Yoseph Tajul Arifin

7353-27890-1-5-20251028

Cakrawala

Document Details

Submission ID
trn:oid:::3618:138805637

Submission Date
Nov 3, 2025, 5:23 PM GMT+7

Download Date
Nov 3, 2025, 5:47 PM GMT+7

File Name
7353-27890-1-5-20251028.docx

File Size
549.5 KB

11 Pages

3,873 Words

23,336 Characters

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

- **52** Not Cited or Quoted 17%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

14% Internet sources
9% Publications
11% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **52** Not Cited or Quoted 17%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

14% Internet sources
9% Publications
11% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	Universitas Bina Sarana Informatika on 2026-01-19 — 1%
2	Student papers	UNIVERSITAS BUDI LUHUR on 2025-12-16 — <1%
3	Student papers	Universitas Pendidikan Ganesha on 2025-12-15 — <1%
4	Student papers	University of Hertfordshire on 2024-05-02 — <1%
5	Publication	Min Hao, Xingtai Cao, Jianying Sun, Yupeng Sun, Jiaxuan Wang, Hao Zhang. "Dete... — <1%
6	Internet	ejournal.kresnamediapublisher.com — <1%
7	Internet	repository.lpkia.ac.id — <1%
8	Publication	Jieling Jin, Helai Huang, Rui Zhou. "Determinants of Autonomous Vehicle Crashes ... — <1%
9	Internet	peerj.com — <1%
10	Internet	www.ncbi.nlm.nih.gov — <1%

11 Internet

par.nsf.gov <1%

12 Internet

public-pages-files-2025.frontiersin.org <1%

13 Internet

alicia.concytec.gob.pe <1%

14 Publication

Danita Divka Sajmira, Khothibul Umam, Maya Rini Handayani. "Enhancing Review... <1%

15 Student papers

UNIVERSITAS BUDI LUHUR on 2025-12-16 <1%

16 Publication

Wail M. Idress, Yuqian Zhao, Khalid A. Abouda, Hiba M. Elhag. "QCNN-Swin-UNet: ... <1%

17 Publication

Mohammad Zubair Khan, Abdulhakim Sabur, Hamza Ghandorh. "A Novel Interne... <1%

18 Internet

eprints.amikompurwokerto.ac.id <1%

19 Internet

jurnal.stkippgritulungagung.ac.id <1%

20 Internet

researchonline.gcu.ac.uk <1%

21 Publication

Anjum Gupta, Shibin Parameswaran, Cheng-Han Lee. "Classification of electroen... <1%

22 Student papers

Baylor University on 2026-04-17 <1%

23 Internet

fti.ars.ac.id <1%

24 Student papers

Higher Education Commission Pakistan on 2022-10-10 <1%

25 Student papers

IUBH - Internationale Hochschule Bad Honnef-Bonn on 2023-10-11 <1%

26 Publication

Siding Li, Hua Wen, Chuyi Peng, Chunyang Ma. "Analysis of light field and flow fiel... <1%

27 Publication

Vannes Wijaya, Nur Rachmat. "Comparison of SVM, Random Forest, and Logistic ... <1%

28 Internet

getmorc.com <1%

29 Internet

repository.ipb.ac.id <1%

30 Publication

Kharisma Kharisma, Irmma Dwijayanti, Ulfi Saidata Aesyi, Alfirna Rizqi Lahitani. "... <1%

31 Publication

Muhammad Mursil, Hatem A. Rashwan, Adnan Khalid, Pere Cavallé-Busquets, Lui... <1%

32 Internet

ejurnal.stmik-budidarma.ac.id <1%

33 Internet

nendensan.web.id <1%

34 Internet

www.spmvv.ac.in <1%

35 Publication

Azka Bima Aditya, Syafri Samsudin, Winahyu Pandu Rizki, Mahir Mahendra, Arif S... <1%

36 Student papers

Moodle2025 on 2026-04-30 <1%

37 Publication

Sayyid Muh. Raziq Olajuwon, Kusrini Kusrini, Kusnawi Kusnawi. "Analyzing Public ... <1%

38 Publication

Sigit Januarto, Aang Alim Murtopo, Zaenul Arif. "Klasifikasi Status Stunting Balita ... <1%

