JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7353

IMPROVING SENTIMENT ANALYSIS OF WOMEN IN STEM DISCOURSE USING SMOTE-ENHANCED SVM–VADER

Dwi Andini Putri[1*]; Siti Nurwahyuni[1]

Informatics, Faculty of Engineering and Informatics[1]
Universitas Bina Sarana Informatika, Jakarta, Indonesia[1]
www.bsi.ac.id[1]
dwi.dwd@bsi.ac.id*, siti.swu@bsi.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract*— The participation of women in Science, Technology, Engineering, and Mathematics (STEM) remains shaped by complex social and structural factors. This study investigates public sentiment regarding the role of technology in supporting women's participation in STEM through a machine learning–based sentiment analysis. Using 1,533 social media comments, sentiment classification was performed by integrating Support Vector Machine (SVM) and VADER-based automatic labeling, with imbalance handling to improve classification reliability. The results indicate a dominance of positive sentiment (98%), suggesting an optimistic tendency within the analyzed dataset, although this may be influenced by dataset characteristics and methodological bias. Among the evaluated models, a linear-kernel SVM achieved the highest accuracy (98.31%). This study contributes methodologically by demonstrating the effectiveness of integrating lexicon-based labeling with supervised learning for public sentiment analysis on gender equality in STEM, offering empirical insights to inform technology-driven policy interventions.*

Keywords*: Sentiment Analysis, SVM Kernels, Vader Lexicon, Women in STEM.*

Intisari— *Partisipasi perempuan dalam bidang Science, Technology, Engineering, and Mathematics (STEM) masih dipengaruhi oleh faktor sosial dan struktural yang kompleks. Penelitian ini bertujuan untuk mengkaji sentimen publik terhadap peran teknologi dalam mendukung partisipasi perempuan di bidang STEM melalui pendekatan analisis sentimen berbasis pembelajaran mesin. Dengan menggunakan 1.533 komentar dari media sosial, klasifikasi sentimen dilakukan melalui integrasi model Support Vector Machine (SVM) dan pelabelan otomatis berbasis leksikon VADER, disertai penanganan ketidakseimbangan data untuk meningkatkan keandalan klasifikasi. Hasil penelitian menunjukkan dominasi sentimen positif sebesar 98%, yang mengindikasikan adanya kecenderungan optimistis dalam dataset yang dianalisis, meskipun hal tersebut dapat dipengaruhi oleh karakteristik dataset dan bias metodologis. Di antara model yang dievaluasi, SVM dengan kernel linear mencapai tingkat akurasi tertinggi sebesar 98,31%. Penelitian ini memberikan kontribusi metodologis dengan menunjukkan efektivitas integrasi pelabelan berbasis lexikon dan supervised learning dalam analisis sentimen publik terkait kesetaraan gender di bidang STEM, serta menawarkan wawasan empiris untuk mendukung perumusan kebijakan berbasis teknologi.*

Kata Kunci*: Analisis Sentimen, Kernel SVM, Leksikon VADER, Perempuan dalam STEM.*

INTRODUCTION

The involvement of women in Science, Technology, Engineering, and Mathematics (STEM) remains a critical global issue. Despite various initiatives aimed at increasing women's participation, gender disparities persist across many countries [1],[2]. These disparities are influenced by factors such as gender stereotypes, limited female role models, and bias in recruitment and promotion processes [3]. Although women's participation in STEM has shown gradual improvement, structural and social barriers continue to restrict their advancement, particularly at professional and leadership levels [4].

STEM plays a central role in driving innovation and technological development; however, women's representation in this sector remains comparatively low. In Indonesia, women constituted only 40.6% of the STEM workforce in 2021, lagging behind Malaysia (48.6%) and Thailand (53.2%) [5]. Globally, women account for 49.3% of the non-STEM workforce but only 29.2% of STEM-related occupations, according to *The Global Gender Gap Report 2023* [6]. These figures indicate that policy interventions alone are insufficient and highlight the importance of understanding social factors and public perceptions surrounding women's roles in STEM [7].

Digital technology has been widely regarded as a potential enabler of gender equality through expanded access to education, technology-based training, and more inclusive work environments [1]. However, public acceptance of women's participation in STEM and the perceived role of technology in supporting gender equality remain underexplored. In this context, sentiment analysis provides a valuable approach for examining societal attitudes reflected in digital discourse.

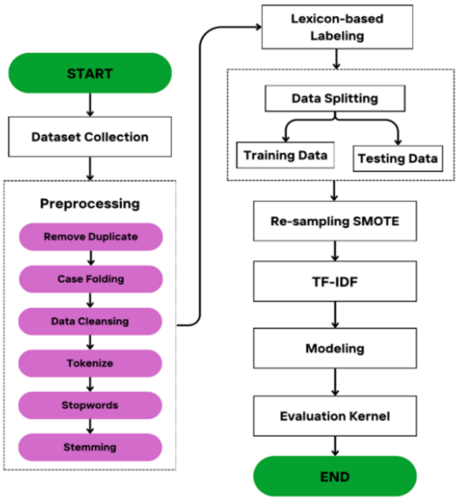
While sentiment analysis has been extensively applied to domains such as politics, customer experience, and education, relatively few studies have focused on gender representation in STEM, particularly using machine learning-based approaches [8][9][10]11]. Moreover, existing research rarely investigates how technology-related narratives shape public sentiment toward women in STEM. This gap underscores the need for computational, data-driven studies that capture public perceptions of gender equality in technology-driven fields.

