Uncased outperformed traditional machine learning methods such as SVM in terms of veracity on e-commerce reviews.

The selection of BERT-Base as the representative modern model is grounded in methodological considerations and its comparative validity against alternative architectures. Recurrent models such as LSTM exhibit limitations in capturing bidirectional context and global semantic dependencies due to their sequential processing and the absence of self-attention mechanisms [28], [29]. Indonesian-language transformer models, including IndoBERT and RoBERTa-ID, although explicitly trained for Indonesian, perform suboptimally on social media text because their pretraining corpora are dominated by formal sources such as news articles and Wikipedia. In contrast, TikTok comments typically contain informal expressions, slang, and abbreviations [30], [31]. Furthermore, TikTok comments often involve code-mixing between Indonesian and English, which monolingual Indonesian-only models cannot effectively handle.

To address these limitations, this study employs BERT-Base-Uncased with a translation-based preprocessing approach combined with hybrid preprocessing to handle code-mixing, slang, and emotional expressions. Similar approaches have proven effective, as demonstrated by Usman *et al.* [32], who achieved an F1-score of 0.84 on multilingual datasets using the Google Translate API. BERT-Base-Uncased is selected due to its case-insensitive nature, making it suitable for social media text with inconsistent capitalization. The model is also pretrained on cross-domain corpora that include informal text, ensuring its relevance to social media contexts and enabling direct comparison with international benchmarks that predominantly use this model as a baseline [32], [33], [34], [35].

This study aims to compare the performance of Support Vector Machine (SVM) and BERT-Base in analyzing sentiment from TikTok comments characterized by informal language and code-mixing. The comparison focuses on evaluating both classification accuracy and computational efficiency after applying hybrid preprocessing. This study is expected to contribute methodologically to sentiment analysis on informal text and provide practical insights for social listening, greenwashing detection, and the development of sustainable digital communication strategies relevant to fashion consumption culture in Indonesia.

II. METHOD

This research follows the workflow illustrated in Figure 1, which includes data collection, preprocessing, translation, labeling, label encoding, dataset splitting, modeling using SVM and BERT, and classification evaluation.

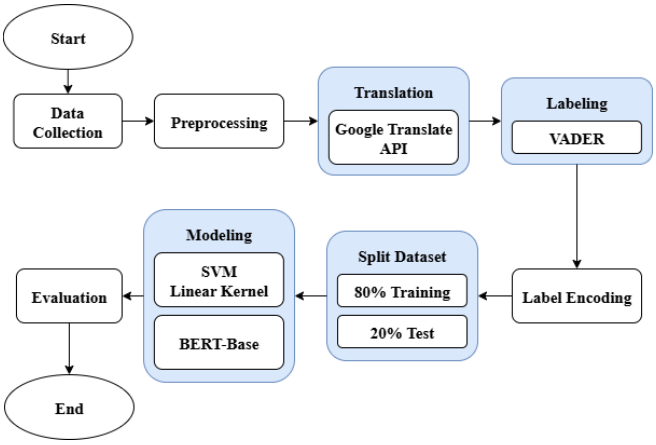

Figure 1. Research workflow

A. Data Collection

The data obtained through web scraping of comments on a TikTok video from the account @dinoaugusto_. The account, owned by Dino Augusto, a lecturer at LaSalle College Jakarta specializing in fashion and consumer psychology, has been active in the fashion industry for more than a decade. It was chosen because of its consistent focus on fast fashion, including its environmental and social impacts, as well as advocacy for conscious consumption and support for local MSMEs, which makes it relevant to the objective of analyzing public sentiment.

A specific video was chosen based on its high engagement and strong relevance to the research topic, ensuring that the collected comments provide quality and representative data for sentiment analysis. Using the Apify API, the process yielded 3,990 informal comments in Indonesian. Only two attributes were utilized in this study, stored in CSV format, as shown in Table 1.

TABLE I
DATASET DESCRIPTION

Attribute	Description
uniqueId	TikTok username
text	Comment text posted by the TikTok user

B. Preprocessing

The preprocessing stage is an essential step in Indonesian text analysis, as social media texts, particularly those from TikTok, often contain abbreviations, slang, spelling variations, and emojis [36], [37]. In this study, the preprocessing stage consists of seven steps, as illustrated in Figure 2.

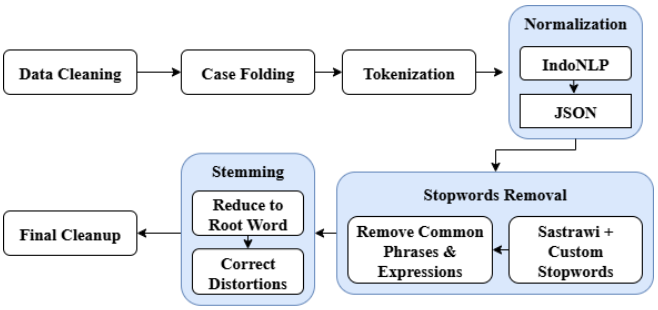

Figure 2. Preprocessing Workflow

1) *Data Cleaning*

The introductory stage is data cleaning to eliminate inconsequential elements from sentiment analysis. This process involves eliminating duplicates, handling missing values, and removing textual noise, including mentions, hashtags, URLs, non-alphabetic characters, emojis, and excessive whitespace. [38], [39].

2) *Case Folding*

All text is converted into lowercase to standardize word formats. This step is essential to avoid inconsistencies in recognizing identical words that differ only in capitalization [18], [36].

3) *Tokenization*

Tokenization is performed to split the text into individual words or phrases using regular expressions, allowing each token to be analyzed separately [21], [39], [40].

4) *Normalization*

Normalization is applied to convert non-standard words into their standard forms according to Indonesian language rules [21]. This process employs the IndoNLP library [41] to handle words with repeated characters and slang terms, along with a JSON-based slang dictionary containing 404 entries that map non-standard words to their standard equivalents.

5) *Stopword Removal*

Stopword removal is applied to eliminate common high-frequency words that do not provide a meaningful contribution to sentiment analysis [37], [40]. A hybrid stopword removal approach is implemented in this study, comprising:

- Modified sastrawi stopwords retain negation words ("tidak", "jangan", "bukan") and emotion terms ("baik", "buruk", "banget") to preserve sentiment polarity,
- Domain-specific entries add thirty-two items covering Indonesian social media patterns, including informal particles, address terms, and laughter expressions,
- Multi-word Expression Removal uses regex rules to eliminate phrases such as "terima kasih" and "tidak apa apa".

Additionally, the preprocessing handles social media-specific linguistic phenomena, including variations of affirmative particles ("ya"), repetitive laughter expressions

("wkwkwk", "hahaha"), character elongation patterns, and reduplication constructs. This stage utilizes IndoNLP utilities to optimize token normalization before stemming operations.

6) *Stemming*

The stemming process normalizes words to their base representations by removing morphological prefixes and suffixes [21], [26], [39], [42]. Additionally, post-stemming mapping is applied using a custom dictionary to revitalize words that were distorted or lost during previous processing stages.

7) *Final Cleanup*

The final stage goal is to provide text consistency by retaining only alphabetic characters and spaces, removing tokens with fewer than two characters, and normalizing whitespace by reducing ensuing spaces to a single space and removing leading or trailing spaces.

C. Translating

The preprocessed text is translated into English by Google Translate API via the googletrans library to ensure compatibility with the BERT-Base model's requirements [43]. This step is necessary because BERT-Base Uncased was initially trained on large-scale English corpora, and translation allows Indonesian social media text to be represented within the same semantic space as the pre-trained model.

D. Labeling

Automatic labeling is conducted using the VADER Sentiment Analyzer, an open-source lexicon and rule-based sentiment analysis method developed in 2014 [44]. Each translated comment is analyzed to determine a polarity score (compound score) as shown in Equation (1).

$$Compound = \frac{x}{\sqrt{(x)^2 + \alpha}} \quad (1)$$

In this equation, x represents the sum of sentiment scores (Positive, Negative, and Neutral), while α is the normalization factor with a default value of 15 [45]. The three-class sentiment mapping utilizes compound score thresholds, such that scores ≥ 0.05 are labeled as Positive, scores ≤ -0.05 are labeled as Negative, and scores within the range of -0.05 to 0.05 are labeled as Neutral [44].

To ensure the reliability of the automatic labeling results, manual labeling was conducted on 1,747 randomly selected comments. The authors assigned sentiment labels based on the contextual meaning of each sentence using three main categories: positive, neutral, and negative. The manual labeling criteria are presented in Table 2.