39 Student papers

University of Sheffield on 2023-09-07 <1%

40 Internet

cdn.juris.id <1%

41 Internet

dspace.jaist.ac.jp <1%

42 Internet

jurnal.atmaluhur.ac.id <1%

43 Internet

jurnal.polinela.ac.id <1%

44 Internet

newinera.com <1%

45 Internet

www.gov.br <1%

46 Internet

www.pythontutorials.net <1%

47 Internet

www.researchsquare.com <1%

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.XXXX

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

Dwi Andini Putri[1*]; Siti Nurwahyuni [2]

Informatics, Faculty of Engineering and Informatics[1*,2]
Universitas Bina Sarana Informatika, Jakarta, Indonesia[1,2]
www.bsi.ac.id[1,2]
dwi.dwd@bsi.ac.id[1], siti.swu@bsi.ac.id[2]

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

Abstract— *The representation of women in Science, Technology, Engineering, and Mathematics (STEM) continues to face various challenges rooted in social, cultural, and structural factors. This study aims to analyze public sentiment regarding the role of technology in promoting women's participation in STEM through a machine learning approach. The research data were obtained from 1,533 social media comments using web scraping techniques. After preprocessing, the data were automatically labeled using the VADER Lexicon-Based approach. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Text features were extracted using Term Frequency–Inverse Document Frequency (TF- IDF) and analyzed using the Support Vector Machine (SVM) algorithm with four kernels: linear, radial basis function (RBF), polynomial, and sigmoid. The VADER labeling results indicated that 98% of the comments were positive, while 2% were negative. The application of SMOTE proved effective in balancing class distribution, thereby improving model performance. Among the evaluated models, the linear kernel achieved the best performance with an accuracy of 98.31%, precision of 98.33%, recall of 98.31%, F1-score of 97.71%, and AUC of 85.81%. The overwhelming dominance of positive sentiment may introduce bias in interpreting the results, thus requiring caution when drawing conclusions about actual public perceptions. These findings confirm that sentiment analysis based on SVM and VADER can provide a clearer understanding of public perceptions and serve as a strategic foundation for developing policies to strengthen women's engagement in STEM sustainably..*

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

Intisari— Keterwakilan perempuan dalam bidang Sains, Teknologi, Teknik, dan Matematika (STEM) masih menghadapi berbagai tantangan yang bersumber dari faktor sosial, budaya, maupun struktural. Penelitian ini bertujuan untuk menganalisis sentimen publik terkait peran teknologi dalam mendorong partisipasi perempuan di STEM melalui pendekatan machine learning. Data penelitian diperoleh dari 1.533 komentar media sosial menggunakan teknik web scraping. Setelah melalui tahap preprocessing, data dilabeli secara otomatis menggunakan pendekatan VADER Lexicon-Based. Untuk mengatasi ketidakseimbangan kelas, digunakan metode Synthetic Minority Over-sampling Technique (SMOTE). Fitur teks diekstraksi menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) dan dianalisis menggunakan algoritma Support Vector Machine (SVM) dengan empat kernel, yaitu linear, radial basis function (RBF), polynomial, dan sigmoid. Hasil pelabelan VADER menunjukkan bahwa 98% komentar bersentimen positif, sedangkan 2% bersentimen negatif. Penerapan SMOTE terbukti efektif dalam menyeimbangkan distribusi kelas sehingga meningkatkan kinerja model. Dari evaluasi model, kernel linear menunjukkan performa terbaik dengan akurasi 98,31%, precision 98,33%, recall 98,31%, F1-score 97,71%, dan AUC 85,81%. Dominasi sentimen positif yang sangat besar berpotensi menimbulkan bias dalam interpretasi hasil, sehingga perlu kehati-hatian dalam menyimpulkan persepsi publik yang sebenarnya. Temuan ini menegaskan bahwa analisis sentimen berbasis SVM dan VADER mampu memberikan gambaran yang lebih jelas mengenai persepsi publik, sekaligus menjadi dasar

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.XXXX

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

strategis bagi penyusunan kebijakan untuk memperkuat keterlibatan perempuan di bidang STEM secara berkelanjutan.

***Kata Kunci**: Perempuan dalam STEM, Analisis Sentimen, Kernel SVM, Veder Lexicon.*

INTRODUCTION

The involvement of women in Science, Technology, Engineering, and Mathematics (STEM) has long been a global issue that requires serious attention. Despite various efforts to increase women's participation in STEM, gender disparities remain significant in many countries [1],[2]. Factors such as gender stereotypes, the lack of female role models, and biases in selection processes are among the major challenges[3]. Although the number of women entering STEM fields has increased, they continue to face structural and cultural barriers that limit their career advancement[4]. STEM represents a sector that drives global innovation and technological progress. However, women's participation in this field remains relatively low. In Indonesia, data from 2021 show that women accounted for only 40.6% of the STEM workforce, a lower proportion compared to Malaysia (48.6%) and Thailand (53.2%) [5]. According to *The Global Gender Gap Report 2023*, while women represent 49.3% of the global workforce outside STEM, they comprise only 29.2% of the workforce in STEM[6]. This disparity is influenced by structural, social, and cultural obstacles such as gender stereotypes, the lack of role models, and limited access to education and employment opportunities in STEM [7]. Digital technology is expected to help reduce this gap by expanding access to education, providing technology-based training, and fostering inclusive work environments[1]. However, evaluations of the effectiveness of such programs remain limited. Thus, public sentiment analysis becomes a relevant approach to understanding societal perceptions of the role of technology in promoting women's participation in STEM.

Sentiments expressed through social media, online forums, digital news, and other online platforms can be analyzed to capture levels of acceptance, support, and perceived barriers [8], [9]. Sentiment analysis enables the identification of sentiment patterns, trends, and factors influencing public perceptions [10]. Public sentiment analysis on social media can offer valuable insights into how society views women in STEM. Social media platforms often reflect people's attitudes and opinions toward certain issues, including women's involvement in STEM [11].

In this study, sentiment analysis is conducted using the Support Vector Machine (SVM) approach, which has proven to be one of the most effective text classification algorithms [12].. Several studies suggest that SVM works by identifying the optimal hyperplane that separates data classes, thereby distinguishing positive and negative opinions with high accuracy [9]. To enhance SVM's ability to handle non-linear data, several kernels are applied, including Linear, Radial Basis Function (RBF), Polynomial, and Sigmoid[13]. These kernels allow data transformation into higher-dimensional spaces, enabling better recognition of complex text patterns.

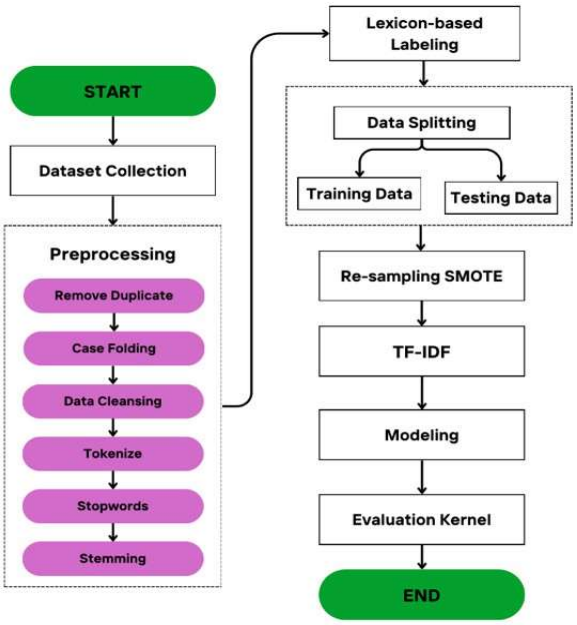
This study also aims to address the issue of class imbalance in sentiment datasets using the Synthetic Minority Over-sampling Technique (SMOTE)[14]. SMOTE improves the representation of minority classes by generating synthetic samples based on existing data, thereby enabling the SVM model to learn more effectively and accurately [15]. By employing TF-IDF for feature extraction, applying SVM with various kernel functions, and integrating SMOTE for re-sampling, this research seeks to provide a more accurate understanding of public perceptions regarding women's participation in STEM. The findings are expected to serve as a strategic foundation for developing more effective policies to sustainably enhance women's engagement in STEM.

The application of the VADER Lexicon method for sentiment labeling, combined with the SMOTE technique and analyzed using the Support Vector Machine (SVM) algorithm, is hypothesized to produce an accurate classification model that effectively represents public perceptions in a more balanced manner regarding the role of technology in promoting women's participation in the STEM fields.

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.

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.XXXX



Source: (Putri, 2025)

Figure 1. Research Procedure

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.

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

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.XXXX

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

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

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

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.XXXX

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. The model's performance was then compared across four different kernels, namely Radial Basis Function (RBF), Linear, Polynomial, and Sigmoid.

H. Evaluation Kernel

The final stage involved evaluating the kernels used in this study, namely RBF, Linear, Sigmoid, and Polynomial. The evaluation was conducted using the confusion matrix by examining the values of accuracy, precision, recall, and F1-score. In addition to the confusion matrix, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) values were also employed. Recall, precision, and F-measure are commonly used metrics for evaluating the performance of machine learning experiments.

RESULTS AND DISCUSSION

A. Preprocessing

Preprocessing was carried out to clean and transform raw text so that it could be more effectively processed by machine learning algorithms and analyzed. Table 1 presents the preprocessing results in this study, which consisted of several steps, namely removing duplicates, case folding, text cleansing, tokenization, stopword removal, and stemming.

Table 1. Preprocessing Results

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.XXXX

y juga tdk memi liki sumb er yg dpt mem banta h statm ent saya (terk ait perbe daan terlet ak pada kema mpua n dlm mem berda yagu naka n nya loh, buka n kema mpua n otak berpi kir logis nya) Ã°ÅŸ Å™Å	juga tdk memi liki sumb er yg dpt mem banta h statm ent saya (terk ait perbe daan terlet ak pada kema mpua n dlm mem berda yagu naka n nya loh, buka n kema mpua n otak berpi kir logis nya) Ã°ÅŸ Å™Å	juga tdk memi liki sumb er yg dpt mem banta h statm ent saya (terk ait perbe daan terlet ak pada kema mpua n dlm mem berda yagu naka n nya loh, buka n kema mpua n otak berpi kir logis nya) Ã°ÅŸ Å™Å	liki sumb er yg dpt mem banta h statm ent saya terka it perbe daan terlet ak pada kema mpua n dlm mem berda yagu naka n nya loh buka n kema mpua n otak berpi kir logis nya	'memi liki', 'sumb er', 'yang' , 'dapat ', 'mem banta h', 'statm ent', 'saya', 'terka it', 'perb edaan ', 'terlet ak', 'pada' , 'kema mpua n', 'dala m', 'mem berda yagun akan', 'buka n', 'kema mpua n', 'otak', 'berpi kir', 'logis nya']	'kema mpua n', 'otak', 'berpi kir', 'logis nya']	

Source : (Putri, 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].. Figure 1 illustrates the results of feature extraction processing using the TF-IDF method implemented in Python.

```
(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 : (Putri, 2025)

Figure 1. hasil pengolahan feature extraction dengan metode TF-IDF

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

1. The first number *(1)* represents the document index (e.g., the 1st document).
2. 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.

Table 2. Sentiment Labeling Results Using the VADER Lexicon

Data	Sentiment Score	Polarity
@kris****_s*** ...factory manager tempat kerja lama cewek, smart dan win solutif bangat malah.		

22

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.XXXX

rut saya kembali ke personal	0.4019	positif

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.XXXX

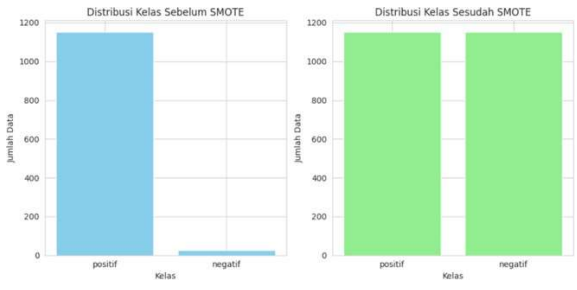
JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

masing2 si		
@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 : (Putri, 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, which may affect the performance of machine learning models, as they tend to be biased toward the majority class (positive). 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: (Putri, 2025)

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

Source: (Putri, 2025)

Figure 3. Word Cloud of Common Themes and Dominant Keywords

The word cloud in Figure 3 shows that the words "women," "technology," and "STEM" are the three most dominant terms, reflecting the main focus of public discussion on women's involvement in science and technology. The prominence of the word "women" emphasizes that gender equality and women's participation remain key issues within the STEM context. Meanwhile, the appearance of the word "technology" indicates that this field is often seen as a tangible representation of progress and innovation, yet one still largely dominated by men. The term "STEM" represents a multidisciplinary space symbolizing intellectual advancement and professional careers, while also revealing the existing gender participation gap. Together, these three words illustrate the social dynamics between potential, opportunity, and challenges faced by women in contributing to the STEM fields.

Evaluation Results of SVM and Kernels

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

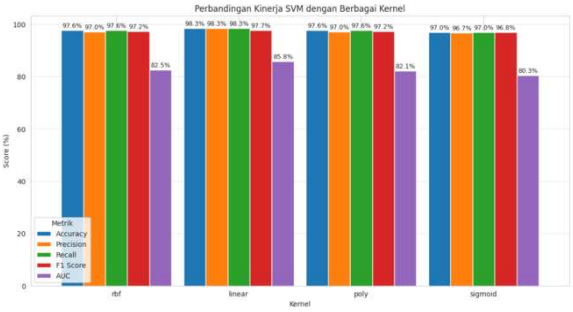
Table 3. Summary of SVM Kernel Comparison

	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: (Putri, 2025)

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

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.XXXX



Source: (Putri, 2025)

Figure 4. Comparison of SVM Models Based on Kernels

Table 3 and Figure 4 present a summary of the performance comparison of Support Vector Machine (SVM) models using four different kernels: RBF, Linear, Polynomial, and Sigmoid. 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%. From these results, the Linear kernel demonstrated the best overall performance, consistent with the findings of [27] which stated that the Linear kernel performs very well on text or structured data. The Linear kernel achieved the highest accuracy of 98.31%, precision of 98.33%, recall of 98.31%, F1-Score of 97.71%, and an AUC value of 85.81%. The RBF and Polynomial kernels showed very similar performance, with accuracy and F1-Score values of approximately 97.63% and 97.24%, respectively, although both recorded lower AUC scores compared to the Linear kernel. Meanwhile, the Sigmoid kernel exhibited the lowest performance among the four, with an accuracy of 96.95%. Based on these findings, it can be concluded that the Linear kernel provides the most optimal classification results for the dataset used, both in terms of accuracy and its ability to correctly identify positive cases (recall) as well as minimize false positives (precision).

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%. The Sigmoid kernel produced an accuracy of 96.95%, precision of 96.68%, F1-score of 96.81%, and an AUC of 80.33%. Meanwhile, the best performance was achieved by the Linear kernel, with an accuracy of 98.31%, precision of 98.33%, F1-score of 97.71%, and an AUC of 85.81%. These findings indicate that the use of different kernels in the SVM method enhances accuracy in analyzing public sentiment on empowering women in STEM. Furthermore, sentiment labeling with the VADER Lexicon revealed that positive sentiments were more dominant than negative ones. However, the class imbalance between positive and negative sentiments was successfully addressed through the application of SMOTE, which generated synthetic samples for the minority class and balanced the data distribution. This study concludes that sentiment analysis provides a clearer understanding of public perceptions and can serve as a strategic foundation for developing effective policies to increase women's participation in STEM. Future research is recommended to compare SVM with other algorithms such as Naïve Bayes, Random Forest, or K-Nearest Neighbors (KNN), to employ larger and more diverse datasets, and to consider deep learning approaches in order to improve accuracy and deepen insights into public perceptions regarding women's involvement in STEM.

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] R. Nur Amelia, A. Delyana Mafikah, And S. Rif, "Equality: Journal Of Gender, Child And Humanity Kesetaraan Gender Dalam Manajemen Sumber Daya Insani: Tantangan Dan Peluang", Doi: 10.58518/Equality.

[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:

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.XXXX

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

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] E. L. Utari And S. H. Wibowo, "Analisis Komparatif Algoritma Svm Naive Bayes Dan Lstm Pada Sentimen Komentar Lagu Labour," *Jurnal Informatika Teknologi Dan Sains*.

[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.

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

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.XXXX

[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