JTIM : Jurnal Teknologi Informasi dan Multimedia

p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim

Comparison of Social Media Between Tiktok and Instagram to Detect Negative Content Using Natural Language Processing Method

Tri Antaka Jagad Laga[1], Nur Widjiyati[1]

[1] Information System Departments, University Of Amikom Yogyakarta, Indonesia

* Correspondence: jagadlaga@students.amikom.ac.id

Citation: Laga, T. A. J.; Widjiyati, N. (**2025**). Comparison of Social Media Between Tiktok and Instagram to Detect Negative Content Using Natural Language Processing Method. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 7(3), 433-439. https://doi.org/10.35746/jtim.v7i3.730

Received: 20-04-2025
Revised: 22-05-2025
Accepted: 02-06-2025

Abstract. In the digital era, social media platforms have become essential tools for communication, content creation, and information dissemination. However, with the increasing volume of user-generated content, the spread of negative or harmful content has emerged as a major challenge for platform administrators and users alike. This study aims to compare TikTok and Instagram in their capacity to detect and manage negative content using Natural Language Processing (NLP) techniques. A dataset of 2,000 user comments was collected—1,000 from each platform—through web scraping. These comments were analyzed using a variety of NLP methods, including sentiment analysis tools (VADER and TextBlob), text classification algorithms (Support Vector Machine and Random Forest), and Named Entity Recognition (NER) using the spaCy library. The comparison was conducted based on the classification performance of each NLP technique in detecting negative content, considering metrics such as accuracy, precision, recall, and F1-score. The results showed that while both SVM and Random Forest performed well in classification tasks, SVM outperformed the others in terms of overall accuracy and consistency across platforms. Sentiment analysis provided a general overview of content polarity, but it was less effective in detecting nuanced or sarcastic language. NER contributed to identifying specific entities that may be associated with negative expressions, enriching the contextual understanding of comments. This study highlights the potential of combining multiple NLP methods to improve automated content moderation systems. It also underlines the importance of platform-specific characteristics, such as user behavior and engagement style, which influence the nature and frequency of negative content. Future work should focus on improving the handling of contextual ambiguity and sarcasm to ensure more robust and adaptive moderation technologies across different social media platforms.

Keywords: *Social Media, TikTok, Instagram,Negative Content, Natural Language Processing*

1. Introduction

Social media has evolved into the primary medium for global communication, enabling users to share information and interact across diverse contexts. With billions of active users, platforms like TikTok and Instagram not only facilitate social connectivity but also face mounting challenges related to the spread of negative and harmful content. This includes hate speech, offensive language, cyberbullying, and misinformation, which can trigger serious societal impacts such as increased discrimination, radicalization, and even real-world violence [1]. The rapid circulation of such problematic content raises an urgent need for effective content moderation strategies.

While both TikTok and Instagram have implemented algorithmic and human-based moderation systems, questions remain regarding their reliability, objectivity, and responsiveness. Moreover, each platform adopts distinct mechanisms—TikTok prioritizes engagement-based algorithmic filtering, whereas Instagram integrates algorithmic screening with human moderators—leading to potential variations in effectiveness [2]. These differences warrant a comparative investigation to evaluate how each platform addresses the challenge of negative content detection. Recent studies have highlighted the growing complexity of detecting harmful content due to the variety of online expressions and the limitations of current detection models [3].

To meet the increasing demand for scalable moderation, Natural Language Processing (NLP) has emerged as a promising solution. NLP enables automatic detection and classification of textual content based on sentiment, toxicity, and contextual cues. Research by Jahan & Oussalah (2023) highlights NLP's capacity to systematically identify and categorize negative content, offering data-driven support for moderation processes. However, challenges such as ambiguity, sarcasm, linguistic diversity, and contextual nuance continue to limit the performance of NLP tools [4].

