Digital Vernaculars in Play: A Machine Learning-Based Linguistic Analysis of Slang and Code-Switching in Indonesian Youth Online Gaming Chats

Muhammad Faqih Dzulqarnain[1,*], Nadia Mirela[2]

[1,2] Information Technology, Aisyiyah Polytechnic of Pontianak, Pontianak, West Kalimantan, Indonesia

ABSTRACT

Online multiplayer games have emerged as critical "third places" for youth socialization, fostering unique digital vernaculars. In Indonesia, this linguistic landscape is a complex blend of formal Indonesian, regional dialects, internet slang, and English code-switching. However, the prevalence of toxic communication, including violence, racism, and harassment, presents a significant barrier to inclusive interaction. Standard automated moderation systems often fail in this context due to a lack of cultural and linguistic nuance. This study addresses this gap by conducting a machine learning-based linguistic analysis to systematically identify and categorize the features of toxic and non-toxic communication in Indonesian gaming chats. Using a manually labeled corpus of 10,702 chat messages, we implemented a supervised classification pipeline. A linear Support Vector Machine (SVM) model, utilizing Term Frequency-Inverse Document Frequency (TF-IDF) and n-gram (unigram and bigram) features, was trained to classify messages into four categories: violence, racism, harassment, and neutral. The model achieved a robust overall accuracy of 82%, demonstrating high efficacy in differentiating between the categories. The central finding of this research is the empirical validation that different forms of toxicity possess distinct and computationally identifiable vernaculars. The language of violence is characterized by general, impersonal insults; racism by specific, identity-based slurs; and harassment by targeted, often sexualized and gendered, terminology. This data-driven comparative analysis provides a nuanced linguistic framework that moves beyond simple keyword flagging. The findings have direct implications for the design of more sophisticated, culturally-aware automated moderation systems capable of understanding the specific nature of toxic behavior in complex digital environments.

Keywords Code-Switching, Computational Linguistics, Hate Speech Detection, Indonesian Language, Online Gaming

Submitted 7 July 2025
Accepted 28 July 2025
Published 1 September 2025

Corresponding author
Muhammad Faqih Dzulqarnain,
mfaqihdz@polita.ac.id

Additional Information and Declarations can be found on page 212

DOI: 10.63913/jds.v1i3.14

Copyright
2025 Dzulqarnain and Mirela

Distributed under
Creative Commons CC-BY 4.0

Introduction

The concept of "third places," as originally articulated by Ray Oldenburg, refers to communal spaces that foster informal interaction beyond the confines of home and work. In recent years, the emergence of online multiplayer games has positioned these digital environments as important third places for youth socialization, rivaling traditional social venues such as cafés and parks. These virtual spaces facilitate unique communication styles and vernaculars shaped by the demands of fast, text-based interactions, thus evolving the nature of social engagement among adolescents.

Research has highlighted the integral role of digital games in performing the functions associated with traditional third places. For instance, Littman et al.

assert that third places must promote opportunities for agency and inclusiveness, aligning with the dynamics seen in online gaming communities where young people actively shape their interactions [1]. Similarly, evidence indicates that online gaming platforms can reduce feelings of loneliness and bolster social connections among youth, particularly in precarious situations, such as unaccompanied refugees [2]. These findings suggest that communicative practices developed within gaming environments serve as vital support networks for marginalized youth, enriching their social experiences and overall well-being.

Moreover, the nature of communication in these online spaces has led to the development of unique in-group vernaculars. The gradual evolution of these communication styles can be attributed to the necessity for efficiency and clarity in the fast-paced environment typical of many multiplayer games. Virtual interactions require players to convey complex ideas succinctly, often leading to the creation of specialized jargon or slang unique to particular gaming communities [3], [4]. For example, avatars in virtual worlds, like those found in VRChat and Minecraft, facilitate social interaction by allowing players to express their identities and engage with one another effectively [5].

The significance of online games as third places extends beyond mere socialization; they also play a crucial role in enhancing emotional well-being. Research exemplifies how "cozy" games, which encourage nurturing environments, can foster positive emotional states among players, functioning similarly to physical third places [3]. Furthermore, studies indicate that such gaming contexts can enrich players' social capital, helping to mitigate social withdrawal and foster community among users [6], [7].

On a broader scale, the pandemic has brought increased attention to gaming as a tool for social connection. As physical gatherings became limited, online games emerged as valuable resources for maintaining relationships and establishing new social connections [8]. This shift underscores the importance for service providers and educators to recognize the evolving landscape of youth interaction and the critical role digital contexts play in fostering community and social wellness.

The linguistic landscape of Indonesian digital spaces reflects the diverse influences within the nation’s youth culture. This unique mix comprises formal Indonesian, regional dialects, internet slang, and a significant amount of English code-switching, creating a dynamic communicative environment. However, this linguistic diversity is accompanied by a prevalence of toxicity, particularly in the forms of violence, racism, and harassment, serving as significant barriers to fostering inclusive social interactions.