TABLE II
MANUAL LABELING CRITERIA

Sentiment Category	Determination Criteria
Positive	Contains praise, support, appreciation, or positive emotion toward sustainability education in fashion.
Neutral	Informative, questioning, or not expressing a clear sentiment orientation.
Negative	Contains criticism, complaints, rejection, or negative emotion toward fast fashion practices or the fashion industry.

The manual labeling results are then used to evaluate the agreement with VADER-based automatic labeling using accuracy and Cohen's Kappa coefficient. The interpretation of Cohen's Kappa in this study follows the criteria proposed by Landis and Koch [46], where values < 0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement. This evaluation is also used to identify potential bias introduced by the translation-based annotation process.

E. Label Encoding

Label encoding converts categorical sentiment labels from text to numerical representations by using the LabelEncoder class from the scikit-learn library. The mapping assigns a value of 0 to Negative, 1 to Neutral, and 2 to Positive [47], [48]. This transformation enables efficient data processing in classification algorithms, such as SVM and BERT.

F. Split Dataset

An 80:20 train-test split is performed on the encoded dataset using the train_test_split function, incorporating stratified sampling to ensure a representative distribution of sentiments. This approach is consistently executed for both the SVM and BERT models.

G. Support Vector Machine (SVM)

The Support Vector Machine approach functions as a supervised learning technique that finds the best separating hyperplane for categorizing data into different classes [38], [44]. The SVM implementation process is illustrated in Figure 3.

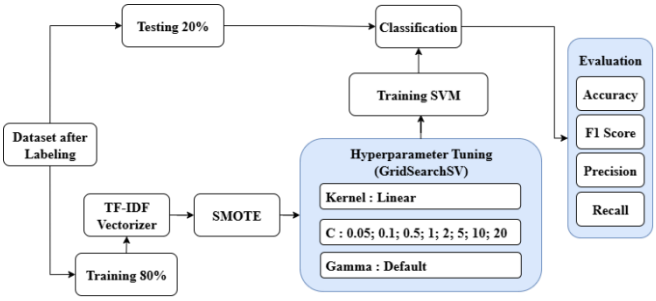

Figure 3. Support Vector Machine (SVM) Workflow

The labeled and split text data are represented in numerical form using the TF-IDF Vectorizer with unigram features. Additionally, character-level n-grams (3-6 characters) are extracted to capture morphological patterns in informal social media text. Term Frequency (TF) calculates the ratio of word occurrences within a document, while Inverse Document Frequency (IDF) evaluates word importance based on its distribution across the entire corpus [27]. The final TF-IDF weight is obtained by multiplying these two components, and both word-level and character-level features are combined using feature union. The TF-IDF formulation is shown in Equations (2) – (4).

$$TF(t,d) = \frac{\text{freq}(t,d)}{\text{number of terms in } d} \quad (2)$$

$$IDF(t) = \log_2\left(\frac{N}{1+df(t)}\right) \quad (3)$$

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (4)$$

Notes:
$\text{freq}(t,d)$ = frequency of term t in document d
N = total number of documents (in the corpus)
$df(t)$ = number of documents containing term t

Subsequently, the class imbalance in the training data was handled using the Synthetic Minority Over-sampling Technique (SMOTE), which oversamples minority classes by generating synthetic instances through linear interpolation between nearest neighbor examples identified via the K-Nearest Neighbors algorithm [21]. The SVM classifier was then trained with a linear kernel, which is computationally efficient for high-dimensional sparse text features, and its decision function is formulated as presented in Equation (5) [17].

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b\right) \quad (5)$$

In this equation, x denotes the input vector, x_i the training data vectors, y_i, the class labels, α_i the weights, and b the bias term.

Following the definition of the model parameters, optimal parameter selection was performed using GridSearchCV [42]. The process examined eight candidate values for the regularization parameter C (0.05, 0.1, 0.5, 1, 2, 5, 10, and 20) with a linear kernel, while the gamma parameter was set to its default value. The resulting model was then applied to three-class sentiment classification comprising Positive, Neutral, and Negative categories.

*H. Bidirectional Encoder Representations from transformers (*BERT-Base*)*

The BERT model, or Bidirectional Encoder Representations from Transformers, is built upon the transformer architecture developed by Google. While traditional approaches process textual input unidirectionally, BERT employs bidirectional processing, allowing it to infer word meanings by simultaneously considering context from both directions. This approach enables BERT to capture semantic relationships and contextual dependencies within sentences more accurately [27], [39].

The training process employs two key self-supervised learning tasks to achieve this bidirectional understanding. Masked Language Model (MLM) focuses on predicting randomly masked tokens within text, and Next Sentence Prediction (NSP) focuses on learning inter-sentence relations [49], [50]. The overall architecture of BERT is shown in Figure 4.

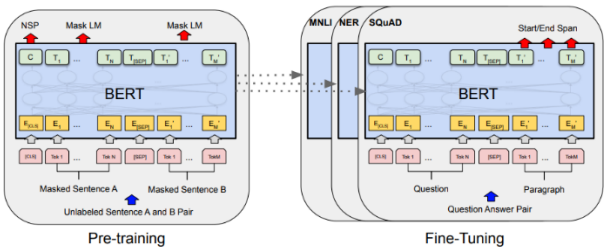

Figure 4. BERT Architecture

In this study, the BERT-Base Uncased variant is employed, comprising 12 transformer encoder blocks, hidden vector dimensions of 768, 12 parallel attention heads, and approximately 110 million parameters [50]. The core mechanism of BERT is self-attention, which is formulated as shown in Equation (6) [48], [50].

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

To improve representation, BERT applies multi-head attention, as presented in Equation (7)[50].

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (7)$$

Where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ with W_i^Q, W_i^K, W_i^V as projection matrices for query, key, and value, respectively, and W^O As the output projection matrix. In BERT-Base, the number of heads is $h = 12$. This mechanism enables the model to attend to information from multiple representation dimensions simultaneously. Figure 5 illustrates the workflow of the BERT-Base method used in this study.

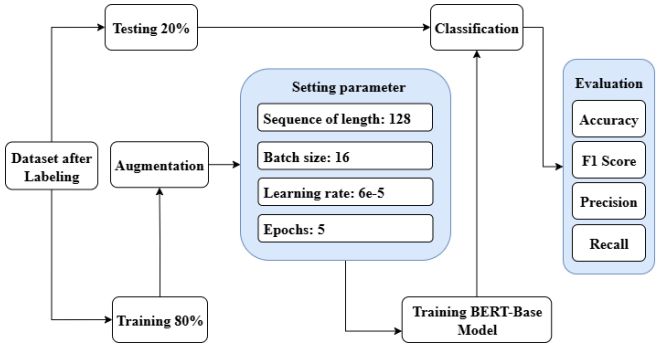

Figure 5. BERT-Base workflow

Training data augmentation was performed by applying the NLPaug library with synonym replacement techniques to enhance text diversity without altering semantic meaning [43]. Tokenization was conducted using the BERT tokenizer with a maximum sequence length of 128 to generate input IDs and attention masks [51]. Training of the BERT-Base Uncased model was conducted with optimization parameters comprising a batch size of 16, a learning rate of 6e-5, and 5 epochs.

I. Evaluation

The performance of the model was evaluated on the test dataset using four standard metrics: accuracy, precision, recall, and F1-score [42], [49], [52]. These metrics were derived from the values in the confusion matrix shown in Table 3.

TABLE III
CONFUSION MATRIX

Class (Actual)	Predicted		
	Negative	**Neutral**	**Positive**
Negative	TNeg	FNn	FPos
Neutral	FNeg	TNn	FPos2
Positive	FNeg2	FNn2	TPos

Notes:
TNeg = correct prediction for Negative
FNeg, FNeg2 = Negative misclassified as Neutral or Positive
FNn, FNn2 = Neutral instances misclassified as Negative or Positive
TNn = correct prediction for the Neutral class
FPos, FPos2 = Positive instances misclassified as Negative or Neutral
TPos = correct prediction for the Positive class

Accuracy is counted as the proportion of accurately classified instances relative to the total number of test samples, as shown in Equation (8).

$$Accuracy = \frac{TNeg + TNn + TPos}{\text{Total number of test instance}} \quad (8)$$

Precision calculates the proportion of accurate positive classifications relative to the total instances assigned to that class, as presented in Equations (9) – (11).

$$Precision_{negative} = \frac{TNeg}{TNeg + FNeg + FNeg2} \quad (9)$$

$$Precision_{neutral} = \frac{TNn}{TNn + FNn + FNn2} \quad (10)$$

$$Precision_{positive} = \frac{TPos}{TPos + FPos + FPos2} \quad (11)$$

Recall evaluates the model's ability to identify instances of each class according to the actual labels, as shown in Equations (12) – (14).

$$Recall_{negative} = \frac{TNeg}{TNeg + FNn + FPos} \quad (12)$$

$$Recall_{\text{neutral}} = \frac{TNn}{TNn + TNeg + FPos2} \quad (13)$$

$$Recall_{positive} = \frac{TPos}{TPos + FNeg2 + FNn2} \quad (14)$$

F1-score assesses classifier performance by calculating the harmonic mean between precision and recall, thus providing an equal-weighted measure of predictive accuracy for imbalanced datasets, as shown in Equation (15)

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \quad (15)$$

Where $c \in \{\text{Negative, Neutral, Positive}\}$ denotes the sentiment class.

