Comparative Analysis of Text Classification Models for Offensive Language Detection on Social Media Platforms

Alan Joseph
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
India
alanjoseph2025@cs.ajce.in

Abhinay A K
Department Of Computer Science and Engineering
Amal Jyothi College of Engineering
India
abhinayak2025@cs.ajce.in

Dr. Gee Varghese Titus
Department Of Electronics and Communication Engineering
Amal Jyothi College of Engineering
India
geevarghesetitus@amaljyothi.ac.in

Anagha Tess B
Department Of Computer Science and Engineering
Amal Jyothi College of Engineering
India
anaghatessb2025@cs.ajce.in

Adham Saheer
Department Of Computer Science and Engineering
Amal Jyothi College of Engineering
India
adhamsaheer2025@cs.ajce.in

Fabeela Ali Rawther
Department of Computer Science and Engineering
Amal Jyothi College Of Engineering
India
fabeelaalirawther@amaljyothi.ac.in

***Abstract*—The detection of offensive language in text has become increasingly crucial in various social media platforms to maintain a respectful and safe environment. In this research we study and present a comparative analysis of different text classification models for identifying offensive and non-offensive language. Specifically, we investigate the performance of Support Vector Classifier (SVC), Compliment model, Gaussian model, and Multinomial model on a dataset curated for this purpose. Each text classification model is implemented and trained using the preprocessed dataset, and their performance is evaluated using standard evaluation metrics such as accuracy. The experimental results display the effectiveness of each model in distinguishing offensive language from non-offensive language. This research contributes to the literature by providing empirical evidence on the performance of various text classification models for offensive language detection, thus aiding in the development of more robust and accurate detection systems for online platforms.**

Keywords—Textclassification,Offensive language detection ,Support Vector Classifier (SVC),Compliment model,Gaussianmodel,Multinomial model,Social media platforms,Empirical analysis,Performance evaluation,Online content moderation

I. INTRODUCTION

The ubiquitous nature of online communication has facilitated widespread interaction and information dissemination across various social media platforms and digital forums [1]. However, this accessibility has also engendered the proliferation of offensive language, posing significant challenges to maintaining respectful and safe online environments [2]. Thus the need for effective methods to detect and mitigate offensive language has become essential.

In response to this imperative, this study endeavors to conduct a comprehensive comparative analysis of different text classification models for the identification of offensive and non-offensive language. The objective is to evaluate the performance of four prominent models in this domain: the Support Vector Classifier (SVC) [3], Compliment model [4], Gaussian model [5], and Multinomial model [6]. By scrutinizing the efficiency of these models, we aim to elucidate their capabilities in detecting offensive language and contributing to the development of robust detection systems.

To facilitate this analysis, a curated dataset specifically tailored for offensive language detection is utilized. Prior to model training, the dataset undergoes preprocessing procedures, including tokenization, removal of emojis etc. These steps are crucial for enhancing the quality of input data and improving the performance of subsequent classification models.

Further each text classification model is implemented and trained using the curated dataset. Subsequently, the performance of these models is evaluated using established evaluation metrics such as accuracy. Through rigorous experimentation and analysis, we try to find the effectiveness of each model in distinguishing offensive language from non-offensive language.

.

DOI:10.5281/zenodo.12515626

II.LITERATURE REVIEW

Text classification techniques have been extensively studied in the realm of natural language processing, with various models proposed to effectively classify text into different categories. Among these models, Support Vector Classifier (SVC), Compliment, Gaussian, and Multinomial models have garnered significant attention due to their robust performance in different classification tasks.

The Support Vector Classifier (SVC) is a widely used model for binary classification tasks, known for its ability to handle high-dimensional data and nonlinear decision boundaries [3]. It works by finding the hyperplane that best separates the data points into different classes, maximizing the margin between classes while minimizing classification errors.

The Compliment model, proposed by Turney [4], is based on a lexicon of positive and negative words. It computes the semantic orientation of a text document by comparing the frequencies of positive and negative words in the document. This simple yet effective approach has been utilized in sentiment analysis and other text classification tasks.

Gaussian models, such as Gaussian Naive Bayes, are probabilistic models based on the assumption that features follow a Gaussian distribution. These models are commonly used for text classification tasks, especially when dealing with continuous-valued features [5]. They have been shown to perform well in practice, particularly when the independence assumption holds true.

Multinomial models, including Multinomial Naive Bayes, are another class of probabilistic models widely used in text classification. Unlike Gaussian models, Multinomial models are suitable for discrete features, making them well-suited for text data represented as word frequency counts or term frequency-inverse document frequency (TF-IDF) vectors [6].