The environment of digital communication in Indonesia is marked by a fluid interplay of languages. Research indicates that linguistic practices in Indonesian digital communities exhibit a blend of public and private linguistic landscapes, where users navigate a multifaceted dialogue composed of various linguistic resources [9]. Evidence from studies shows that while formal Indonesian and local dialects dominate public communication, the digital realm presents a landscape where internet slang and English thrive [10]. These shifts not only symbolize linguistic evolution, but they also highlight the challenges posed by

toxic interactions which can marginalize certain groups.

Toxic communication has emerged as a critical issue in this multicultural and multisocial environment. Toxicity in digital spaces is often expressed through violent language, racist remarks, and harassment, significantly undermining positive engagement among users [11]. Such behavior is particularly observed on platforms popular among Indonesian youth, where anonymous interactions can embolden individuals to engage in harmful behavior without being held accountable by social norms [12]. The pervasive nature of toxicity contributes to a fragmented sense of community, where individuals may feel alienated or unsafe, counteracting the potential for inclusive social interaction [13].

Addressing the multifaceted issue of toxicity in Indonesian digital spaces necessitates a deeper understanding of the specific linguistic practices at play. This involves fostering awareness of the language choices and code-switching behaviors that Indonesian youth employ, which can either reinforce or challenge existing power dynamics within these communities. Initiatives aimed at promoting digital literacy and respectful communication could play a critical role in mitigating toxic behavior by educating users about the impacts of their communication styles and leveraging the linguistic versatility of youth to enhance emotional and cultural intelligence within online interactions [14] .

Moreover, there is a pressing need for community engagement strategies that harness the linguistic dynamism characteristic of Indonesian youth. Promoting positive discourse through inclusive community dialogues, moderated forums, and campaigns against toxic behavior could foster a more harmonious digital environment, allowing diverse linguistic expressions to coexist while minimizing harassment and discrimination [15]. This approach may empower users to communicate effectively and respectfully and facilitate the removal of barriers that inhibit inclusive social interactions.

This study adopts a supervised machine learning approach as its primary research method to systematically identify and categorize the linguistic features of both toxic and non-toxic communication. The primary objective is to move beyond anecdotal observations and develop an empirical, data-driven model that can recognize the nuanced patterns inherent in Indonesian online gaming chats. By training a classifier on a manually labeled dataset, we can computationally determine the specific words, phrases, and n-grams that are most predictive of distinct communication categories, thereby creating a structured framework for understanding a complex and fluid linguistic environment.

The central contribution of this paper is its departure from simplistic, keyword-based toxicity detection toward a more sophisticated comparative analysis. Rather than treating toxic language as a monolithic category, this research provides a data-driven deconstruction of the distinct vernaculars of violence, racism, and harassment within Indonesian gaming culture. By identifying the unique linguistic fingerprints of each sub-category, this study offers a more granular understanding of how different forms of harmful speech manifest. This nuanced perspective is critical for developing more effective, culturally-aware moderation technologies and contributes a novel sociolinguistic analysis of digital communication in a Southeast Asian context.

Literature Review

Theoretical Framework of Computer-Mediated Communication (CMC) and Sociolects

The study CMC has proven instrumental in understanding social interactions that occur in digital spaces, especially regarding how group identity and anonymity can lead to both innovative linguistic practices and toxic behaviors. Central to this discourse is the Social Identity model of Deindividuation Effects (SIDE), which posits that anonymity in CMC environments diminishes personal accountability, facilitating behaviors that may align with group norms rather than individual morals [4]. This model provides a framework for examining how digital anonymity can foster communal creativity, making platforms like gaming chats ripe for the emergence of distinct sociolects, or social dialects.

Research highlights that the anonymity afforded by CMC can catalyze both positive and negative social dynamics. While it can promote innovative forms of communication, through the formation of new vernaculars and community-specific lexicons, it simultaneously enables toxic communication such as harassment and hate speech [7]. For instance, during online interactions, players often utilize unique terminologies that evolve quickly, reflecting collective experiences and shared identities [16] . This phenomenon creates distinct sociolects characterized by a blend of formal language, slang, and code-switching, reflecting the diverse backgrounds of participants while also serving as a barrier to inclusion for those not familiar with the vernacular [17].

Platforms like Twitter and Reddit have been studied extensively, demonstrating the formation of sociolects through communal interactions and the subsequent influence these linguistic forms have on community identity and cohesion [18]. On these platforms, users curate their language to fit in with community norms, which is similarly seen in gaming environments [19]. Previous work has shown that such interactions are often laden with communal values that reinforce group identity. In the context of gaming, these interactions can manifest as players develop a shared lexicon that simultaneously creates opportunities for connection and inclusivity while also potentially breeding toxicity against outsiders or members who violate group norms [7].