III. RESULTS AND DISCUSSION

The results of this study are systematically presented, beginning with data collection and preprocessing, followed by the evaluation of the classification models. The analysis focuses on the comparative performance of SVM and BERT in classifying sentiment from TikTok comments related to fast fashion.

A. Data Collection

A total of 3,990 TikTok comments were collected through web scraping from the @dinoaugusto_ account using the Apify API. The dataset consists of two main attributes, where uniqueId refers to the username and *text* refers to the comment content. This dataset formed the basis for the preprocessing stage and was stored in CSV format. Sample entries from the collected data are presented in Table 4.

TABLE IV
SAMPLE OF COLLECTED TIKTOK DATA

uniqueId	text
a.canddd	normalisasi outfit repeater
iniraitsugalea	abang ini susah payah edukasi ,tp dluar sana org belomba2 jualan sisa fast fashion dg harga murah dr luar,sampe umkm bnyak yg kalah saing 😭
influencer_magang	Terbaik bang. Udah Nerapin capsule wardrobe solusi ngurangin beli2 baju terus
buguru99_	Pengaruh selebtok dan selebgram yang tiap hari racunin baju 😭
msy_als	plis itu perusahaan ultra fast fashion di banned...jgn ksih izin produksinyaaa

An exploratory analysis was conducted to characterize the dataset. As shown in Figure 6, the comments are predominantly written in Indonesian (86.1%), with smaller proportions in English (2.1%), mixed-language expressions (1.7%), and undetected text (10.2%). This linguistic distribution indicates that the audience engaging with the video is largely local and communicates using informal conversational patterns.

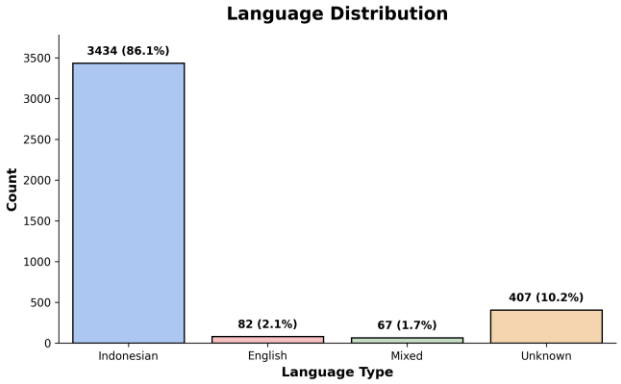

Figure 6. Language Distribution in TikTok Comments

Further lexical analysis highlights recurring themes in fashion consumption. The top 20 most frequent words, illustrated in Figure 7, show that terms such as “baju” (2,148 occurrences) and “beli” (1,589 occurrences) dominate the corpus. The frequent appearance of pronouns (e.g., “aku”) and informal connectors (e.g., “yg”, “aja”) reinforces the conversational nature of TikTok interactions. The presence of culturally specific terms such as “lebaran” reflects temporal and social nuances relevant to fashion-related behavior.

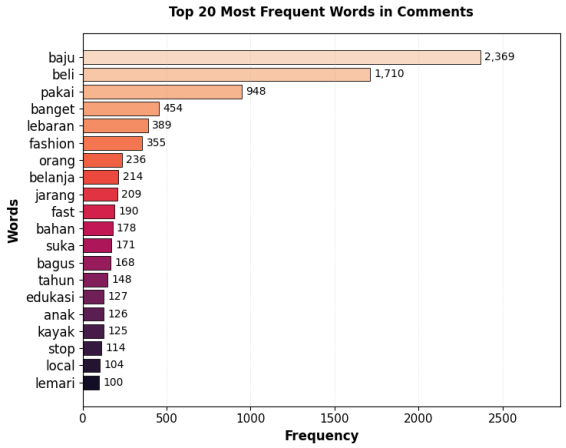

Figure 7. Top 20 Most Frequent Words in TikTok Comments

B. *Preprocessing*

The preprocessing stage produced a cleaner and more consistent text corpus, which was then prepared for feature extraction in SVM and tokenization in BERT. The outcomes of each step are described in detail as follows.

1) *Data Cleaning*

From the initial 3,990 raw comments, 3,565 valid comments remained after the data cleaning process. A total of 425 comments (10.65%) were removed, including empty entries, duplicates, and irrelevant elements such as mentions, hashtags, URLs, numbers, and non-alphabetic characters. A comparison of the results is presented in Table 5, which shows the transformation of comments before and after cleaning.

TABLE V
COMPARISON BEFORE AND AFTER DATA CLEANING

Before Data Cleaning	After Data Cleaning
normalisasi outfit repeater	normalisasi outfit repeater
abang ini susah payah edukasi ,tp dluar sana org belomba2 jualan sisa fast fashion dg harga murah dr luar,sampe umkm bnyak yg kalah saing 😭	abang ini susah payah edukasi tp dluar sana org belomba jualan sisa fast fashion dg harga murah dr luar sampe umkm bnyak yg kalah saing
Terbaik bang. Udah Nerapin capsule wardrobe solusi ngurangin beli2 baju terus	Terbaik bang Udah Nerapin capsule wardrobe solusi ngurangin beli baju terus
Pengaruh selebtok dan selebgram yang tiap hari racunin baju 😭	Pengaruh selebtok dan selebgram yang tiap hari racunin baju
plis itu perusahaan ultra fast fashion di banned...jgn ksih izin produksinyaaa	plis itu perusahaan ultra fast fashion di banned jgn ksih izin produksinyaaa

2) *Case Folding*

The case-folding step converts all comment text into lowercase letters to ensure a uniform writing format. This process prevents identical words from being treated differently due to variations in capitalization, for example,

"Terbaik" becomes "terbaik" in this way, the meaning of the sentence is preserved, while the text format becomes consistent. A comparison of the case before and after folding is presented in Table 6.

TABLE VI
COMPARISON BEFORE AND AFTER CASE FOLDING

Before Case Folding	After Case Folding
normalisasi outfit repeater	normalisasi outfit repeater
abang ini susah payah edukasi tp dluar sana org belomba jualan sisa fast fashion dg harga murah dr luar sampe umkm bnyak yg kalah saing	abang ini susah payah edukasi tp dluar sana org belomba jualan sisa fast fashion dg harga murah dr luar sampe umkm bnyak yg kalah saing
Terbaik bang Udah Nerapin capsule wardrobe solusi ngurangin beli baju terus	terbaik bang udah nerapin capsule wardrobe solusi ngurangin beli baju terus
Pengaruh selebtok dan selebgram yang tiap hari racunin baju	pengaruh selebtok dan selebgram yang tiap hari racunin baju
plis itu perusahaan ultra fast fashion di banned jgn ksih izin produksinyaaa	plis itu perusahaan ultra fast fashion di banned jgn ksih izin produksinyaaa

3) *Tokenization*

After case folding, TikTok comments in lengthy and variable text formats were processed through tokenization to decompose text into individual word tokens. This process generates a structured word-level representation required for subsequent sentiment analysis. The tokenization results are presented in Table 7.