To address this gap, this study employs sentiment analysis using the Support Vector Machine (SVM) algorithm, which has demonstrated strong performance in text classification tasks [12]. SVM distinguishes sentiment classes by identifying an optimal hyperplane, enabling accurate separation of positive and negative opinions [9]. To accommodate non-linear data patterns, multiple kernel functions—Linear, Radial Basis Function (RBF), Polynomial, and Sigmoid—are applied [13]. Additionally, this study addresses class imbalance in sentiment data using the Synthetic Minority Over-sampling Technique (SMOTE) [14], which enhances minority class representation and improves model performance [15]. The results are expected to contribute empirical evidence to AI-based sentiment analysis on gender equality and provide insights to support data-driven policy formulation aimed at strengthening women's long-term engagement in STEM.

Based on the theoretical framework and methodological design, this study hypothesizes that integrating the VADER lexicon for sentiment labeling, the SMOTE technique for addressing class imbalance, and the Support Vector Machine (SVM) algorithm for classification can enhance sentiment analysis performance in social issue contexts. Previous sentiment analysis studies on social issues have predominantly relied on either lexicon-based methods or single machine learning classifiers without explicitly addressing data imbalance or combining polarity-aware lexicons with supervised learning models. In contrast, the proposed approach combines rule-based sentiment labeling (VADER) with data balancing (SMOTE) and robust classification (SVM), enabling more reliable representation of public perceptions regarding the role of technology in supporting women's participation in STEM fields. This integrated framework is expected to offer improved accuracy and interpretability compared to conventional sentiment analysis approaches used in prior research.

MATERIALS AND METHODS

This study was conducted to analyze public sentiment regarding the issue of women's involvement in STEM by utilizing the Support Vector Machine (SVM) algorithm with Linear, RBF, Polynomial, and Sigmoid kernels. The methods employed include data collection, preprocessing, lexicon-based labeling, TF-IDF feature extraction, model development, and kernel evaluation.



Source: (Research Results, 2025)
Figure 1. Research Procedure

The research data were obtained from publicly accessible social media platforms without accessing private accounts or restricted content. This study did not collect or retain any personally identifiable information, and all user identities were anonymized during the preprocessing stage to ensure data privacy and confidentiality. As the study relied solely on publicly available data and did not involve direct interaction with research participants, formal ethical approval was not required in accordance with applicable institutional research ethics guidelines.

A. Dataset Collecting

This study uses data obtained from social media through web scraping techniques. The scraping process resulted in 1,533 public comments from social media users. These data were then used as the primary dataset to be analyzed in order to explore public sentiment regarding the issue of women's participation in STEM.

B. Preprocessing

At this stage, the dataset was processed through several steps to ensure data quality and to prevent potential issues during the training process [16]. The preprocessing steps were carried out as follows:

1. **Remove Duplicate.**

This step was performed to check the dataset for missing values or duplicate entries. Redundant or irrelevant data may affect the analysis results and therefore must be removed.

2. **Case Folding**

In this step, all letters were converted into lowercase. The purpose is to standardize the representation of words that are essentially the same but written in different formats, thereby improving consistency.

3. **Data Cleansing**

This process cleans the data by removing unnecessary elements such as hashtags (#), emoticons, URLs (e.g., www.), or certain symbols. Data cleansing is performed to make the dataset more structured and ready for analysis.

4. **Tokenization**

Tokenization splits text or sentences into the smallest units called tokens (words or phrases). These tokens are then used in the analysis process.

5. **Stopwords Removal**

At this stage, common words with no significant meaning, such as conjunctions or connectors, were removed. Eliminating stopwords allows the model to focus more on important words in sentiment analysis.

6. **Stemming**

The final step is stemming, which reduces words to their root form using the Sastrawi stemmer.

C. Lexicon-Based Labeling

At this stage, sentiment labeling was carried out using the VADER lexicon-based approach. Each text in the stemming column was analyzed using the polarity_scores() function from the Sentiment Intensity Analyzer to generate sentiment scores [17]. Among the results, the compound score was used to indicate the overall polarity of the sentence. If the compound score ≥ 0, the text was labeled as positive, whereas if the compound score < 0, it was labeled as negative. These scores and labels were then stored in new columns, namely *sentiment score* and *sentiment* [18]. In this way, each text that had gone through preprocessing and stemming could be automatically categorized as either a positive or negative opinion based on the VADER lexicon-based approach [19]. However, the labeling results should be interpreted with caution, as it remains unclear whether this phenomenon truly reflects public perception or is merely the result of sampling bias, given the predominance of positive sentiments.

Although alternative approaches such as other sentiment lexicons or deep learning-based models have shown strong performance in sentiment analysis, they often require large labeled datasets and higher computational resources. In contrast, VADER is specifically designed for short, informal social media text and incorporates linguistic features such as negation and intensity handling, making it a practical and interpretable choice for large-scale public sentiment analysis in social issue contexts.