The importance of these findings is further underscored by studies exploring the psychiatric implications of these behaviors. The relationship between the use of gaming and social media and symptoms of psychiatric disorders suggests that the drive for social linkage in gaming can sometimes conflict with individual mental health concerns, creating an environment where toxic behaviors may flourish due to the lower social accountability afforded by anonymity [4]. This inconsistency can inhibit the potential for CMC platforms, such as gaming chats, to serve as spaces of positive social interaction and community support, ultimately affecting players' experiences.

Given these complexities, investigating sociolects within gaming chats as distinct linguistic environments allows for a deeper understanding of how these digital spaces function socially and psycholinguistically. This research bears significance, as it aligns with broader studies on the dynamics of CMC, providing

a nuanced lens for analyzing group identity and assertive community practices while recognizing the dual potential for innovation and toxicity within these interactions [20], [21].

Computational Linguistics and Hate Speech Detection

The detection of hate speech and toxicity in online spaces has evolved significantly in its methodological approaches, transitioning from initial dictionary-based techniques to advanced machine learning models. These developments reflect a growing awareness of the complexities inherent in analyzing online communication, especially in contexts laden with cultural nuances and ambiguity.

Early approaches to hate speech detection primarily relied on keyword or dictionary-based systems, which identified hate speech based on the presence of specific terms or phrases that were categorized as offensive. Saleem et al. note the limitations of these keyword-based methods, highlighting their susceptibility to context loss and the challenges in accurately identifying hate speech beyond recognizable keywords [22]. This inadequacy demonstrated the need for models that incorporate broader patterns of language usage and social context, paving the way for machine learning techniques that leverage vast datasets to discern nuanced language patterns and contexts [22].

As the field has evolved, researchers have employed machine learning algorithms that not only analyze linguistic features but also integrate sociolinguistic understanding. For instance, machine learning models can be trained on data from self-identifying hateful communities, improving their performance by reducing the need for exhaustive and sometimes biased annotation processes typical of earlier methods [22] This approach allows for a more dynamic analysis of language use that adapts over time and across various platforms, such as Twitter and Reddit, where vernaculars emerge from community-driven interactions [23].

Despite advancements in machine learning, significant limitations remain. Current models frequently struggle with understanding context, irony, and culturally-specific slang, drawbacks emphasized by Mandryk et al., who argue that without accounting for the diverse linguistic practices prevalent in online environments, models may underperform in real-world applications [24]. Moreover, the emergence of irony and humor in hate speech necessitates sophisticated models that understand contemporary cultural dynamics, which many existing frameworks fail to achieve [25]. The subtlety of language, especially in multilingual contexts, can further confound detection mechanisms that do not adapt to the fluidity of language use.

These limitations underscore the necessity for an approach that utilizes sophisticated algorithms while also incorporating linguistic and cultural insights unique to specific online communities. An interdisciplinary methodology that engages with computational linguistics alongside social media studies may offer a more robust solution to the challenges of toxicity and hate speech detection in digital interactions, enabling a comprehensive exploration of sociolects evolving on platforms where youth and diverse communities engage [26].

Sociolinguistics of Modern Indonesia

The phenomenon of code-switching between Indonesian and English, particularly among the youth, serves as a crucial lens for understanding the sociolinguistic dynamics of modern Indonesia. This code-switching is not merely a reflection of linguistic ability but functions as a significant marker of social status, modernity, and topic-specificity. As Indonesian youth increasingly engage in globalized discourses, particularly in digital spaces such as social media and gaming forums, their linguistic choices reveal layers of identity and cultural expression.

Research confirms that code-switching can indicate social positioning. For example, Martin-Anatias discusses how language selection in Indonesian contexts often reflects socio-political sentiments and personal identity construction, particularly in an era of cultural shifts post-Reformasi [27]. English, in this context, symbolizes modernity and global connectivity, allowing speakers to align with contemporary trends and promote a cosmopolitan persona. Such use of English is prevalent in technical discussions, especially within gaming culture, where specific jargon is often in English due to the global nature of the industry. However, the reference on English use within the context of educational settings [28] does not support the claim made in this paragraph and has been removed.

This aligns with the findings of Maha et al., who note that code-switching manifests prominently in multimedia contexts such as talk shows, where mixing languages allows participants to convey nuanced meanings appropriate for varied audiences [29]. In addition to code-switching, the use of bahasa gaul, or Indonesian slang, has been further enhanced by digital communication platforms, serving as a linguistic evolution reflective of youth culture. Studies like that of Kandiawan highlight that Gen Z Indonesians frequently mix traditional Bahasa Indonesia with English and new slang terms, often creating a hybrid form that is distinct from both standard Indonesian and formal English [30]. This blend not only facilitates more authentic self-expression among youths but also creates community among peers who share a similar cultural background and digital literacy.