TABLE VII
COMPARISON BEFORE AND AFTER TOKENIZATION

Before Tokenization	After Tokenization
normalisasi outfit repeater	['normalisasi', 'outfit', 'repeater']
abang ini susah payah edukasi tp dluar sana org belomba jualan sisa fast fashion dg harga murah dr luar sampe umkm bnyak yg kalah saing	['abang', 'ini', 'susah', 'payah', 'edukasi', 'tp', 'dluar', 'sana', 'org', 'belomba', 'jualan', 'sisa', 'fast', 'fashion', 'dg', 'harga', 'murah', 'dr', 'luar', 'sampe', 'umkm', 'bnyak', 'yg', 'kalah', 'saing']
terbaik bang udah nerapin capsule wardrobe solusi ngurangin beli baju terus	['terbaik', 'bang', 'udah', 'nerapin', 'capsule', 'wardrobe', 'solusi', 'ngurangin', 'beli', 'baju', 'terus']
pengaruh selebtok dan selebgram yang tiap hari racunin baju	['pengaruh', 'selebtok', 'dan', 'selebgram', 'yang', 'tiap', 'hari', 'racunin', 'baju']
plis itu perusahaan ultra fast fashion di banned jgn ksih izin produksinyaaa	['plis', 'itu', 'perusahaan', 'ultra', 'fast', 'fashion', 'di', 'banned', 'jgn', 'ksih', 'izin', 'produksinyaaa']

4) *Normalization*

Text normalization was applied to standardize TikTok comments, which often contain slang words, character repetitions, and emotional expressions. This process utilized the IndoNLP library functions replace_word_elongation to reduce repetitive characters (e.g., "produksinyaaa" becomes "produksinya") and replace_slang to convert slang terms into standard words. Additionally, a custom JSON-based slang dictionary comprising 404 entries was developed to enhance the mapping of social media slang, particularly those prevalent on TikTok. The normalization results are presented in Table 8.

TABLE VIII
COMPARISON BEFORE AND AFTER NORMALIZATION

Before Normalization	After Normalization
['normalisasi', 'outfit', 'repeater']	['normalisasi', 'pengulang', 'outfit']
['abang', 'ini', 'susah', 'payah', 'edukasi', 'tp', 'dluar', 'sana', 'org', 'belomba', 'jualan', 'sisa', 'fast', 'fashion', 'dg', 'harga', 'murah', 'dr', 'luar', 'sampe', 'umkm', 'bnyak', 'yg', 'kalah', 'saing']	['abang', 'ini', 'susah', 'payah', 'edukasi', 'tapi', 'diluar', 'sana', 'orang', 'belomba', 'jualan', 'sisa', 'fast', 'fashion', 'dengan', 'harga', 'murah', 'dari', 'luar', 'sampai', 'UMKM', 'banyak', 'yang', 'kalah', 'saing']
['terbaik', 'bang', 'udah', 'nerapin', 'capsule', 'wardrobe', 'solusi', 'ngurangin', 'beli', 'baju', 'terus']	['terbaik', 'abang', 'sudah', 'menerapkan', 'capsule', 'wardrobe', 'solusi', 'mengurangi', 'beli', 'baju', 'terus']
['pengaruh', 'selebtok', 'dan', 'selebgram', 'yang', 'tiap', 'hari', 'racunin', 'baju']	['pengaruh', 'selebritas', 'TikTok', 'dan', 'selebritas', 'Instagram', 'yang', 'tiap', 'hari', 'meracuni', 'baju']
['plis', 'itu', 'perusahaan', 'ultra', 'fast', 'fashion', 'di', 'banned', 'jgn', 'ksih', 'izin', 'produksinyaaa']	['tolong', 'itu', 'perusahaan', 'ultra', 'fast', 'fashion', 'di', 'banned', 'jangan', 'kasih', 'izin', 'produksinya']

5) *Stopword Removal*

Stopword removal was applied to eliminate common words that do not contribute to sentiment analysis. A customized Sastrawi stopword list was employed, with modifications to retain negation words such as "tidak" (not), "jangan" (don't), and "bukan" (not) to preserve polarity. Additional entries characteristic of TikTok comments, including laughter expressions and profanity, were incorporated. Regex-based rules were further implemented to address word repetitions. The results of stopword removal are presented in Table 9.

TABLE IX
COMPARISON BEFORE AND AFTER STOPWORD REMOVAL

Before Stopword Removal	After Stopword Removal
['normalisasi', 'pengulang', 'outfit']	['normalisasi', 'pengulang', 'outfit']
['abang', 'ini', 'susah', 'payah', 'edukasi', 'tapi', 'diluar', 'sana', 'orang', 'belomba', 'jualan', 'sisa', 'fast', 'fashion', 'dengan', 'harga', 'murah', 'dari', 'luar',	['susah', 'payah', 'edukasi', 'diluar', 'sana', 'orang', 'berlomba', 'jualan', 'sisa', 'fast', 'fashion', 'harga', 'murah', 'luar', 'umkm', 'banyak', 'kalah', 'saing']

'sampai', 'UMKM', 'banyak', 'yang', 'kalah', 'saing']	
['terbaik', 'abang', 'sudah', 'menerapkan', 'capsule', 'wardrobe', 'solusi', 'mengurangi', 'beli', 'baju', 'terus']	['terbaik', 'sudah', 'menerapkan', 'capsule', 'wardrobe', 'solusi', 'mengurangi', 'beli', 'baju', 'terus']
['pengaruh', 'selebritas', 'TikTok', 'dan', 'selebritas', 'Instagram', 'yang', 'tiap', 'hari', 'meracuni', 'baju']	['pengaruh', 'selebritas', 'tiktok', 'selebritas', 'instagram', 'tiap', 'hari', 'meracuni', 'baju']
['tolong', 'itu', 'perusahaan', 'ultra', 'fast', 'fashion', 'di', 'banned', 'jangan', 'kasih', 'izin', 'produksinya']	['perusahaan', 'ultra', 'fast', 'fashion', 'banned', 'jangan', 'kasih', 'izin', 'produksinya']

6) *Stemming*

After stopword removal, the process continues with stemming to convert inflected words into their base forms. This step is performed using the Sastrawi library with additional post-stemming mapping, such as "terap" becoming "menerapkan" and "moga" becoming "semoga". This step ensures that stemming results remain contextually appropriate for the Indonesian language and maintain consistent meaning before subsequent processing steps. The stemming results are presented in Table 10.

TABLE X
COMPARISON BEFORE AND AFTER STEMMING

Before Stemming	After Stemming
['normalisasi', 'pengulang', 'outfit']	['normalisasi', 'ulang', 'outfit']
['susah', 'payah', 'edukasi', 'diluar', 'sana', 'orang', 'berlomba', 'jualan', 'sisa', 'fast', 'fashion', 'harga', 'murah', 'luar', 'umkm', 'banyak', 'kalah', 'saing']	['susah', 'payah', 'edukasi', 'luar', 'sana', 'orang', 'lomba', 'jual', 'sisa', 'fast', 'fashion', 'harga', 'murah', 'luar', 'umkm', 'banyak', 'kalah', 'saing']
['terbaik', 'sudah', 'menerapkan', 'capsule', 'wardrobe', 'solusi', 'mengurangi', 'beli', 'baju', 'terus']	['baik', 'sudah', 'menerapkan', 'capsule', 'wardrobe', 'solusi', 'kurang', 'beli', 'baju', 'terus']
['pengaruh', 'selebritas', 'tiktok', 'selebritas', 'instagram', 'tiap', 'hari', 'meracuni', 'baju']	['pengaruh', 'selebritas', 'tiktok', 'selebritas', 'instagram', 'tiap', 'hari', 'racun', 'baju']
['perusahaan', 'ultra', 'fast', 'fashion', 'banned', 'jangan', 'kasih', 'izin', 'produksinya']	['usaha', 'ultra', 'fast', 'fashion', 'banned', 'jangan', 'kasih', 'izin', 'produksi']

7) *Final Cleanup*

The final cleanup step produces comment text that is completely clean from numbers, symbols, and non-alphabetic characters, while retaining relevant words with a minimum length of two characters. This cleanup process makes the data more concise and consistent. The total number of data after the final preprocessing step is 3,513. The final preprocessing results are presented in Table 11.

TABLE XI
COMPARISON BEFORE AND AFTER FINAL CLEANUP

Before Final Cleanup	After Final Cleanup
['normalisasi', 'ulang', 'outfit']	normalisasi ulang outfit
['ini', 'susah', 'payah', 'edukasi', 'luar', 'sana', 'orang', 'lomba', 'jual', 'sisa', 'fast', 'fashion', 'harga', 'murah', 'luar', 'umkm', 'banyak', 'kalah', 'saing']	susah payah edukasi luar sana orang belomba jual sisa fast fashion harga murah luar umkm banyak kalah saing
['baik', 'sudah', 'menerapkan', 'capsule', 'wardrobe', 'solusi', 'kurang', 'beli', 'baju', 'terus']	baik menerapkan capsule wardrobe solusi kurang beli baju terus
['pengaruh', 'selebritas', 'tiktok', 'selebritas', 'instagram', 'tiap', 'hari', 'racun', 'baju']	pengaruh selebritas tiktok selebritas instagram setiap hari racun baju
['usaha', 'ultra', 'fast', 'fashion', 'banned', 'jangan', 'kasih', 'izin', 'produksi']	usaha ultra fast fashion banned jangan kasih izin produksi

After preprocessing, comments become cleaner and more focused on meaningful words relevant to fast fashion issues. Figure 8 displays the 20 most frequent words that illustrate the primary focus of comments, particularly on clothing consumption activities such as purchasing and wearing.