D. Data Splitting

At this stage, the sentiment-labeled dataset was divided into two main parts: the training set and the testing set [20]. The training set was used to build and train the classification model, while the testing set was used to evaluate the model's performance on previously unseen data [21].

E. Re-sampling with SMOTE

This study applied the Synthetic Minority Over-sampling Technique (SMOTE) to address the issue of data imbalance. In imbalanced datasets, one class contains significantly fewer samples compared to the dominant class. Algorithm-based approaches typically adjust classification mechanisms to account for such conditions [22]. To mitigate this issue, a resampling strategy was employed using SMOTE oversampling, which is recognized as one of the most widely used techniques for enhancing the effectiveness of oversampling [23]. Accordingly, the application of SMOTE strengthened the model's ability to recognize the minority class, ultimately leading to more effective detection [14].

F. TF-IDF

At this stage, text feature extraction was carried out using the Term Frequency–Inverse Document Frequency (TF-IDF) method [24]. The text in the stemming column was transformed into a numerical representation so that it could be processed by machine learning algorithms [25]. This process employed the TF-IDF Vectorizer with an n-gram setting of (1,2) to capture both single words and two-word combinations. As a result, each document was represented in the form of word weights that indicate the level of importance of each term within the overall text corpus.

G. Modeling

This stage was carried out to develop a classification model using a data split ratio of 80:20 for training and testing. To address class imbalance in the training data, the Synthetic Minority Oversampling Technique (SMOTE) was applied prior to model training. The model's performance was then compared across four different kernel functions, namely Radial Basis Function (RBF), Linear, Polynomial, and Sigmoid.

H. Evaluation Kernel

he final stage involved evaluating the kernels used in this study, namely RBF, Linear, Sigmoid, and Polynomial. The evaluation was conducted using a confusion matrix by examining accuracy, precision, recall, and F1-score values. In addition, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were employed to provide a comprehensive assessment of classification performance. Although SMOTE was implemented to improve minority class representation and enhance classification performance, it is important to acknowledge its potential limitations. SMOTE generates synthetic samples based on existing data points, which may introduce noise or increase the risk of overfitting, particularly when the minority class distribution is complex. Consequently, while SMOTE can improve model performance, its impact on model generalization should be interpreted with caution. Recall, precision, and F-measure remain commonly used metrics for evaluating the performance of machine learning experiments.

RESULTS AND DISCUSSION

A. Preprocessing

The preprocessing stage was carried out to clean and transform raw text so that it could be processed and analyzed more effectively by machine learning algorithms. As shown in Table 1, the preprocessing steps in this study include removing duplicate data, case folding, cleaning special characters (text cleansing), tokenization, removing common meaningless words (stopword removal), and stemming. This process plays an essential role in improving the quality of input data, allowing the model to learn from clean and consistent text representations.

These results are in line with the findings of previous studies [1] showing that proper preprocessing can significantly improve the accuracy of text classifi cation through noise reduction and data redundancy. In addition, studies [2] report that the combination of stopword removal and stemming in social media data can improve classification performance by more than 10%. The findings of this study reinforce these results by showing that the application of preprocessing as a whole is particularly relevant in the context of informal and unstructured social media text data.

In the context of this study, the systematic implementation of preprocessing not only supports the effectiveness of the Support Vector Machine (SVM) model in recognizing sentiment patterns but also minimizes errors caused by linguistic variations and spelling inconsistencies commonly found in social media texts. Therefore, the preprocessing stage serves as a fundamental foundation to ensure more accurate and representative sentiment analysis results that truly reflect public opinion.

Table 1. Text Preprocessing Steps Applied to Social Media Data for Sentiment Analysis

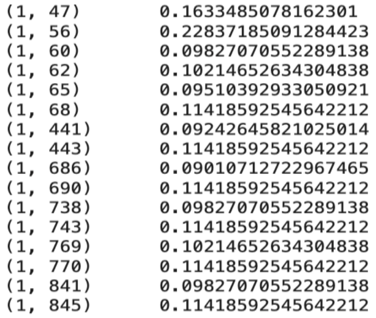
Data	Remove Duplicate	Case Folding	Cleansing	Tokenize	Stopword Removal	Stemming
@krisnanda_sean ...factory manager tempat kerja lama cewek,	@krisnanda_sean ...factory manager tempat kerja lama cewek,	@krisnanda_sean ...factory manager tempat kerja lama cewek,	factory manager tempat kerja lama cewek smart dan win	['factory', 'manager', 'tempat', 'kerja', 'lama', 'cewek', 'smart',	['factory', 'manager', 'kerja', 'cewek', 'smart', 'menang',	factory manager kerja cewek smart menang solutif