For instance, Davidson et al. [2] proposed a supervised learning approach for hate speech detection, utilizing a dataset annotated with offensive language labels. They experimented with different feature representations and classification algorithms, demonstrating the effectiveness of machine learning techniques in detecting offensive language. Similarly, Schmidt and Wiegand [7] explored the use of deep learning models for hate speech detection, achieving competitive performance compared to traditional machine learning approaches.Overall, previous studies have demonstrated the efficacy of various text classification models, including SVC, Compliment, Gaussian, and Multinomial models, in identifying offensive language in text. These models serve as valuable tools for automated content moderation and fostering a respectful online environment.

III.METHODOLOGY

A.Dataset Description and Preprocessing

The selected dataset comprises tweets collected from Twitter and is specifically designed for text classification tasks, focusing on distinguishing between offensive and non-offensive tweets. The dataset is obtained from the TweetEval dataset, a widely used benchmark dataset from Hugging Face for evaluating text classification models. Further,the model is tested using a dataset that is obtained from Hugging Face.Before training the text classification models, the dataset undergoes several preprocessing steps to enhance the quality of the input data.These steps include:

1 Tokenization: The tweets are tokenized to split them into individual words or tokens, which serves as the basic units of analysis.

2. Removal of Emojis: Emojis are removed from the tweets as they do not contribute to the semantic meaning of the text and may introduce noise into the dataset.

3. Lowercasing: All text is converted to lowercase to ensure consistency in feature representation and to prevent duplication of words due to case differences.

By performing these preprocessing steps, the dataset is cleaned and standardized, making it suitable for training the text classification models.

B.Text Classification Models

1. Support Vector Classifier (SVC):

The Support Vector Classifier (SVC) is a widely used algorithm for binary classification tasks, including text classification. SVC aims to find the optimal hyperplane that separates data points into different classes by maximizing the margin between the classes while minimizing the classification error [3]. This approach makes SVC particularly effective for handling high-dimensional data and nonlinear decision boundaries. During the training phase, SVC identifies support vectors, which are the data points closest to the decision boundary, to define the optimal hyperplane.

2. Compliment Model:

The Compliment model, proposed by Turney [4], is a lexicon-based approach for sentiment analysis and text classification tasks. It calculates the semantic orientation of a text document by comparing the frequencies of positive and negative words in the document. The model assigns a positive or negative score to the document based on the relative frequencies of positive and negative words. This simple yet effective approach has been widely used in sentiment analysis tasks and has demonstrated competitive performance compared to more complex machine learning models.

3. Gaussian Model:

The Gaussian model, such as Gaussian Naive Bayes, is a probabilistic model commonly used for text classification tasks. It is based on the assumption that features follow a Gaussian distribution, making it suitable for continuous-valued features derived from text data [5]. In the context of text classification, Gaussian models are effective when dealing with continuous features, such as word embeddings or word frequency counts. Despite its simplicity, the Gaussian model has been shown to perform well in practice, particularly when the independence assumption holds true.

4.Multinomial Model:

The Multinomial model, including Multinomial Naive Bayes, is another probabilistic model commonly used for text classification tasks [6]. Unlike Gaussian models, Multinomial models are suitable for discrete features, making them well-suited for text data represented as word frequency counts or term frequency-inverse document frequency (TF-IDF) vectors. In the context of text classification, Multinomial models are effective when dealing with discrete features derived from text data. Despite its simplicity, the Multinomial model has been shown to perform well in practice, particularly for tasks such as document classification and sentiment analysis.

IV. RESULT

In this section, we present a detailed analysis of the results obtained from the experiments conducted with each text classification model, including Support Vector Classifier (SVC), Compliment, Gaussian, and Multinomial models.

1. Support Vector Classifier (SVC):

The SVC model achieved the highest accuracy among all models, with an impressive training accuracy rate of 96% and testing accuracy rate of 76%. The accuracy rate indicates a balanced performance in correctly classifying offensive and non-offensive tweets.

2. Compliment Model:

The Compliment model demonstrated competitive performance, achieving a training accuracy rate of 89% and testing accuracy rate of 75%. While the accuracy was slightly lower compared to SVC. The model exhibited a balanced performance in classifying offensive and non-offensive tweets.

3. Gaussian Model:

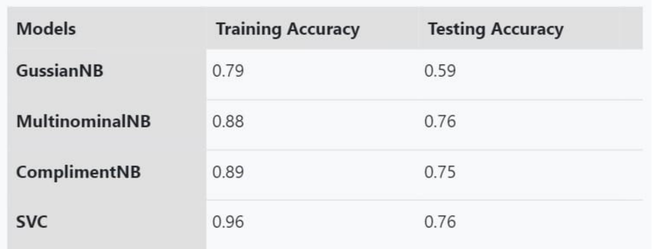
The Gaussian model, implemented using Gaussian Naive Bayes, achieved a training accuracy rate of 79% and with 59% of testing accuracy rate. The accuracy was lower compared to SVC and Compliment models. The model demonstrated a reasonable ability to distinguish between offensive and non-offensive tweets.

4. Multinomial Model:

The Multinomial model exhibited strong performance, with a training accuracy rate of 88% and testing accuracy rate of 76%. During the evaluation of accuracy rates, it became evident that SVC achieved the highest accuracy, closely followed by the Multinomial model. Despite slightly lower accuracies, both the Compliment and Gaussian models remained competitive overall.

Visualizations:

Overall, the experimental results (see Fig. 1 and Fig. 2) demonstrate the effectiveness of all text classification models in accurately distinguishing between offensive and non-offensive tweets. While SVC and Multinomial models exhibited slightly superior performance, Compliment and Gaussian models also showed competitive performance, highlighting their potential utility in real-world applications.



Models	Training Accuracy	Testing Accuracy
GussianNB	0.79	0.59
MultinominalNB	0.88	0.76
ComplimentNB	0.89	0.75
SVC	0.96	0.76

Fig. 1. Accuracy Table

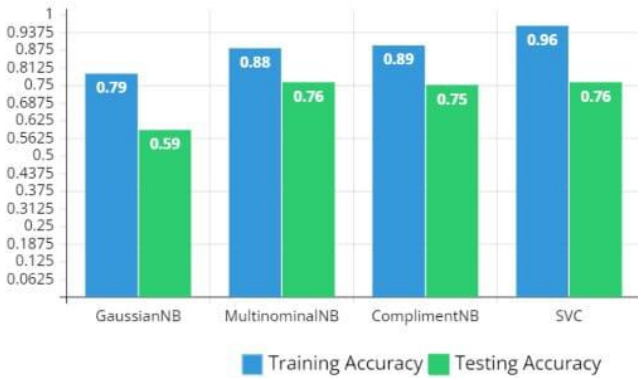


Fig. 2. Tabulated Accuracy Graph

V.DISCUSSION

Interpreting the results obtained from the experiments sheds light on the performance of each text classification model. The Support Vector Classifier (SVC) emerged as the top performer, boasting the highest accuracy in identifying offensive and non-offensive language. Its ability to handle high-dimensional data and nonlinear decision boundaries contributes to its robust performance, although with potential computational intensity and hyperparameter sensitivity. The Multinomial model also exhibited strong performance, particularly effective in handling discrete features like word frequency counts, yet it assumes feature independence and may struggle with rare words. The Compliment model, while simple and computationally efficient with its lexicon-based approach, is reliant on the quality of its lexicon and may overlook contextual nuances. Similarly, the Gaussian model, effective for continuous-valued features and simple to implement, may falter with non-Gaussian distributions. Understanding these strengths and weaknesses informs the selection of the most suitable model for specific tasks and underscores the importance of considering various factors influencing their performance.

VI. CONCLUSION

This research explored the effectiveness of various text classification models in identifying offensive and non-offensive languages.The analysis points out the capabilities and limitations of each model.

The Support Vector Classifier (SVC) performed the best, showing high accuracy in distinguishing between offensive and non-offensive content. The Multinomial model also performed well, indicating the effectiveness of probabilistic models in this task.While the Compliment and Gaussian models had slightly lower accuracy, they still showed competitive results. This highlights the versatility of lexicon-based and probabilistic approaches in identifying offensive language.

Overall, our research provides valuable insights into text classification for offensive language detection. By understanding the strengths and limitations of each model, we can develop more reliable systems for online platforms to maintain respectful environments. Looking forward, exploring ensemble methods and deep learning architectures could further improve detection accuracy. Additionally, considering multilingual contexts and diverse social media platforms will enhance the applicability of our findings and promote cross-cultural understanding.By advancing text classification techniques, we aim to create safer and more inclusive online spaces for everyone.

References

[1] J. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of Computer-Mediated Communication, vol. 13, no. 1, pp. 210-230, 2007.

[2] Z. W. Davidson et al., "Automated hate speech detection and the problem of offensive language," Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 512-515, 2017.

[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[4] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Proceedings of the Association for Computational Linguistics, pp. 417-424, 2002.

[5] E. Riloff and R. Jones, "Learning dictionaries for information extraction by multi-level bootstrapping," Proceedings of the National Conference on Artificial Intelligence, vol. 98, pp. 474-479, 1999.

[6] C. D. Manning et al., "Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?" Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 4, pp. 63-70, 2002.

[7] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1-10, 2017.