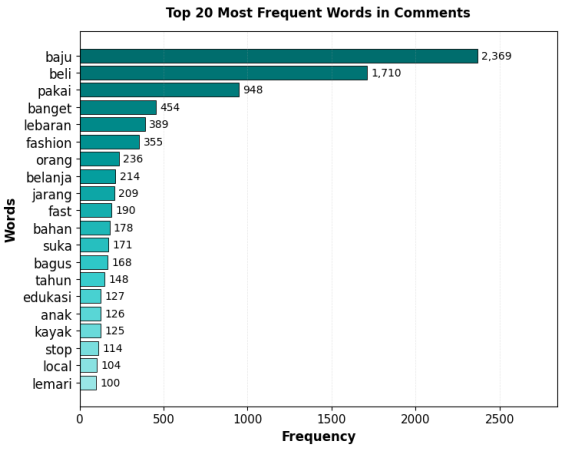

Figure 8. Most Frequent Words After Preprocessing

The words "baju" (clothes), "beli" (buy), and "pakai" (wear) have the highest frequency, indicating that clothing consumption activities are the primary focus of comments. In other words, terms such as “lebaran” (Eid celebration), “fashion”, and “belanja” (shopping) reveal cultural context and seasonal consumption patterns. Additionally, the appearance of words like “sampah” (trash), “stop, and “edukasi” (education) indicates public attention toward environmental issues and calls for more mindful consumption. These results demonstrate that the preprocessing stage not only cleans the text but also emphasizes words relevant to sentiment analysis.

C. Translation

After completing the preprocessing stages, Indonesian comments are translated into English to ensure compatibility with the BERT-Base model, which is trained using an English corpus. The translation process is performed automatically using the googletrans library with parameters src='id' and dest='en'. Examples of translation results are presented in Table 12.

TABLE XII
COMPARISON BEFORE AND AFTER FINAL CLEANUP

Before Translation	After Translation
normalisasi ulang outfit	normalization of outfit
susah payah edukasi luar sana orang belomba jual sisa fast fashion harga murah luar umkm banyak kalah saing	hard to education outside people who sell the rest of the fast fashion at low prices msme are less competitive
baik menerapkan capsule wardrobe solusi kurang beli baju terus	both applying wardrobe capsule solutions to buy clothes
pengaruh selebritas tiktok selebritas instagram setiap hari racun baju	the influence of tiktok celebrities in instagram celebrities every day poison clothes.
usaha ultra fast fashion banned jangan kasih izin produksi	that's an ultra fast fashion banned business, don't give a production permit

D. Labeling

Sentiment labeling is performed on the translated_text column using VADER (Valence Aware Dictionary and sEntiment Reasoner). Label determination is based on compound scores, with thresholds of Positive ≥ 0.05, Negative ≤ -0.05, and Neutral in between.

Analysis of sentiment distribution across the entire dataset shows 654 data points (18.6%) classified as Negative, 1,648 data points (46.9%) as Neutral, and 1,211 data points (34.5%) as Positive, as presented in Figure 9. This distribution represents the diversity of public opinion toward fast fashion.

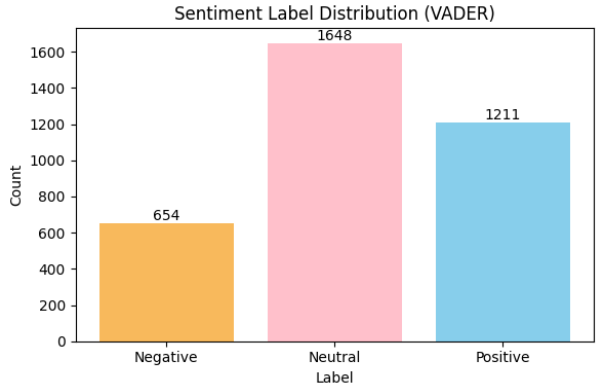

Figure 9. Sentiment Label Distribution (VADER)

Examples of labeling results are presented in Table 13, which illustrates how comments with informal language can be mapped into three sentiment categories based on compound scores.

TABLE XIII
LABELING RESULT

text	translated_text	compound	sentimen
normalisasi ulang outfit	normalization of outfit	0.0000	Neutral
susah payah edukasi luar sana orang belomba jual sisa fast fashion harga murah luar umkm banyak kalah saing	hard to education outside people who sell the rest of the fast fashion at low prices msmes are less competitive	-0.2716	Negative
baik menerapkan capsule wardrobe solusi kurang beli baju terus	both applying wardrobe capsule solutions to buy clothes	0.1779	Positive
pengaruh selebritas tiktok selebritas instagram setiap hari racun baju	the influence of tiktok celebrity celebrities instagram clothing poison	-0.5423	Negative
usaha ultra fast fashion banned jangan kasih izin produksi	ultra fast fashion banned business give production permit	-0.4588	Negative

To evaluate VADER's reliability, a stratified random sample of 1,747 comments (50%) was manually annotated by two independent raters. The comparison yielded an accuracy of 82.03% and Cohen's Kappa of 0.7155, which indicates substantial agreement (0.61-0.80 range) according to Landis and Koch's guidelines. These results validate VADER-generated labels as reliable ground truth for model training.

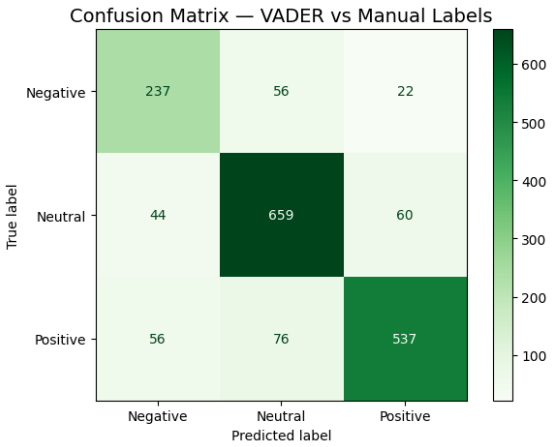

Figure 10. Confusion Matrix VADER vs Manual Labels

Figure 10 presents the confusion matrix between VADER and manual labels, showing strong agreement on the diagonal. Misclassifications primarily involve neutral comments being assigned sentiment polarity, reflecting the difficulty of distinguishing informational from evaluative content. Sample comparisons are shown in Table 14.

TABLE XIV
SAMPLE COMPARISON BETWEEN VADER AND MANUAL LABEL

Text Raw (Simplified)	VADER	Manual
hisabnya gimana KLO bajunya banyak bgt 😭	Negative	Neutral
itu krn modernisasi.. modernisasi itu menghasilkan limbah yg sulit didaur ulang.. dan orang2 Indonesia cenderung lebih gampang utk membuang.. blinisan ga? modernisasi tanpa pengetahuan lingkungan	Negative	Negative
Please normalisasi pake baju itu itu ajaa	Positive	Neutral

E. Label Encoding

Sentiment classes are converted into numerical form using label encoding with the scheme Negative = 0, Neutral = 1, Positive = 2. This process ensures that the data can be processed by classification algorithms while allowing for the interpretation of prediction results in terms of their original categories.

F. Dataset Split

The dataset is split into training data, consisting of 2,810 instances (80%), and testing data, consisting of 703 instances (20%), using stratified sampling to maintain class proportions in both subsets. The class distribution of each subset is presented in Table 15.

TABLE XV
CLASS DISTRIBUTION IN TRAINING AND TESTING DATA

Code	Sentiment Class	Training Data	Testing Data
0	Negative	523	131
1	Neutral	1318	330
2	Positive	969	242

The consistency of distribution between training and testing data confirms the effectiveness of stratified sampling in maintaining class representation. However, there is a class imbalance where the Neutral class dominates nearly half of the total data, followed by the Positive class, while the Negative class has the smallest number. This imbalance may potentially affect model performance in recognizing minority sentiments, particularly the Negative class, requiring special handling techniques during the model training process.

G. SVM

The SVM model with a linear kernel is implemented on the training data using a hybrid TF-IDF feature representation that combines words and characters. The word-based component uses unigrams with min_df=3, max_df=0.9, and sublinear TF scaling. The character-based component utilizes the char_wb analyzer, which employs 3-6 character n-grams to capture remaining linguistic patterns after preprocessing. Both components are combined using FeatureUnion to capture linguistic patterns at the word level as well as spelling variations characteristic of TikTok comments.