Data	Remove Duplicate	Case Folding	Cleansing	Tokenize	Stopword Removal	Stemming
smart dan win solutif bangat malah. Menurut saya kembali ke personal masing2 si	smart dan win solutif bangat malah. menurut saya kembali ke personal masing2 si	smart dan win solutif bangat malah. menurut saya kembali ke personal masing2 si	solutif bangat malah menurut saya kembali ke personal masing si	'menang', 'solutif', 'bangat', 'malah', 'menurut', 'saya', 'kembali', 'personal', 'masing']	'solutif', 'bangat', 'personal']	bangat personal
@brooobrown bagaiman kita bisa setuju dengan respon? Kondisinya, saya berstatmen berdasarkan pengetahuan saya. Dan saya diawal belum menunjukkan sumber secara detail, bukan berarti tdk memiliki, dan beddy juga tdk memiliki sumber yg dpt membantah statment saya (terkait perbedaan terletak pada kemampuan dlm memberdayagunakan nya loh, bukan kemampuan otak berpikir logisnya) Ã°ÂŸÂ™Â□	@brooobrown bagaiman kita bisa setuju dengan respon? kondisinya, saya berstatmen berdasarkan pengetahuan saya. dan saya diawal belum menunjukkan sumber secara detail, bukan berarti tdk memiliki, dan beddy juga tdk memiliki sumber yg dpt membantah statment saya (terkait perbedaan terletak pada kemampuan dlm memberdayagunakan nya loh, bukan kemampuan otak berpikir logisnya) Ã°ÂŸÂ™Â□	@brooobrown bagaiman kita bisa setuju dengan respon? kondisinya, saya berstatmen berdasarkan pengetahuan saya. dan saya diawal belum menunjukkan sumber secara detail, bukan berarti tdk memiliki, dan beddy juga tdk memiliki sumber yg dpt membantah statment saya (terkait perbedaan terletak pada kemampuan dlm memberdayagunakan nya loh, bukan kemampuan otak berpikir logisnya) Ã°ÂŸÂ™Â□	bagaiman kita bisa setuju dengan respon kondisinya saya berstatmen berdasarkan pengetahuan saya dan saya diawal belum menunjukkan sumber secara detail bukan berarti tdk memiliki dan beddy juga tdk memiliki sumber yg dpt membantah statment saya terkait perbedaan terletak pada kemampuan dlm memberdayagunakan nya loh bukan kemampuan otak berpikir logisnya	['bagaiman', 'kita', 'bisa', 'setuju', 'dengan', 'respon', 'kondisinya', 'saya', 'berstatmen', 'berdasarkan', 'pengetahuan', 'saya', 'saya', 'diawal', 'belum', 'menunjukkan', 'sumber', 'adalah', 'detail', 'bukan', 'berarti', 'tidak', 'memiliki', 'beddy', 'juga', 'tidak', 'memiliki', 'sumber', 'yang', 'dapat', 'membantah', 'statment', 'saya', 'terkait', 'perbedaan', 'terletak', 'pada', 'kemampuan', 'dalam', 'memberdayagunakan', 'bukan', 'kemampuan', 'otak', 'berpikir', 'logisnya']	['bagaiman', 'setuju', 'respon', 'kondisinya', 'berstatmen', 'berdasarkan', 'pengetahuan', 'diawal', 'sumber', 'detail', 'memiliki', 'beddy', 'memiliki', 'sumber', 'membantah', 'statment', 'terkait', 'perbedaan', 'terletak', 'kemampuan', 'memberdayagunakan', 'kemampuan', 'otak', 'berpikir', 'logisnya']	bagaiman tuju respon kondisi berstatmen dasar tahu awal sumber detail milik beddy milik sumber ban statment kait beda letak mampu memberdayagunakan mampu otak pikir logis

Source : (Research Results, 2025)

B. TF-IDF

The use of the TF-IDF feature extraction technique applied in this study is beneficial for identifying important words in a document and supporting the text analysis process[26]. TF-IDF was used to convert raw textual data into a numerical representation suitable for processing by machine learning algorithms, particularly the Support Vector Machine (SVM) model. Figure 1 illustrates the role of the TF-IDF feature extraction stage within the sentiment analysis pipeline, where preprocessed textual data are transformed into numerical feature vectors that serve as input for the SVM classification model. This visualization displays the structure of the TF-IDF matrix, where each index pair and numerical weight represents the contribution of a specific word to a given document. The figure plays a key role in explaining how raw text data are transformed into numerical feature vectors that serve as the primary input for the SVM model.



```
(1, 47)       0.1633485078162301
(1, 56)       0.22837185091284423
(1, 60)       0.09827070552289138
(1, 62)       0.10214652634304838
(1, 65)       0.09510392933050921
(1, 68)       0.11418592545642212
(1, 441)      0.09242645821025014
(1, 443)      0.11418592545642212
(1, 686)      0.09010712722967465
(1, 690)      0.11418592545642212
(1, 738)      0.09827070552289138
(1, 743)      0.11418592545642212
(1, 769)      0.10214652634304838
(1, 770)      0.11418592545642212
(1, 841)      0.09827070552289138
(1, 845)      0.11418592545642212
```

Source : (Research Results, 2025)

Figure 1. TF-IDF Feature Extraction Stage in the Sentiment Analysis

Figure 1 Pipeline Showing the Transformation of Preprocessed Text into Numerical Feature Vectors for SVM Classification.

Parts such as *(1, 56)* indicate a position in the TF-IDF matrix, which means:

a) The first number *(1)* represents the document index (e.g., the 1st document).
b) The second number *(56)* represents the word index (the 56th feature in the vocabulary generated by TF-IDF).