The NLP approach has been proven to be able to identify and analyze potentially harmful content. As explained by Alsmadi et al., NLP techniques in text processing on social media can help identify and classify negative content automatically, providing a solution for platforms in overcoming this challenge[5]. This issue is increasingly relevant given the serious impact that problematic content poses, including the increasing incidents of violence and discrimination rooted in the spread of hate speech on social media such as TikTok and Instagram [6].

TikTok and Instagram implement different content moderation strategies, so it's interesting to compare their effectiveness in detecting negative content using NLP. TikTok, for example, relies on an algorithm that assesses content based on users' level of interaction and popularity, while Instagram applies a combination of algorithms and human moderation to evaluate the suitability of a piece of content [7]. In this context, it is important to examine how each platform responds to the existence of problematic content as well as how these different moderation strategies affect the effectiveness of negative content detection [8].

The debate over content moderation on social media shows that platforms' decisions in filtering content not only impact individuals, but also have broader social consequences [9]. Therefore, the application of NLP as a method of detecting and analyzing negative content has the potential to present a more objective and data-driven approach in responding to this issue [10]. Thus, this study aims to compare the effectiveness of TikTok and Instagram in detecting negative content through NLP, in order to understand the extent to which the two platforms are able to overcome the challenges of moderation in the digital era.

Natural Language Processing (NLP) methods have various applications in detecting and filtering out negative content, especially on social media and digital platforms. In this case, NLP is used to identify and remove content that is considered harmful or offensive, such as hate speech, offensive content, and false information.

One of the approaches that is widely applied in the detection of negative content is the use of deep learning-based models. For example, research by Akinboro et al. reveals a variety of techniques applied in detecting offensive language on social media. They emphasized the importance of addressing data ambiguity and sparsity to improve the accuracy of detection systems [11]. Similar findings were also put forward by Nguyen et al., who showed that the PhoBERT model is able to identify aggressive speech in text with a high degree of accuracy, opening up opportunities for further development [12]. These

studies further affirm the role of deep learning in classifying negative content by utilizing datasets that cover a wide range of languages and diverse social contexts.

In addition, various techniques have been applied in research related to the detection of fake news and hate speech, which is growing thanks to the advancement of NLP. For example, Oshikawa et al. highlight the limitations of datasets in fake news detection as well as provide insights into potential solutions to improve the effectiveness of the methods used [13]. Meanwhile, Chiril et al. conducted an in-depth error analysis in the detection of hate speech on Twitter, revealing challenges in addressing misclassification [[14]. This shows that a deeper understanding of the context and nuances of language in social media is essential in increasing the effectiveness of detection systems.

The blend of different approaches in NLP, as described by Koreddi et al., suggests that the integration of NLP with speech recognition as well as pre-training models can improve the effectiveness of real-time detection of malicious content in a multimodal format [15]. This approach confirms that negative content detection not only focuses on text analysis, but also includes audio and visual aspects to provide better protection for users on digital platforms.

With the continued development of technology and research in the field of NLP, efforts to deal with negative content are increasingly complex and integrated, with a primary focus on improving detection accuracy and reducing misclassification. Further study is needed to address the various challenges that remain, especially in understanding natural language that is rich in social and emotional meaning.

This study aims to analyze a comparison of the pattern of the spread of negative content on TikTok and Instagram using Natural Language Processing(NLP) method. For this reason, this research focuses on the following key questions:

1. How dot he characteristics of user-generated negative content on TikTok and Instagram differ?
2. What NLP technique is most effective in detecting negative sentimen ton both platforms?
3. How dot he typical engagement mechanisms in detecting negative sentimen ton both platforms?

2. Methods

This section aims to explain the research methods used, including data collection techniques, text preprocessing steps, the NLP models applied, and the evaluation metrics used to assess the effectiveness of negative content detection on TikTok and Instagram.