Moreover, the digital context demands a topic-specific understanding of language, especially in fields laden with English terminology. Sumaryanti and Yuniar found that social media facilitates emotional articulation and the exchange of knowledge, reinforcing code-switching as a practical tool for conveying complex technical information pertaining to modern subjects, such as gaming [31]. This usage further broadens the spectrum of communication styles and sociolects prevalent in Indonesian youth interactions, allowing for an inclusive yet distinct identity to emerge.

Despite the positive aspects of this linguistic fluidity, certain challenges persist. The rapid evolution of slang and mixed codes can alienate those not fully immersed in these linguistic trends, potentially creating barriers in communication and access to social networks. Moreover, the cultural nuances tied to specific terms often go unrecognized by outsiders, complicating the landscape of Indonesian-English communication [32].

Method

This study employed a supervised machine learning approach to systematically classify toxic and non-toxic communication within a corpus of Indonesian online gaming chats. The selection of a supervised methodology was deliberate, predicated on the necessity of predefined, distinct categories of communication (racism, violence, harassment, neutral) to guide the model's learning process. The overall methodology was executed through a multi-stage computational pipeline, encompassing data preprocessing, feature engineering, model training, and rigorous evaluation. Each stage was designed to methodically transform the raw, unstructured, and often chaotic text data into a structured, numerical format suitable for computational analysis. This systematic refinement is critical for building a robust classification model capable of discerning the nuanced linguistic features that characterize each communication category, ultimately enabling a data-driven analysis of digital vernaculars.

Corpus and Ethical Considerations

The foundation of this research is a corpus consisting of 10,702 individual chat messages collected from Indonesian youth online gaming sessions. The creation of this corpus involved a meticulous manual annotation process, where each message was carefully examined and assigned one of the four aforementioned labels. This manual labeling, while labor-intensive, is fundamental to the study's validity as it establishes a high-quality, human-validated ground truth essential for training and evaluating the machine learning models. The reliability of this ground truth is paramount for ensuring that the model learns from accurate and contextually appropriate examples. In adherence with ethical research standards and to mitigate potential harm, several considerations were upheld. All data was anonymized to protect the identities of the users involved. Furthermore, to ensure reader sensitivity and prevent the perpetuation of harmful language, all vulgar and profane terms identified during the analysis are presented in a partially masked format throughout this paper. This practice respects the dignity of the communities being studied while still allowing for a transparent linguistic analysis.

Preprocessing Pipeline

To prepare the text data for feature engineering, a comprehensive and sequential preprocessing pipeline was implemented to clean, normalize, and standardize the corpus. This process is crucial for reducing data noise and focusing the models' learning on the most semantically significant features. The pipeline began with case folding, where all text was converted to lowercase. Subsequently, Python's regular expression (re) library was used for a cleaning step to remove non-essential textual elements that add noise without contributing meaning, such as URLs, user mentions (@username), and any characters that were not part of the standard alphabet.

The cleaned text was then subjected to tokenization using the word_tokenize function from the Natural Language Toolkit (nltk) library, a fundamental step that breaks down each message into a list of individual words, or tokens. Following

this, a curated and extensive list of Indonesian stopwords was used to filter the tokens. Finally, to consolidate morphological variations of words into a common base form, stemming was applied using the Sastrawi library's StemmerFactory. This step reduces words to their root (e.g., "bermain," "memainkan," and "permainan" all become "main"), which significantly minimizes the dimensionality of the feature space and allows the model to learn the core concept of a word regardless of its inflectional form.

Feature Engineering

After preprocessing, the cleaned and stemmed text was transformed into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique, implemented via the TfidfVectorizer class from the scikit-learn library. TF-IDF was specifically chosen for its ability to create a weighted feature matrix that reflects not just how frequently a word appears in a message (Term Frequency), but also how unique or important that word is across the entire corpus (Inverse Document Frequency). To capture crucial multi-word phrases and collocations, the vectorizer was configured with the parameter ngram_range=(1, 2), instructing it to generate features for both unigrams (single words) and bigrams (two-word sequences). The inclusion of bigrams is particularly vital, as many forms of toxicity are expressed through phrases where the combination of words is what carries the harmful intent (e.g., the racist slur c**a k***r has a specific, charged meaning that is lost when analyzing "c**a" and "k***r" in isolation).

Experimental Design and Evaluation

For the experimental phase, the dataset was partitioned using the train_test_split function from scikit-learn. A test_size of 0.2 was specified, allocating 80% of the data to the training set and the remaining 20% to the testing set. To ensure the experiment's reproducibility, a random_state of 42 was used. Crucially, the stratify=y parameter was enabled to ensure that the proportional distribution of the four categories was maintained in both the training and testing subsets, preventing potential biases during evaluation.

Two distinct classification algorithms from scikit-learn were trained and compared. The first was MultinomialNB, a probabilistic model used as an efficient baseline. The second was a SVC (Support Vector Machine), configured with a kernel='linear'. A linear kernel was chosen as it is highly effective for text classification tasks where the data is often linearly separable in a high-dimensional feature space.