The decimal value on the right (e.g., *0.22837185091284423*)is the TF-IDF weight of that word in the 1st document. The higher the value, the more important the word is for that particular document compared to other documents.

C. Labeling Results with the VADER Method

Classification using the VADER lexicon produced 1,501 positive reviews and 31 negative reviews. The labeling process conducted with the VADER lexicon showed that 98% of the reviews were categorized as positive, while 2% were categorized as negative. Table 2 presents the sentiment scoring results generated by the VADER lexicon, which include negative scores, positive compound scores, and polarity values. On the other hand, this imbalance may also be influenced by methodological limitations associated with the use of the VADER lexicon in an Indonesian-language context. As VADER was originally developed for English-language social media, it may not fully capture linguistic nuances, implicit negativity, sarcasm, or culturally specific expressions in Indonesian text, potentially leading to an overestimation of positive sentiment. Therefore, the high proportion of positive labels should be interpreted cautiously, as it may represent a combination of actual public sentiment and lexicon-induced bias rather than an entirely accurate reflection of sentiment polarity.

Table 2. Sentiment Labeling Results Using the VADER Lexicon (Representative Samples Showing Polarity Scores and Class Distribution Before SMOTE)

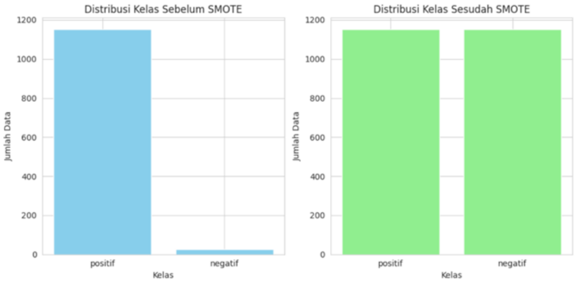
No	Sample Text (Excerpt)	Sentiment Score (Compound)	Polarity Classification
963	@kris****_s*** ...factory manager tempat kerja lama cewek, smart dan win solutif bangat malah. Menurut saya kembali ke personal masing2 si	0.4019	positif
1071	@bro*******bagaiman kita bisa setuju dengan respon? Kondisinya, saya berstatmen berdasarkan pengetahuan saya. Dan saya diawal belum menunjukkan sumber secara detail, bukan berarti tdk memiliki, dan beddy juga tdk memiliki sumber yg dpt membantah statment saya (terkait perbedaan terletak pada kemampuan dlm memberdayagunakan nya loh, bukan kemampuan otak berpikir logisnya) Ã°ÂÂÂ	-0.5574	negatif

Source : (Research Results, 2025)

D. Re-Sampling with SMOTE

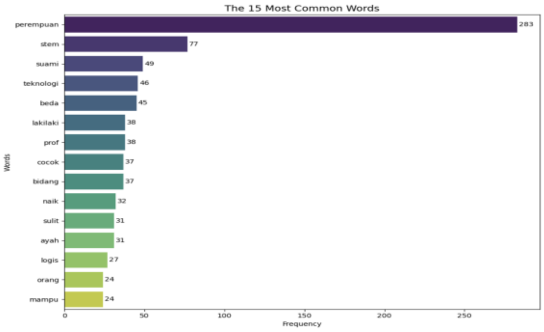
Figure 2 presents a comparison of class distribution before and after applying the Synthetic Minority Over-sampling Technique (SMOTE). In the left graph (Class Distribution Before SMOTE), the positive class dominates with 1,153 samples, while the negative class contains only 25 samples. This imbalance indicates a skewed dataset that may affect the performance of machine learning models, as classifiers tend to favor the majority class. While the application of balancing techniques aims to reduce such bias, the overwhelming dominance of positive sentiment (98%) requires critical examination. This distribution may reflect genuinely favorable public attitudes toward women's participation in STEM. However, it may also represent a methodological artifact arising from data collection strategies or the limitations of lexicon-based sentiment labeling in capturing nuanced or context-dependent expressions. Therefore, the observed sentiment distribution should be interpreted cautiously, as it likely reflects an interaction between real-world social dynamics and methodological constraints.

This could be due to the nature of social media platforms, where users tend to post more positive comments about women's involvement in STEM. According to [30], the prevalence of positive comments on such platforms is a common phenomenon, while [31] argues that positive representations of women on social media and STEM-related campaigns can enhance public perceptions of women in STEM fields. In the right graph (Class Distribution After SMOTE), the number of samples in the negative class was increased by generating synthetic data through SMOTE. As a result, both positive and negative classes became balanced, each with 1,153 samples. With this balanced distribution, the machine learning model can learn more effectively without bias toward one class.



Source: (Research Results, 2025)

Figure 2. Comparison of Class Distribution Before and After Applying SMOTE



Source: (Research Results, 2025)

Figure 3. Frequency of word occurrences in the research dataset

Source: (Research Results, 2025)

Figure 4. Word Cloud of Common Themes and Dominant Keywords

The results of the word frequency analysis indicate that discussions about women in STEM are dominated by the terms "women" (283 occurrences) and "STEM" (77 occurrences), highlighting that gender-related issues constitute the primary focus of public discourse. The appearance of words such as "husband" (49), "father" (31), and "men" (38) suggests a strong association between domestic roles and societal perceptions of women's participation in STEM.