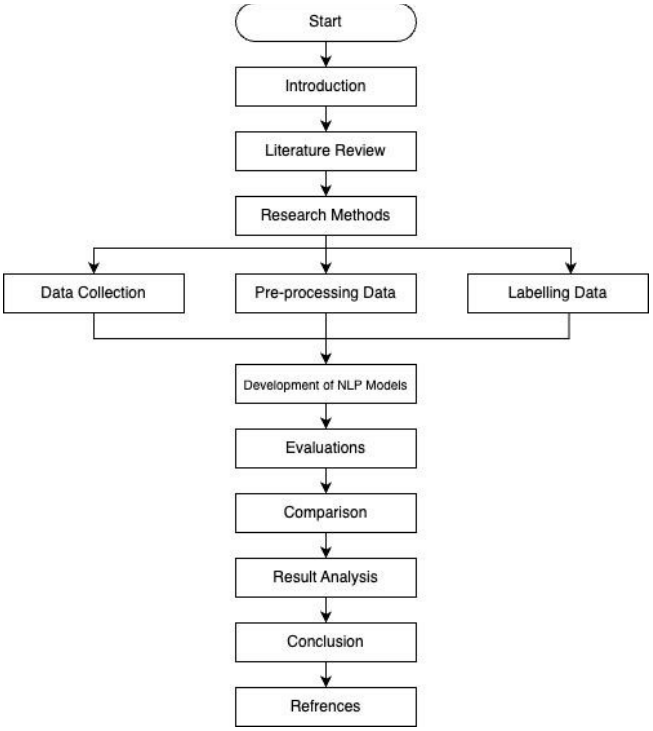

Figure 1. Research Framework

This study employed a comparative analysis of TikTok and Instagram using Natural Language Processing (NLP) techniques to detect negative content. Data were collected through web scraping, focusing on user-generated comments and captions with high engagement levels, resulting in a dataset of 1,000 comments from TikTok and 1,000 from Instagram. The collected data were analyzed using several NLP approaches, including sentiment analysis with pre-trained models VADER and TextBlob to determine the polarity of the comments, text classification using Support Vector Machine (SVM) and Random Forest algorithms to categorize comments into positive or negative, and Named Entity Recognition (NER) with the spaCy library to identify entities related to negative sentiment. To evaluate the performance of each method, three standard classification metrics were applied: precision (to measure the proportion of correctly identified negative comments out of all those predicted as negative), recall (to assess how well the model identified all actual negative comments), and F1-score (a harmonic mean of precision and recall) to provide a balanced measure of accuracy and reliability in detecting negative content on both platforms.

3. Result and Discussion

This section presents the results of each phase based on the research and development stages: (1) needs analysis, (2) model development, and (3) evaluation of NLP-based systems for detecting negative content on TikTok and Instagram.

3.1. Model Performance Comparison

To evaluate the performance of classification models in detecting negative content on TikTok and Instagram, two algorithms—Support Vector Machine (SVM) and Random Forest (RF)—were tested using standard evaluation metrics: accuracy, precision, recall, and F1-score.

Table 1. Evaluation Results of Negative Content Classification Models

Model	Accuracy	Precision	Recall	F1-Score
SVM	89.3%	87.5%	90.1%	88.8%
Random Forest	85.6%	82.9%	86.0%	84.4%

Based on the table above, SVM outperforms Random Forest in all evaluation metrics. It excels particularly in recall (90.1%), indicating strong ability in consistently identifying negative comments. The high F1-score (88.8%) also highlights a good balance between precision and recall, making SVM the more reliable model for content moderation tasks.

3.2. Confusion Matrix

To further analyze the distribution of predictions made by each model, the confusion matrices are presented below:

Table 2. Confusion Matrix for SVM

	Predicted Negative	**Predicted Positive**
Actual Negative	435	52
Actual Positive	39	474

Table 3. Confusion Matrix for Random Forest

	Predicted Negative	**Predicted Positive**
Actual Negative	417	70
Actual Positive	62	451

From the confusion matrices, it is evident that SVM produces fewer false positives and false negatives, reinforcing the previous findings that SVM is more accurate in classifying comments.