Class imbalance in the training data is handled using SMOTE. The initial distribution is presented in Table XVI. After applying SMOTE, each class becomes balanced with 1,318 comments as presented in Figure 11. This technique is used to reduce the dominance of majority classes and prevent bias toward them while maintaining a balanced data distribution.

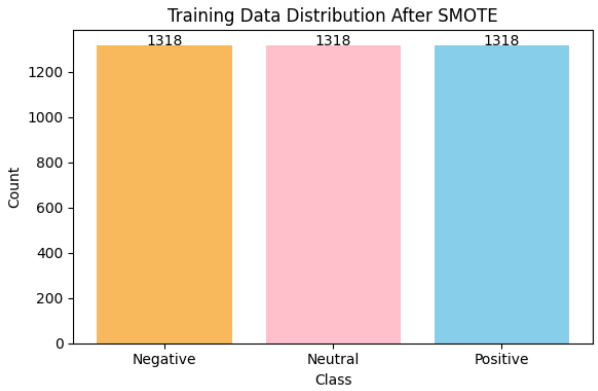

Figure 11. Training Data Distribution After SMOTE

Hyperparameter optimization is performed using GridSearchCV with stratified 7-fold cross-validation to find the optimal regularization C value. Eight candidate values are tested [0.05, 0.1, 0.5, 1, 2, 5, 10, 20] with F1-weighted as the evaluation metric.

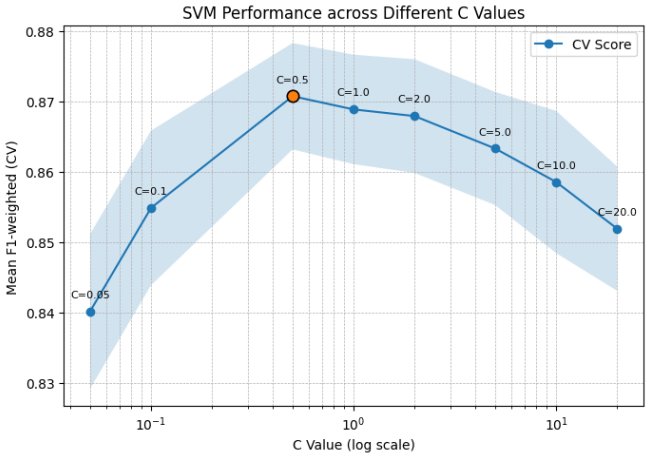

Figure 12. SVM Performance on C Value Variations

The optimization results are presented in Figure 12, which shows an inverted U-curve pattern. Low C values (0.05) result in an F1-weighted score of 0.840, indicating underfitting. Conversely, a high C value (20.0) decreases the score to 0.852 due to overfitting. The best performance is achieved at C=0.5 with an F1-weighted score of 0.871, demonstrating that moderate regularization provides an optimal equity between bias and variance for the TikTok comment dataset.

The model with parameter C = 0.5 is selected as the final configuration based on the most stable validation results, then utilized for sentiment classification across three polarity categories (Negative, Neutral, Positive) on the test data. Thus, the complete SVM modeling stages, including dataset splitting, TF-IDF feature representation, balancing with SMOTE, and hyperparameter optimization, result in a final model ready for comparison with the BERT-Base approach.

H. BERT

Class balancing in the BERT model is performed through synonym-based augmentation using the nlpaug library. This approach is chosen because BERT works directly on raw text through subword tokenization. Words in comments from minority classes are replaced with synonyms from WordNet to generate new sentence variations that preserve the sentiment meaning.

The data distribution before augmentation is presented in Table XVI. After augmentation, each class has a balanced distribution with 1,648 comments as presented in Figure 13. This process ensures that BERT training is not biased toward dominant classes, providing an equal opportunity for each sentiment category.

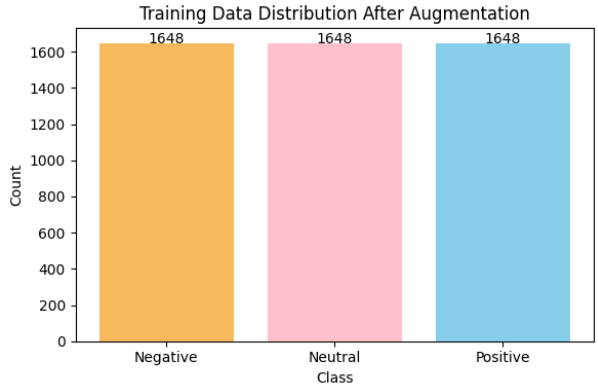

Figure 13. Training Data Distribution After Augmentation

The BERT-base-uncased tokenizer is used for text tokenization, with a maximum sequence length of 128 tokens. The BERT model is then trained using a batch size of 16, a learning rate of 6e-5, a weight decay of 0.01, and early stopping with a patience of 2 epochs based on the F1 score. The training process includes evaluation and model saving after each epoch, using the load_best_model_at_end function to select the model with the highest F1 score. The training progress is shown in the evaluation curve in Figure 14.

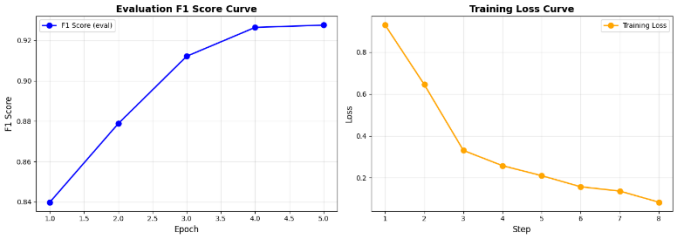

Figure 14. BERT Model F1-Score and Training Loss Curve

The F1-score curve on the evaluation data begins at 0.84 in the first epoch, rises to 0.88 in the second epoch, and continues to improve steadily, reaching 0.91 in the third epoch and 0.93 in the fourth epoch. The model then achieves its highest F1-score of 0.93 in the fifth epoch, indicating that performance gains begin to stabilize. This consistent upward trend demonstrates the model's ability to progressively learn more effective text representations throughout the training process.

I. Evaluation

The performance evaluation of both models is presented by confusion matrices in Figures 15 and 16.

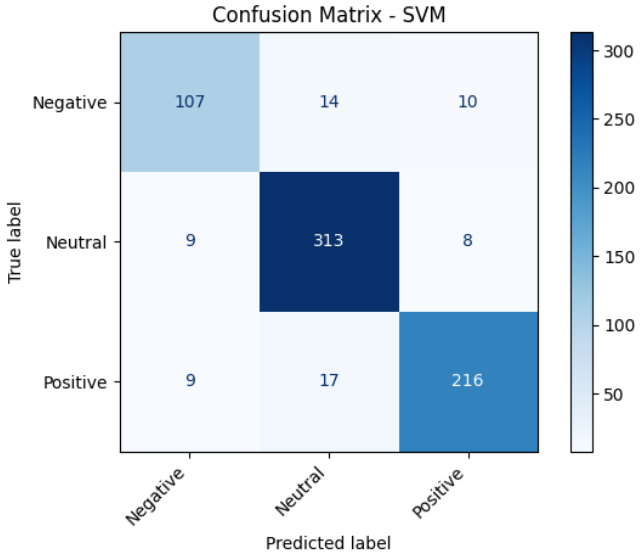

Figure 15. SVM Confusion Matrix

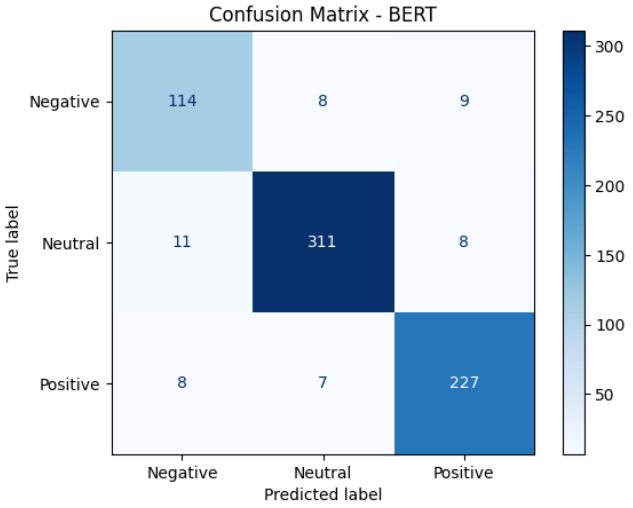

Figure 16. BERT Confusion Matrix

The SVM confusion matrix shows that the model correctly predicts 107 Negative, 313 Neutral, and 216 Positive samples. BERT yields higher performance with 114 correct Negative predictions, 311 Neutral, and 227 Positive, showing clear improvements, particularly for the Negative class. Based on these results, Table 16 summarizes the precision, recall, and F1-score for each sentiment category.