These findings are consistent with Gender Role Theory, which posits that gender-based social expectations continue to shape public views regarding women's roles in professional domains. Furthermore, terms such as "difficult" (31), "different" (45), and "suitable" (37) reflect normative evaluations of women's abilities and perceived suitability for technical fields, which may contribute to the emergence of stereotype threat. This condition has the potential to undermine women's confidence and performance as a result of the internalization of negative stereotypes.

Therefore, this sentiment analysis not only reflects public opinion but also underscores the importance of policy- and education-based interventions aimed at reducing gender bias and fostering a more inclusive and supportive STEM environment for women.

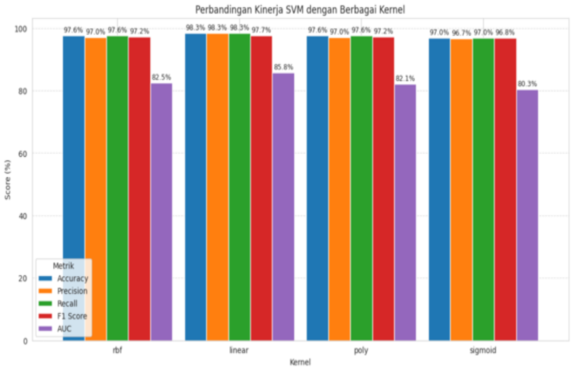
Evaluation Results of SVM and Kernels

The best performance of the SVM model with the applied kernels can be seen in Table 3 below.

Table 3. Summary of SVM Kernel Comparison

	kernel	Accuracy	Precision	Recall	F1 Score	AUC
0	rbf	97.63	96.97	97.63	97.24	82.53
1	Linear	98.31	98.33	98.31	97.71	85.81
2	Poly	97.63	96.97	97.63	97.24	82.06
3	sigmoid	96.95	96.68	96.95	96.81	80.33

Source: (Research Results, 2025)



Source: (Research Results, 2025)

Figure 4. Comparison of SVM Models Based on Kernels

Table 3 and Figure 4 present a comparison of the performance of Support Vector Machine (SVM) models with four types of kernels, namely Radial base Function (RBF), Linear, Polynomial, and Sigmoid. Figure 4 visually shows a comparison of five key evaluation metrics-Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC) for each kernel, making it easier to analyze performance differences between models.

The table shows the evaluation results based on five key classification metrics: Accuracy, Precision, Recall, F1-Score, and AUC (Area Under the Curve). From the results, the Linear kernel achieved the best overall performance, with the highest accuracy of 98.31%, precision of 98.33%, recall of 98.31%, F1-Score of 97.71%, and an AUC of 85.81%. Comparatively, these findings align with prior studies such as [27] and [28], which highlight that the Linear kernel performs exceptionally well in text classification due to the inherently linear structure of TF-IDF-based feature vectors. Similarly, [29] demonstrated that Linear SVM models outperform RBF-based ones in large-scale social media sentiment datasets, where textual data often exhibit linearly separable patterns.

The superior performance of the Linear kernel in this study can be attributed to the characteristics of the dataset, which consists of high-dimensional, sparse TF-IDF representations derived from textual data. In such feature spaces, sentiment-related patterns are often distributed in a manner that allows classes to be separated using linear decision boundaries. As a result, the Linear kernel can effectively capture discriminative features without introducing unnecessary model complexity, leading to better generalization and more stable performance. However, in contrast to [30], which reported superior RBF kernel performance for datasets containing high semantic variability, this study found that both RBF and Polynomial kernels achieved stable yet lower AUC values (82.53% and 82.06%, respectively).

This indicates that while non-linear kernels are capable of modeling complex relationships, they may be less effective when the feature space already encodes sufficient discriminatory information, as is the case with TF-IDF-based sentiment features. In such scenarios, non-linear transformations can introduce overfitting or reduce generalization performance. The Sigmoid kernel recorded the lowest accuracy (96.95%), consistent with [31], which noted that this kernel tends to be unstable in high-dimensional text data due to its sensitivity to parameter scaling. Overall, these results reinforce the argument in previous literature that the Linear kernel is the most optimal choice for text-based sentiment analysis. Its effectiveness stems from its ability to handle linearly separable TF-IDF features efficiently while maintaining high interpretability and computational scalability.

CONCLUSION

The Polynomial kernel recorded an accuracy of 97.63%, precision of 96.97%, F1-score of 97.24%, and an AUC of 82.06%, while the Sigmoid kernel achieved an accuracy of 96.95%, precision of 96.68%, F1-score of 96.81%, and an AUC of 80.33%. The best performance was obtained using the Linear kernel, with an accuracy of 98.31%, precision of 98.33%, F1-score of 97.71%, and an AUC of 85.81%, indicating that kernel selection in SVM influences sentiment classification performance. Sentiment labeling using the VADER Lexicon showed a predominance of positive sentiment; however, this result may be affected by the domain-dependent nature of lexicon-based approaches. In addition, although SMOTE effectively addressed class imbalance by generating synthetic samples, it may introduce a risk of overfitting, which should be considered when interpreting the results.