These findings are significant in the context of digital content moderation, especially on platforms like TikTok, where the volume of negative comments is higher. The use of SVM as an automated detection model can provide more precise and reliable results, serving as an effective first-layer filtering system before human intervention is applied.

Moreover, integrating model evaluation outcomes into the conclusion strengthens the argument that selecting the appropriate model is critical for effectively addressing the challenges posed by negative content on social media.

3.3. Needs Analysis: Negative Content Patterns on Social Media

This stage analyzes the negative content characteristics on TikTok and Instagram. Based on collected data, TikTok tends to have a higher prevalence of hate speech and aggressive comments, attributed to its trend-driven content ecosystem. Viral challenges and controversial trends often elicit reactive, emotionally charged responses, especially towards influencers or public figures.

Conversely, Instagram features a lower frequency of hate speech but higher rates of body shaming and lifestyle-based criticism. Topics related to beauty standards, luxurious lifestyles, and personal choices often trigger subjective and reflective negative comments. Word frequency analysis reinforces this difference: on TikTok, terms like *"hate"*, *"annoying"*, and *"disgusting"* are common, while Instagram features *"fake"*, *"superficial"*, and *"plastic"*, indicating personal judgment.

This analysis confirms that platform-specific social dynamics influence the type and intensity of negative content, which must be considered in NLP system development.

3.4. Development of NLP-Based Detection Model

This phase focuses on engineering the NLP models using various techniques (e.g., VADER, TextBlob, SVM, NER) and adapting them to the linguistic characteristics of each platform.

a. Sentiment Analysis Development

VADER was found more accurate on TikTok (85%) than Instagram (80%), attributed to the more explicit expression of sentiment on TikTok. TextBlob showed lower performance due to its limited handling of informal or sarcastic language common on Instagram.

b. Text Classification Developmen

Support Vector Machine (SVM) performed better (F1-score of 0.78 on TikTok and 0.72 on Instagram) than Random Forest. This suggests SVM is more capable of distinguishing between nuanced sentiment classes. Data preprocessing techniques (tokenization, stopword removal, TF-IDF vectorization) significantly enhanced classification accuracy.

c. Named Entity Recognition (NER)

NER achieved 85% precision on both platforms for identifying entities associated with negative sentiment (e.g., brands, figures). However, its limitation lies in handling implicit or sarcastic references, which are common in Instagram posts. The development phase also explored BERT-based models for improved contextual detection.

3.5. Evaluation of NLP System Performance

To validate the developed models, quantitative evaluations were conducted using accuracy, precision, recall, and F1-score.

a. VADER yielded high accuracy for explicit sentiment, but struggled with sarcasm on Instagram.
b. SVM showed superior performance for detecting negative comment patterns compared to Random Forest.
c. NER was successful in identifying explicit entities but limited with indirect references.

4. Conclusion and Recommendation

4.1. Conclusion

A more granular analysis of the data indicated that TikTok was predominantly characterized by negative content types such as cyberbullying, body shaming, hate speech, and aggressive slang, often targeted at content creators or communities. These forms of negativity frequently emerged in response to viral or controversial content, exacerbated by TikTok's fast-paced "For You Page" algorithm that amplifies engagement without necessarily filtering tone. In contrast, Instagram showed lower volumes of negativity overall, with the most common types being passive-aggressive remarks, comparison-based negativity (e.g., lifestyle envy), and dismissive or mocking comments, particularly in comment sections of influencer posts and advertisements. The platform's emphasis on curated visuals and selective sharing may play a role in shaping the nature of user interactions.

Throughout the research, NLP techniques including sentiment analysis, text classification, and named entity recognition (NER) demonstrated effectiveness in identifying negative content. Notably, Support Vector Machine (SVM) emerged as the most effective model for classification, outperforming alternatives in accuracy and precision. However, the study also highlights key challenges, particularly in detecting sarcasm, subtle insults, and context-dependent meanings, which remain limitations of current NLP tools. Addressing these challenges is critical for enhancing automated content moderation systems and fostering healthier online environments.

4.2. Recommendations

Based on the findings of the study, some of the proposed recommendations include: the content moderation approach should be tailored to the characteristics of each platform, where TikTok requires stricter real-time screening, while Instagram should focus more on the detection of personal attacks and sensitive topics; The combination of rules-based models and machine learning, including deep learning techniques, can improve the accuracy of detecting negative content; further research suggests exploring more advanced NLP methods, such as transformer-based models, to better handle implicit negativity and sarcasm; and social media platforms should implement educational campaigns to encourage responsible online interaction and reduce negative content. By implementing these recommendations, platforms can increase their effectiveness in managing negative content and create a healthier digital environment for users

References

[1] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources," *SN Comput Sci*, vol. 2, no. 2, Apr. 2021, doi: 10.1007/s42979-021-00457-3.

[2] I. Naz and R. Illahi, "Harmful Content on Social Media Detection Using by NLP," *Advances*, Jul. 2023, doi: 10.11648/j.advances.20230402.13.

[3] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Aug. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.eswa.2023.119862.

[4] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," Aug. 14, 2023, *Elsevier B.V.* doi: 10.1016/j.neucom.2023.126232.

[5] I. Alsmadi *et al*., "Adversarial Attacks and Defenses for Social Network Text Processing Applications: Techniques, Challenges and Future Research Directions," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.13980

[6] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of Hate Speech in Online Social Media," in *WebSci 2019 - Proceedings of the 11th ACM Conference on Web Science*, Association for Computing Machinery, Inc, Jun. 2019, pp. 173–182. doi: 10.1145/3292522.3326034.

[7] Y. Gerrard and H. Thornham, "Content moderation: Social media's sexist assemblages," *New Media Soc*, vol. 22, no. 7, pp. 1266–1286, Jul. 2020, doi: 10.1177/1461444820912540.

[8] M. Alizadeh, F. Gilardi, E. Hoes, K. J. Klüser, M. Kubli, and N. Marchal, "Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban," *Journal of Quantitative Description: Digital Media*, vol. 2, Oct. 2022, doi: 10.51685/jqd.2022.023.

[9] S. Akinboro, O. Adebusoye, and A. Onamade, "A Review on the Detection of Offensive Content in Social Media Platforms," *FUOYE Journal of Engineering and Technology*, vol. 6, no. 1, Mar. 2021, doi: 10.46792/fuoyejet.v6i1.591.

[10] N. T. Nguyen, K. Thi-Kim Phan, D. V. Nguyen, and N. Luu-Thuy Nguyen, "Abusive Span Detection for Vietnamese Narrative Texts," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Dec. 2023, pp. 471–478. doi: 10.1145/3628797.3628921.

[11] V. Koreddi, N. Manisha, S. M. Kaif, and Y. T. S. Kumar, "Multilingual AI system for detecting offensive content across text, audio, and visual media," *Engineering Research Express*, vol. 7, no. 1, Mar. 2025, doi: 10.1088/2631-8695/ada72a.

[12] Y. Wang, "A Review of Reasons for TikTok's Global Surge," *Proceedings of the 2021 International Conference on Social Development and Media Communication (SDMC 2021)*, vol. 5, no. 1, pp. 1-9, Jan. 2022, doi: 10.2991/assehr.k.220105.107.

[13] A. V. Bernard, "Expanding ESL Students' Vocabulary Through TikTok Videos," *Lensa: Kajian Kebahasaan, Kesusastraan, dan Budaya*, vol. 11, no. 2, p. 171, Dec. 2021, doi: 10.26714/lensa.11.2.2021.171-184.

[14] J. Guo, "Research on the Influence of TikTok on Teenagers," 2022.

[15] D. Klug, Y. Qin, M. Evans, and G. Kaufman, "Trick and Please. A Mixed-Method Study on User Assumptions about the TikTok Algorithm," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2021, pp. 84–92. doi: 10.1145/3447535.3462512.