The performance of each model was assessed using standard metrics from scikit-learn's metrics module. Accuracy was used to measure overall correctness, while the classification_report function provided per-class Precision, Recall, and F1-Score. Finally, a confusion_matrix was generated for each model and visualized using the matplotlib and seaborn libraries to visually analyze specific error patterns and highlight which categories were most frequently confused with one another.

Result and Discussion

This section presents the performance of the two machine learning models and provides a comparative linguistic analysis of the results. The findings are

discussed in two parts: first, a quantitative evaluation of model performance to select the superior algorithm, and second, a qualitative discussion of what these results reveal about the distinct vernaculars of toxicity within the Indonesian online gaming community. This dual approach allows for a transition from statistical validation to a richer, context-aware interpretation of the communication patterns.

Quantitative Performance and Model Selection

The experimental phase yielded clear and compelling results, demonstrating the efficacy of the machine learning pipeline in classifying toxic communication. Both the Multinomial Naive Bayes (MNB) and the linear Support Vector Machine (SVM) performed significantly better than a random baseline, but the SVM emerged as the decisively superior model for this task. The MNB classifier achieved a respectable overall accuracy of 79%, while the SVM reached a higher accuracy of 82%, indicating a more robust capability to generalize from the training data to unseen chat messages.

A more granular analysis of the per-class metrics, detailed in the classification reports, reinforces the SVM's superiority and reveals critical differences in how each model handles nuance. The SVM consistently achieved more balanced and often higher F1-scores across all four categories: Harassment (0.88), Neutral (0.75), Racist (0.84), and Violence (0.81). This balance is crucial, as it signifies a model that does not sacrifice performance in one category to boost another. In contrast, the MNB model showed notable performance imbalances. For instance, while the MNB had a very high recall (0.88) for the violence category, its precision was the lowest of any class (0.69). This specific imbalance suggests that while the MNB was effective at finding most instances of violence, it did so at the cost of frequently misclassifying other forms of toxic or even neutral speech as violent, a tendency that in a real-world moderation system would lead to excessive and inaccurate user penalties. Conversely, its precision for harassment was high (0.93), but its recall was lower (0.77), indicating it was highly confident when it identified harassment but failed to detect nearly a quarter of all instances.

The SVM model rectified these imbalances, showing a more harmonious relationship between precision and recall across all categories. This superior performance can be attributed to the SVM's underlying mechanism, which seeks to find an optimal separating hyperplane with the maximum margin between classes. In a high-dimensional feature space like TF-IDF vectorized text, this approach is often more robust than the MNB's probabilistic assumptions, which can be challenged by the complex dependencies between words. The SVM's ability to create a more effective decision boundary indicates that it is not only more accurate overall but also a more reliable and robust classifier for this specific, nuanced task. Given its superior and more balanced performance, the subsequent linguistic analysis is based on the patterns identified by the SVM model.

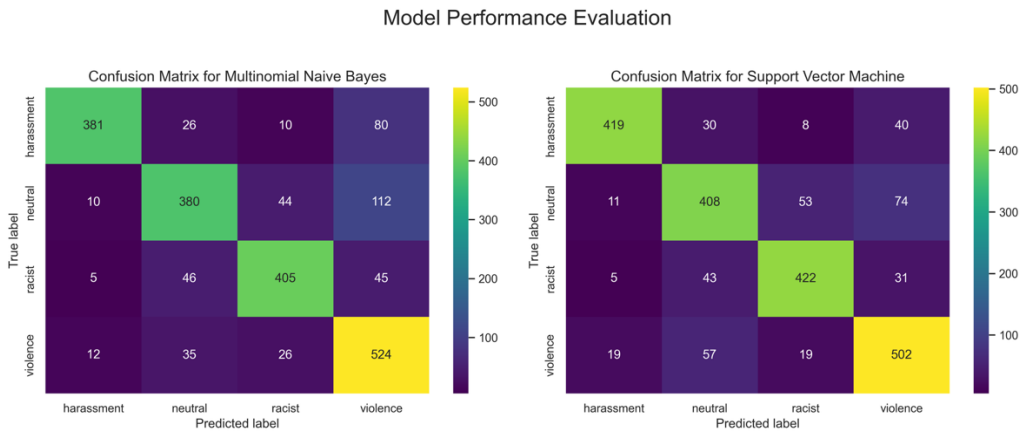


Figure 1 Model Performance Evaluation

Figure 1 provides a visual comparison of the performance of the Multinomial Naive Bayes and Support Vector Machine models through their respective confusion matrices. The matrices map the true labels against the models' predicted labels, with the main diagonal representing correct classifications. A qualitative inspection immediately reveals the SVM's superior performance, evidenced by the higher concentration of values along its main diagonal (e.g., correctly classifying 419 harassment instances versus the MNB's 381) and lower values in the off-diagonal cells, indicating fewer misclassifications. Notably, the SVM significantly reduces the confusion between the neutral and violence categories compared to the MNB. This visual data corroborates the quantitative metrics, confirming that the SVM is not only more accurate but also a more robust and reliable model for differentiating the nuanced categories of communication in the dataset.

Comparative Linguistic Analysis of Category Vernaculars

The quantitative results provide strong empirical evidence for the central hypothesis of this study that different forms of toxicity are not linguistically monolithic but possess their own distinct and computationally identifiable vernaculars. The SVM's ability to differentiate these categories with high accuracy allows for a data-driven exploration of their unique linguistic features and social functions within the gaming context.

The SVM's balanced F1-score of 0.81 for the violence category suggests this form of toxicity is characterized by a set of consistently used, high-impact terms. This vernacular is best understood as the language of immediate frustration, often triggered by in-game events like a tactical failure or a perceived lack of skill from teammates. Linguistically, it consists of general, often impersonal insults (e.g., g****k, t***l), expletives, and animal-related vulgarities (e.g., a****g, b**i). Unlike more targeted forms of toxicity, the language of violence is typically broader, less specific to the target's identity, and serves as a raw, unfiltered expression of anger.

Both models performed exceptionally well in identifying racist language, with the SVM achieving a strong F1-score of 0.84. This high performance strongly suggests that the vernacular of racism within this dataset is lexically specific and unambiguous. The language relies on potent keywords and, crucially,

bigrams that combine ethnic or religious slurs with derogatory qualifiers (e.g., d***r j**a). These terms have little to no ambiguity in the Indonesian sociopolitical context and function as powerful, low-context signals of hostility and othering. Their computational identifiability stems from their specificity; these phrases rarely appear outside of a racist context, making them highly predictive features for a machine learning model.

With the highest F1-score of 0.88, the harassment category appears to be the most linguistically distinct and formulaic. This suggests a vernacular characterized by highly specific, targeted, and often sexualized or gendered language. The features defining this category are likely vulgar terms for anatomy and sexual acts that are used to directly demean and objectify an individual, frequently targeting female players. The model's exceptionally high precision (0.92) indicates that when these features are present, the classification is almost certain, as this vocabulary is rarely used for any other communicative purpose within the game. This vernacular is not just toxic; it is a tool of personal degradation.

As anticipated, the neutral category proved to be the most challenging for the model, achieving the lowest F1-score (0.75). This is not a sign of model failure but rather a reflection of the category's inherent linguistic diversity. "Neutral" is a catch-all class encompassing a vast range of communicative acts: game-specific strategic communication (e.g., "push mid," "hati-hati di jungle"), technical questions, social greetings, and general chatter. It lacks the repetitive, specialized, and high-impact vocabulary that defines the toxic categories, making it inherently more difficult to classify against their highly distinct patterns. Its features are spread thinly across a wide vocabulary, whereas toxic vernaculars concentrate their features in a small set of powerful terms.

In synthesizing these findings, this analysis confirms that a computational approach can successfully move beyond simple keyword flagging. The SVM model did not just learn a list of "bad words"; it learned the distinct vocabularies and contextual patterns that differentiate the intent behind the language, whether it be general aggression (violence), identity-based attacks (racism), or targeted personal degradation (harassment). This nuanced understanding is a critical step toward developing more intelligent and culturally aware automated moderation systems.

Limitations and Future Research

Despite the promising results, this study has several limitations that offer clear directions for future research. First, the dataset is static, representing a snapshot of communication at a specific point in time. Online vernaculars, particularly gaming slang, evolve with remarkable speed, meaning the model's performance may degrade as new terms and phrases emerge. Second, the current model operates on a message-by-message basis, rendering it blind to the broader conversational context. It cannot reliably interpret sarcasm, irony, or messages that are only toxic in response to a preceding comment. Finally, this study is unimodal, focusing exclusively on text while ignoring non-textual cues like emojis, which can significantly alter a message's tone and intent.

These limitations inform several avenues for future work. A longitudinal study involving data collection over an extended period would be invaluable for

building dynamic models that can adapt to linguistic evolution. To address the issue of context, future research should explore more advanced neural network models, such as BERT or other transformer-based architectures, which are designed to understand contextual relationships within text. Furthermore, a multimodal analysis that incorporates emoji and other visual data alongside text would provide a more holistic and accurate classification system. Finally, a follow-up qualitative analysis of the current model's misclassifications could yield rich insights into the most ambiguous and nuanced forms of toxic communication, guiding the development of more sophisticated and culturally-attuned moderation tools.