Beyond confirming the effectiveness of sentiment analysis as an analytical tool, the findings of this study provide meaningful implications for policy and practice aimed at supporting women's participation in STEM. The dominance of positive sentiment suggests a growing level of societal acceptance and openness toward women's roles in science and technology. This insight can be utilized by policymakers, educational institutions, and non-governmental organizations to design evidence-based interventions, such as targeted STEM outreach programs for young women, inclusive curricula, public awareness campaigns, and scholarship schemes that address remaining barriers to participation.

Moreover, sentiment analysis can function as a continuous monitoring instrument to evaluate the impact of gender equality initiatives and public campaigns over time. By identifying shifts in public perception as well as areas where negative sentiment persists, stakeholders can implement more focused and adaptive strategies, including mentoring programs, promotion of female STEM role models, and community-based engagement initiatives.

Future research is encouraged to expand this work by incorporating larger and more diverse datasets, comparing SVM with alternative machine learning and deep learning approaches, and exploring longitudinal sentiment trends. Such efforts would further enhance the role of data-

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

driven insights in informing sustainable policies and interventions that strengthen women's empowerment and long-term engagement in STEM fields.

REFERENCE

[1] A. Suryaningsih And A. H. Sanjaya, "Pemberdayaan Perempuan Dalam Mewujudkan Kesetaraan Gender: Strategi Dan Tantangan Di Era Globalisasi," *Jurnal Pendidikan Sejarah Dan Riset Sosial Humaniora*, Vol. 4, No. 2, Pp. 2621–119, 2024.

[2] C. Dwi Anggola, F. Prawita, And D. Putri Lestarika, "Peran Pendidikan Dalam Mengurangi Kesenjangan Gender Di Tempat Kerja," Vol. 02, No. 1, Pp. 531–537, 2024, [Online]. Available: Https://Jurnal.Kopusindo.Com/Index.Php/Jkhkp

[3] Amelia, R. N., Mafikah, A. D., and Rif'ah, S., "Kesetaraan Gender dalam Manajemen Sumber Daya Insani: Tantangan dan Peluang," EQUALITY: Journal of Gender, Child, and Humanity Studies, vol. 2, no. 1, pp. 30–40, 2024.

[4] F. Hotman, S. Damanik, O. Sukmana, And W. Winarjo, "Sosiologi Kritis Dan Transformasi Pendidikan: Menggugat Ketidaksetaraan Gender Di Indonesia," 2025. [Online]. Available: Https://Jurnaldidaktika.Org2031

[5] East.Vc, "Hari Perempuan Sedunia: Menyoroti Kontribusi Perempuan Di Bidang Stem," East.Vc. Accessed: Mar. 22, 2025. [Online]. Available: Https://East.Vc/Id/Berita/Insights-Id/Hari-Perempuan-Sedunia-Perempuan-Stem/

[6] Word Economic Forum, "Global Gender Gap Report 2023," Jun. 2023. Accessed: Mar. 22, 2025. [Online]. Available: Https://Www.Weforum.Org/Publications/Global-Gender-Gap-Report-2023/

[7] L. Sonia And K. Sassi, "Menjelajahi Kesenjangan Gender Dalam Pendidikan: Studi Perbandingan Antara Swedia Dan Afghanistan," Vol. 5, No. 4, Nov. 2024, [Online]. Available: Https://Ejurnals.Com/Ojs/Index.Php/

[8] A. Permata, "Analisis Sentimen Media Sosial: Mengurai Opini Publik Dengan Data," Teknologipintar.Org, Vol. 4, No. 3, Pp. 2024–2025, 2024.

[9] D. Andini Putri And D. Ayu Muthia, "Implementasi Metode Lexicon Based Dan Support Vector Machine Pada Analisis Sentimen Ulasan Pengguna Chatgpt," Ijcit (Indonesian Journal On Computer And Information Technology), Vol. 9, No. 2, Pp. 80–86, 2024.

[10] L. Geni, E. Yulianti, And D. I. Sensuse, "Sentiment Analysis Of Tweets Before The 2024 Elections In Indonesia Using Bert Language Models," *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, Vol. 9, No. 3, Pp. 746–757, Aug. 2023, Doi: 10.26555/Jiteki.V9i3.26490.

[11] S. Mariam And I. Nurhaida, "Edumatic: Jurnal Pendidikan Informatika Analisis Sentimen Berbasis Deep Learning Terhadap Kesetaraan Gender Di Bidang Stem: Perspektif Dan Implikasinya," Vol. 9, No. 1, Pp. 69–78, 2025, Doi: 10.29408/Edumatic.V9i1.29071.

[12] A. Saepudin Et Al., "Analisis Sentimen Pemanfaatan Artificial Intelligence Di Dunia Pendidikan Menggunakan Svm Berbasis Particle Swarm Optimization," 2024. [Online]. Available: Http://Jurnal.Bsi.Ac.Id/Index.Php/Co-Science

[13] S. Ernawati And R. Wati, "Evaluasi Performa Kernel Svm Dalam Analisis Sentimen Review Aplikasi Chatgpt Menggunakan Hyperparameter Dan Vader Lexicon," 2024.