TABLE XVI
EVALUATION RESULT SUMMARY OF SVM AND BERT

Model	Class	Precision	Recall	F1-score
SVM	Negative	0.86	0.82	0.84
	Neutral	0.91	0.95	0.93
	Positive	0.92	0.89	0.91
BERT	Negative	0.86	0.87	0.86
	Neutral	0.95	0.93	0.94
	Positive	0.93	0.94	0.93

The evaluation results show that BERT performs more consistently across all sentiment classes, achieving an average F1-score of 0.91 and an overall accuracy of 92.7%. Meanwhile, SVM achieves an average F1-score of 0.89 with 90.4% accuracy. BERT also shows better performance in the negative class, obtaining a recall of 0.87, compared to SVM's 0.82. These results demonstrate BERT's stronger ability to capture contextual information for sentiment analysis, while SVM relies on lexical patterns derived from TF-IDF features.

The computational efficiency comparison between the two models is presented in Table 17.

TABLE XVII
COMPUTATIONAL EFFICIENCY COMPARISON

Model	Device	Train Time (s)	Inference Time (s)	Accuracy
SVM	CPU	25.9	0.051	0.904
BERT-Base	GPU (Tesla T4)	230.2	3.449	0.927

The results in Table 17 highlight a clear trade-off between performance and computational cost. BERT reaches 92.7% accuracy but requires substantially higher resources, with 230.2 seconds of GPU-based training and 3.449 seconds of inference time. In contrast, SVM operates efficiently on a CPU, completing training in 25.9 seconds and inference in 0.051 seconds. Although SVM is less accurate, its speed and minimal hardware requirements make it suitable for real-time monitoring and deployment in resource-constrained environments. The hybrid preprocessing pipeline further supports SVM by enhancing its ability to capture meaningful patterns from informal social-media text without relying on heavy contextual modeling, offering a practical alternative where GPU availability is limited.

The findings offer practical insights for sustainability stakeholders in the fashion industry. The dominance of neutral sentiment at 46.9% indicates that public awareness remains in an exploratory phase, presenting opportunities for targeted educational campaigns. Word frequency patterns reveal strong cultural influences such as Lebaran (Eid al-Fitr) on consumption behavior, suggesting that sustainability messaging should be adapted to local contexts and seasonal cycles. For greenwashing detection, sentiment analysis enables monitoring of discrepancies between brand sustainability claims and consumer responses, providing early warning signals when negative sentiment clusters around green marketing initiatives. From an operational perspective, organizations can deploy model-appropriate systems based on resource availability and analytical depth requirements. SVM provides cost-efficient continuous monitoring for real-time social listening across multiple platforms with minimal infrastructure, while BERT offers higher accuracy for in-depth periodic analysis such as quarterly sustainability assessments or campaign impact evaluation. This flexibility enables scalable implementation from small NGO initiatives to enterprise-level brand monitoring, supporting evidence-based strategies for promoting sustainable consumption culture through data-driven communication interventions tailored to Indonesian social media users.

IV. CONCLUSION

This research compares the performance of Support Vector Machine (SVM) and Bidirectional Encoder Representations from Transformers (BERT-Base) for sentiment analysis of 3,513 Indonesian TikTok comments on fast fashion using a hybrid preprocessing approach. The comments underwent normalization, slang cleaning using a 404-entry dictionary, and the handling of elongated and repetitive characters with the IndoNLP library to address informal linguistic patterns common in social media text.

BERT-Base achieves 92.7% accuracy with F1-scores of 0.86, 0.94, and 0.93 for negative, neutral, and positive classes. SVM attains 90.4% accuracy with F1-scores of 0.84, 0.93, and 0.91. BERT demonstrates stronger negative-sentiment detection, achieving a recall of 0.87 compared to SVM's 0.82,

reflecting its superior bidirectional contextual modeling. However, SVM's competitive performance indicates that hybrid preprocessing substantially narrows the gap between classical machine learning and transformer-based architectures.

Computational analysis shows clear resource trade-offs. SVM requires only 25.9 seconds of CPU training, far more efficient than BERT's 230.2 seconds of GPU processing and faster inference by a factor of 67.6×. Validation of automatic VADER labeling against 1,747 manually annotated samples achieved a Cohen's Kappa of 0.7155, confirming the reliability of the preprocessing and translation pipeline.

The sentiment distribution of 46.9% neutral, 34.5% positive, and 18.6% negative comments indicates that Indonesian society remains in the early stages of understanding sustainability issues. Organizations can select SVM for cost-efficient, real-time monitoring or BERT for in-depth, periodic analysis that requires nuanced sentiment detection. The effectiveness of hybrid preprocessing across both model paradigms confirms its essential role in social media sentiment analysis for low-resource languages. Future research should develop domain-adapted Indonesian models pretrained on social media corpora and explore lightweight architectures that leverage preprocessing-enhanced features to balance contextual understanding with computational efficiency for scalable sustainability monitoring.

REFERENCE

[1] F. Bonelli, R. Caferra, and P. Morone, "In need of a sustainable and just fashion industry: identifying challenges and opportunities through a systematic literature review in a Global North/Global South perspective," Discov. Sustain., vol. 5, no. 1, 2024, doi: 10.1007/s43621-024-00400-5.

[2] N. Olivar Aponte, J. Hernández Gómez, V. Torres Argüelles, and E. D. Smith, "Fast fashion consumption and its environmental impact: a literature review," Sustain. Sci. Pract. Policy, vol. 20, no. 1, p., 2024, doi: 10.1080/15487733.2024.2381871.

[3] United Nations Environment Programme, Catalysing Science-based Policy Action On Sustainable Consumption And Production: The value-chain approach & its application to food, construction and textiles. 2025. [Online]. Available: https://www.unep.org/resources/publication/catalysing-science-based-policy-action-sustainable-consumption-and-production

[4] M. Stenton, V. Kapsali, R. S. Blackburn, and J. A. Houghton, "From clothing rations to fast fashion: Utilising regenerated protein fibres to alleviate pressures on mass production," Energies, vol. 14, no. 18, pp. 1–18, 2021, doi: 10.3390/en14185654.

[5] European Parliament, "The impact of textile production and waste on the environment (infographics)." Accessed: Sep. 11, 2025. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20201208STO93327/the-impact-of-textile-production-and-waste-on-the-environment-infographics

[6] K. Bailey, A. Basu, and S. Sharma, "The Environmental Impacts of Fast Fashion on Water Quality: A Systematic Review," Water (Switzerland), vol. 14, no. 7, 2022, doi: 10.3390/w14071073.

[7] K. Khurana and S. S. Muthu, "Are low- and middle-income countries profiting from fast fashion?," J. Fash. Mark. Manag., vol. 26, no. 2, pp. 289–306, 2022, doi: 10.1108/JFMM-12-2020-0260.

[8] Y. Defrita Rufikasari, "Telaah Teologi, Ekonomi Dan Ekologi Terhadap Fenomena Fast Fashion Industry," J. Kepemimp. Kristen, Teol. dan Entrep., vol. 1, no. 2, pp. 64–83, 2023, doi: 10.61660/tep.v1i2.23.

[9] B. Ozbay, "Fast Fashion Market Report | Fashionbi," Fashionbi. Accessed: Sep. 11, 2025. [Online]. Available: http://fashionbi.com/market/fast-fashion/all

[10] T. Widari, Aliffianti, and M. Indra, "Fast fashion: Consumptive behavior in fashion industry Generation Z in Yogyakarta," IAS J. Localities, vol. 1, no. 2, pp. 104–113, 2023, doi: 10.62033/iasjol.v1i2.18.

[11] C. A Lin, X. Wang, and L. Dam, "TikTok Videos and Sustainable Apparel Behavior: Social Consciousness, Prior Consumption and Theory of Planned Behavior," Emerg. Media, vol. 1, no. 1, pp. 46–69, 2023, doi: 10.1177/27523543231188279.

[12] B. Zhong, J. Deng, and X. Liu, "Analyzing the influence of TikTok on sustainable choices: the moderating role of environmental consciousness," Acta Psychol. (Amst)., vol. 258, no. September 2024, p. 105182, 2025, doi: 10.1016/j.actpsy.2025.105182.

[13] D. El-Shihy and S. Awaad, "Leveraging social media for sustainable fashion: how brand and user-generated content influence Gen Z's purchase intentions," Futur. Bus. J., vol. 11, no. 1, 2025, doi: 10.1186/s43093-025-00529-3.

[14] Z. Cheng and Y. Li, "Like, Comment, and Share on TikTok: Exploring the Effect of Sentiment and Second-Person View on the User Engagement with TikTok News Videos," Soc. Sci. Comput. Rev., vol. 42, no. 1, pp. 201–223, 2024, doi: 10.1177/08944393231178603.

[15] M. He, C. Ma, and R. Wang, "A Data-Driven Approach for University Public Opinion Analysis and Its Applications," Appl. Sci., vol. 12, no. 18, 2022, doi: 10.3390/app12189136.

[16] H. M. Abiola, A. Iyanuoluwa, A. A. A., A. M. Gadafi, and A. Ishaq, "Tiktok Through AI Eyes: A Deep Learning Approach to Sentiment Analysis," Kwaghe Int. J. Eng. Inf. Technol., vol. 2, no. 2, pp. 57–77, 2025, doi: 10.58578/kijeit.v2i2.5485.

[17] V. Piccialli and M. Sciandrone, "Nonlinear optimization and support vector machines," Ann. Oper. Res., vol. 314, no. 1, pp. 15–47, 2022, doi: 10.1007/s10479-022-04655-x.

[18] T. Ahmed Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. Mohd Su'ud, "Sentiment Analysis using Support Vector Machine and Random Forest," J. Informatics Web Eng., vol. 3, no. 1, pp. 67–75, 2024, doi: 10.33093/jiwe.2024.3.1.5.

[19] Y. Song, X. Liu, and Z. Zhou, "A Comprehensive Review of Text Classification Algorithms," J. Electron. Inf. Sci., vol. 9, no. 2, pp. 34–42, 2024, doi: 10.23977/jeis.2024.090205.

[20] M. B. Sitepu, I. R. Munthe, and S. Z. Harahap, "Implementation of Support Vector Machine Algorithm for Shopee Customer Sentiment Anlysis," Sinkron, vol. 7, no. 2, pp. 619–627, 2022, doi: 10.33395/sinkron.v7i2.11408.

[21] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, pp. 534–539, 2022, doi: 10.14569/IJACSA.2022.0130665.

[22] A. Aitim, M. Abdulla, and A. Altayeva, "Sentiment Analysis Using Natural Language Processing," vol. 3567, 2024.

[23] K. Puh and M. Bagi, "Predicting sentiment and rating of tourist reviews using machine learning," vol. 6, no. 3, pp. 1188–1204, 2025, doi: 10.1108/JHTI-02-2022-0078.

[24] M. T. Stow, C. Ugwu, and L. N. Onyejegbu, "Improved Hybrid Model for Classification of Text Documents," vol. 2, no. 2, pp. 17–23, 2023.

[25] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using BERT," CEUR Workshop Proc., vol. 2841, 2021.

[26] A. Khan, D. Majumdar, and B. Mondal, "Sentiment analysis of emoji fused reviews using machine learning and Bert," Sci. Rep., vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-92286-0.

[27] M. P. Geetha and D. Karthika Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model," Int. J. Intell. Networks, vol. 2, no. July, pp. 64–69, 2021, doi: 10.1016/j.ijin.2021.06.005.

[28] Z. Yin et al., "DPG-LSTM: An Enhanced LSTM Framework for Sentiment Analysis in Social Media Text Based on Dependency Parsing and GCN," 2023.

[29] C. Raskoti and W. Li, "Exploring Transformer-Augmented LSTM for Temporal and Spatial Feature Learning in Trajectory Prediction," arXiv, 2024.
[30] W. Suwarningsih, R. A. Pratama, and F. Y. Rahadika, "RoBERTa : language modelling in building Indonesian question-answering systems Language modelling," vol. 20, no. 6, pp. 1248–1255, 2022, doi: 10.12928/TELKOMNIKA.v20i6.24248.
[31] A. F. Hidayatullah, R. Anna, A. Daphne, T. Ching, and L. Atika, "Pre - trained language model for code - mixed text in Indonesian , Javanese , and English using transformer," Soc. Netw. Anal. Min., 2025, doi: 10.1007/s13278-025-01444-9.
[32] M. Usman, M. Ahmad, M. Shahiki, I. Gelbukh, and R. Quintero, "Multilingual Hate Speech Detection in Social Media Using Translation-Based Approaches with Large Language Models," arXiv Prepr. arXiv2506.08147, 2025.
[33] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," vol. 2022, 2022, doi: 10.1155/2022/3498123.
[34] U. K. Das et al., "Enhancing sentiment analysis accuracy on social media comments using a tuned BERT model," Discov. Comput., vol. 28, no. 1, p. 198, 2025, doi: 10.1007/s10791-025-09599-x.
[35] A. Agrawal, S. Tripathi, M. Vardhan, V. Sihag, and G. Choudhary, "BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling," Appl. Sci., vol. 12, no. 3, 2022, doi: 10.3390/app12030976.
[36] M. A. Palomino and F. Aider, "Evaluating-the-Effectiveness-of-Text-PreProcessing-in-Sentiment-AnalysisApplied-Sciences-Switzerland.pdf," Mdpi, vol. 12, p. 8765, 2022.
[37] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review," RSF Conf. Ser. Eng. Technol., vol. 1, no. 1, pp. 363–367, 2021, doi: 10.31098/cset.v1i1.397.
[38] K. Makkar, P. Kumar, M. Poriye, and S. Aggarwal, "Improvisation in opinion mining using data preprocessing techniques based on consumer's review," Int. J. Adv. Technol. Eng. Explor., vol. 10, no. 99, pp. 258–278, 2023, doi: 10.19101/IJATEE.2021.875886.
[39] A. A. Aladeemy et al., "Advancements and challenges in Arabic sentiment analysis: A decade of methodologies, applications, and resource development," Heliyon, vol. 10, no. 21, p. e39786, 2024, doi: 10.1016/j.heliyon.2024.e39786.
[40] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar, and M. A. Shah, "Improving Sentiment Analysis in Social Media by Handling Lengthened Words," IEEE Access, vol. 11, no. December 2022, pp. 9775–9788, 2023, doi: 10.1109/ACCESS.2023.3238366.
[41] Hyuto, "indoNLP: Indonesian Natural Language Processing." [Online]. Available: https://hyuto.github.io/indo-nlp/
[42] M. M. Danyal, S. S. Khan, M. Khan, M. B. Ghaffar, B. Khan, and M. Arshad, "Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques," J. Big Data, vol. 5, no. September, pp. 1–18, 2023, doi: 10.32604/jbd.2023.041319.
[43] M. S. Mayaleh, S. A. Mayaleh, M. S. Mayaleh, S. A. Mayaleh, E. Sentiment, and S. Datasets, "Enhancing Sentiment Classification on Small Datasets through Data Augmentation and Transfer Learning : A Comparative Study To cite this version : HAL Id : hal-05090101 Enhancing Sentiment Classification on Small Datasets through Data Augmentation and Tran," pp. 0–16, 2025.
[44] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," IEEE Access, vol. 11, no. January, pp. 11811–11826, 2023, doi: 10.1109/ACCESS.2023.3242290.
[45] C. A. Cruz and F. Balahadia, "Analyzing Public Concern Responsesfor Formulating Ordinances and Lawsusing Sentiment Analysis through VADER Application," Int. J. Comput. Sci. Res., vol. 6, pp. 842–856, 2022, doi: 10.25147/ijcsr.2017.001.1.77.
[46] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, no. 1, pp. 159–174, Nov. 1977, doi: 10.2307/2529310.
[47] A. Vohra and R. Garg, "Deep learning based sentiment analysis of public perception of working from home through tweets," J. Intell. Inf. Syst., vol. 60, no. 1, pp. 255–274, 2023, doi: 10.1007/s10844-022-00736-2.
[48] M. A. Talukder et al., "A hybrid deep learning model for sentiment analysis of COVID-19 tweets with class balancing," Sci. Rep., vol. 15, no. 1, pp. 1–19, 2025, doi: 10.1038/s41598-025-97778-7.
[49] A. K. Durairaj and A. Chinnalagu, "Transformer based Contextual Model for Sentiment Analysis of Customer Reviews: A Fine-tuned BERT A Sequence Learning BERT Model for Sentiment Analysis," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 474–480, 2021, doi: 10.14569/IJACSA.2021.0121153.
[50] C. M. Greco and A. Tagarelli, Bringing order into the realm of Transformer-based language models for artificial intelligence and law, vol. 32, no. 4. Springer Netherlands, 2024. doi: 10.1007/s10506-023-09374-7.
[51] A. Areshey and H. Mathkour, "Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model," Sensors, vol. 23, no. 11, 2023, doi: 10.3390/s23115232.
[52] R. Obiedat et al., "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," IEEE Access, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.