Conclusion

This study successfully developed and validated a machine learning model capable of classifying toxic communication in Indonesian online gaming chats with 82% accuracy. By employing a Support Vector Machine with TF-IDF and n-gram features, we have moved beyond simple toxicity detection to provide empirical evidence for the existence of distinct digital vernaculars. The findings demonstrate that violence, racism, and harassment are not linguistically interchangeable; rather, they are characterized by unique vocabularies and patterns. The vernacular of violence is marked by general, frustration-driven insults, racism by specific identity-based slurs, and harassment by targeted, often gendered and sexualized, language. This differentiation is the core linguistic contribution of this research, revealing how intent and social function are encoded in specific lexical choices within this digital community.

The implications of these findings are significant for the field of automated content moderation. By proving that different forms of toxicity have unique linguistic fingerprints, this research provides a blueprint for creating more intelligent, nuanced, and culturally-aware moderation systems. Such systems can move beyond simplistic keyword blacklists, which are often ineffective against the complexities of slang and code-switching, toward models that understand the specific nature of a toxic act. While acknowledging the limitations of a static dataset and the need for more context-aware models, this study serves as a crucial step in developing technologies that can foster safer and more inclusive online social spaces for youth in Indonesia and beyond.

Declarations

Author Contributions

Conceptualization: M.F.D.; Methodology: N.M.; Software: N.M.; Validation: M.F.D.; Formal Analysis: N.M.; Investigation: M.F.D.; Resources: N.M.; Data Curation: M.F.D.; Writing Original Draft Preparation: N.M.; Writing Review and Editing: M.F.D.; Visualization: N.M.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. M. Littman, T. Milligan, R. R. Berry, B. T. Holloway, and M. L. Scott, "What Do Recently Housed Young People Imagine for the Future of Third Places? Using Game-based Inquiry to (Re)imagine Affirming, Youth-centered Third Places," *Am. J. Community Psychol.*, 2024, doi: 10.1002/ajcp.12750.

[2] G. Seddighi, "Taking a Dialogical Approach to Guiding Gaming Practices in a Non-Family Context," *Media Commun.*, 2022, doi: 10.17645/mac.v10i4.5727.

[3] H. Gunderman, "Cosy Video Games as Digital Third Places for Emotional Well-Being: Case Studies of Stardew Valley, Coffee Talk Episode 2 and Kinder World," *J. Gaming Virtual Worlds*, 2024, doi: 10.1386/jgvw_00108_1.

[4] Y. Du, T. Grace, K. Jagannath, and K. Salen, "Connected Play in Virtual Worlds: Communication and Control Mechanisms in Virtual Worlds for Children and Adolescents," *Multimodal Technol. Interact.*, 2021, doi: 10.3390/mti5050027.

[5] X. Wang, W. Xing, and J. M. Laffey, "Autistic Youth in 3D Game-based Collaborative Virtual Learning: Associating Avatar Interaction Patterns With Embodied Social Presence," *Br. J. Educ. Technol.*, 2018, doi: 10.1111/bjet.12646.

[6] L. L. Liu, T. M. H. Li, A. R. Teo, T. A. Kato, and P. W. Wong, "Harnessing Social Media to Explore Youth Social Withdrawal in Three Major Cities in China: Cross-Sectional Web Survey," *Jmir Ment. Health*, 2018, doi: 10.2196/mental.8509.

[7] L. Kaye, R. Kowert, and S. Quinn, "The Role of Social Identity and Online Social Capital on Psychosocial Outcomes in MMO Players," 2018, doi: 10.31219/osf.io/32uvr.

[8] H. R. Marston and R. Kowert, "What Role Can Videogames Play in the COVID-19 Pandemic?," *Emerald Open Res.*, 2020, doi: 10.35241/emeraldopenres.13727.1.

[9] D. Purnamasari, E. S. Sholekhah, P. Nurkamaliah, A. Apandi, and H. Herlina, "Investigation of the Linguistic Landscape of Local People at Pantura," *J. Engl. Lang. Teach. Appl. Linguist.*, 2024, doi: 10.32996/jeltal.2024.6.2.10.

[10] N. Yannuar and Y. Febrianti, "Walikan in the Linguistic Landscape of Malang: The Rise of a Local Youth Language," 2021, doi: 10.2991/assehr.k.211226.016.

[11] G. Razali and Y. Yulianti, “The Influence of Digital Communication on TikTok Addictive Behavior on Elementary School,” *J. Komun. Ikat. Sarj. Komun. Indones.*, 2022, doi: 10.25008/jkiski.v7i2.760.

[12] Y. W. Riani, A. W. Ningsih, M. Novitasari, and M. S. Samudra Zulkarnaen, “A Linguistic Landscapes Study in Indonesian Sub-Urban High School Signages: An Exploration of Patterns and Associations,” *J. Appl. Stud. Lang.*, 2021, doi: 10.31940/jasl.v5i1.2434.

[13] K. Miller, E. Champion, L. Summers, A. Lugmayr, and M. Clarke, “The Role of Responsive Library Makerspaces in Supporting Informal Learning in the Digital Humanities,” 2018, doi: 10.1016/b978-0-08-102023-4.00007-0.

[14] C. Cocq and E. Liliequist, “Digital Ethnography: A Qualitative Approach to Digital Cultures, Spaces, and Socialites,” *First Monday*, 2024, doi: 10.5210/fm.v29i5.13196.

[15] B. V. Winkle, N. Carpenter, and M. Moscucci, “Why Aren’t Our Digital Solutions Working for Everyone?,” *Ama J. Ethic*, 2017, doi: 10.1001/journalofethics.2017.19.11.stas2-1711.

[16] N. Crenshaw, J. LaMorte, and B. Nardi, “‘Something We Loved That Was Taken Away’: Community and Neoliberalism in World of Warcraft,” 2017, doi: 10.24251/hicss.2017.247.

[17] C. Phetphum, O. Keeratisiroj, and A. Prajongjeep, “The Association Between Mobile Game Addiction and Mental Health Problems and Learning Outcomes Among Thai Youths Classified by Gender and Education Levels,” 2023, doi: 10.21203/rs.3.rs-2863303/v1.

[18] K. Yusuf, Z. Rohmah, and O. I. Alomoush, “The Commodification of Arabic in the Commercial Linguistic Landscape of Leipzig,” *Pertanika J. Soc. Sci. Humanit.*, 2022, doi: 10.47836/pjssh.30.4.13.

[19] C. S. Andreassen et al., “The Relationship Between Addictive Use of Social Media and Video Games and Symptoms of Psychiatric Disorders: A Large-Scale Cross-Sectional Study.,” *Psychol. Addict. Behav.*, 2016, doi: 10.1037/adb0000160.

[20] F. Ahmed, J. R. Carrión, F. Bellotti, G. Barresi, F. Floris, and R. Berta, “Applications of Serious Games as Affective Disorder Therapies in Autistic and Neurotypical Individuals: A Literature Review,” *Appl. Sci.*, 2023, doi: 10.3390/app13084706.

[21] T. S. Um, R. L. Ownby, and S. Chou, “Mapping the Mind in the Virtual Metaverse: An Initial in-Depth Thematic Exploration of Youth Mental Health Within VRChat,” 2024, doi: 10.1101/2024.09.28.24314456.

[22] H. M. Saleem, K. Dillon, S. Benesch, and D. Ruths, “A Web of Hate: Tackling Hateful Speech in Online Social Spaces,” 2017, doi: 10.48550/arxiv.1709.10159.

[23] P. B. Thomas, D. Riehm, M. Glenski, and T. Weninger, “Behavior Change in Response to Subreddit Bans and External Events,” 2021, doi: 10.48550/arxiv.2101.01793.

[24] R. L. Mandryk et al., “Combating Toxicity, Harassment, and Abuse in Online Social Spaces: A Workshop at CHI 2023,” 2023, doi: 10.1145/3544549.3573793.

[25] J. Klapproth, S. Unger, J. Pohl, S. Boberg, C. Grimme, and T. Quandt, “Immunize the Public Against Disinformation Campaigns: Developing a Framework for

Analyzing the Macrosocial Effects of Prebunking Interventions," 2023, doi: 10.24251/hicss.2023.298.

[26] A. Hess and C. D. Flores, "Simply More Than Swiping Left: A Critical Analysis of Toxic Masculine Performances On *Tinder Nightmares*," *New Media Soc.*, 2016, doi: 10.1177/1461444816681540.

[27] N. Martin-Anatias, "Language Selection in the Indonesian Novel: *Bahasa Gado-Gado* in Expressions of Love," *South East Asia Res.*, 2018, doi: 10.1177/0967828x18809592.

[28] M. Fareed, S. Humayun, and H. Akhtar, "English Language Teachers' Code-Switching in Class: ESL Learners' Perceptions," *J. Educ. Soc. Sci.*, 2016, doi: 10.20547/jess0411604101.

[29] E. R. Maha, Z. Zainuddin, and I. W. Dirgeyasa, "Code Switching in Sarah Sechan Talk Show Program on Net Tv," *Linguist. Terap.*, 2018, doi: 10.24114/lt.v14i1.8358.

[30] A. B. Kandiawan, "Code-Switching and Slang Used by Gen Z Indonesians on Social Media," *Eltr J.*, 2022, doi: 10.37147/eltr.v7i1.165.

[31] I. U. Sumaryanti and J. Yuniar, "The Implication of Social Media Toward College Students' Online Behavior in Bandung," *Mediat. J. Komun.*, 2022, doi: 10.29313/mediator.v15i1.8845.

[32] H. Sahib, W. Hanafiah, M. Aswad, A. H. Yassi, and F. Mashhadi, "Syntactic Configuration of Code-Switching Between Indonesian and English: Another Perspective on Code-Switching Phenomena," *Educ. Res. Int.*, 2021, doi: 10.1155/2021/3402485.