[14] M. Ibnu Choldun Rachmatullah And S. Armiati, "Menerapkan Smote Pada Klasifikasi Data Penyakit Stroke," Vol. 17, No. 1, 2025.

[15] F. S. Pratiwi, M. Agung Barata, And A. D. Ardianti, "Implementasi Metode Smote Dan Random Over-Sampling Pada Algoritma Machine Learning Untuk Prediksi Customer Churn Di Sektor Perbankan," Jurnal Sistem Informasi Dan Informatika (Simika), Vol. 8, No. 1, 2025, [Online]. Available: Https://Www.Kaggle.Com/Datasets/Gauravtopre/Bank-Customer-Churn-Dataset/Data

[16] F. Dewi, N. Cahyo, H. Wibowo, M. R. Handayani, And K. Umam, "Evaluasi Hyperparamter Tuning Pada Support Vector Machine (Svm) Dalam Klasifikasi Ulasan Hotel Di Tripadvisor," Vol. 10, No. 3, Pp. 2584–2593, 2025, Doi: 10.29100/Jipi.V10i3.7774.

[17] V. Renedominick And S. Barus, "Analisis Sentimen Pada Trailer Deadpool Vs Wolverine Menggunakan Model Machine Learning," Jurnal Pustaka Ai (Pusat Akses Kajian Teknologi Artificial Intelligence), Vol. 5, No. 1, Pp. 01–06, Apr. 2025, Doi: 10.55382/Jurnalpustakaai.V5i1.892.

[18] Utari, E. L. and Wibowo, S. H., "Analisis Komparatif Algoritma SVM, Naive Bayes, dan

LSTM pada Sentimen Komentar Lagu Labour," Jurnal Informatika Teknologi dan Sains (Jinteks), vol. 7, no. 3, pp. 1276–1286, 2025.

[19] N. Fauziah, "Analisis Sentimen Publik Terhadap Kenaikan Tarif Ppn Di Indonesia Dengan Pendekatan Vader," Jurnal Akuntansi Dan Keuangan, Vol. 12, No. 2, P. 228, Sep. 2024, Doi: 10.29103/Jak.V12i2.16796.

[20] D. Nasien Et Al., "Perbandingan Implementasi Machine Learning Menggunakan Metode Knn, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," 2024.

[21] A. R. Hanum Et Al., "Analisis Kinerja Algoritma Klasifikasi Teks Bert Dalam Mendeteksi Berita Hoaks," Vol. 11, No. 3, Pp. 537–546, 2024, Doi: 10.25126/Jtiik938093.

[22] Hizbul Izzi, Arief Setyanto, And Anggit Dwi Hartanto, "Optimalisasi Akurasi Algoritma Naïve Bayes Dengan Metode Syntetic Minority Oversampling Technique (Smote) Pada Data Numerik," Infotek: Jurnal Informatika Dan Teknologi, Vol. 8, No. 1, Pp. 217–227, Jan. 2025, Doi: 10.29408/Jit.V8i1.28340.

[23] I. Maulana And S. Ernawati, "Meningkatkan Klasifikasi Penyakit Diabetes Menggunakan Metode Ensemble Softvoting Dengan Smote-Enn Dan Optimasi Bayesian," Jurnal Sains Dan Manajemen, Vol. 13, No. 1, 2025.

[24] K. Tri Putra, S. Anggraini, L. Sutriani, A. Impron, And J. Informatika, "Analisis Sentimen Masyarakat Kalimantan Tengah Terhadap Perkebunan Kelapa Sawit Menggunakan Tf-Idf Dan Support Vector Machine," 2025.

[25] E. Rifut Nur Mustaqim, U. Pagalay, And C. Crysdian, "Prediksi Tingkat Kepercayaan Masyarakat Terhadap Pilpres 2024 Menggunakan Tf-Idf Dan Bow Menggunakan Metode Svm."

[26] T. Baskoro And S. R. Nuddin, "Analisa Kinerja Chatgpt Dalam Menghasilkan Teks Bahasa Indonesia Menggunakan Metode Support Vector Machines (Svm)," Journal Of Informatics And Computer Science, Vol. 06, 2024.

[27] M. A. R. N. M. Celine Mutiara Putri, "Perbandingan Evaluasi Kernel Support Vector Machine dalam Analisis Sentimen Chatbot AI pada Ulasan Google Play Store," Jurnal Teknologi Sistem Informasi dan Aplikasi, vol. 7, Jul. 2024.

[28] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 375–380, Oct. 2019, doi: 10.22219/kinetik.v4i4.912.

[29] T. Hevianto Saputro and A. Hermawan, "The Accuracy Improvement of Text Mining Classification on Hospital Review through The Alteration in The Preprocessing Stage," 2021. [Online]. Available: www.ijcit.com140

[30] M. A. Rosulan and R. Rosli, "Key Dimensions and Impact Factors on STEM Identity Among Female Students: A Systematic Literature Review", doi: 10.47772/IJRISS.

[31] M. Stella, "Text-mining forma mentis networks reconstruct public perception of the STEM gender gap in social media," Mar. 2020, doi: 10.7717/peerj-cs.295.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri