Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1834~1842
e-ISSN: 2548-6861

Sentiment Classification for Tabletop Entertainment Discussions Using BERT and VADER Weak-Agreement Rules

Alberto Halim Limantoro [1]*, Adi Suryaputra Paramita [2]*
***Information System for Business, Ciputra University, Indonesia
alimantoro@student.ciputra.ac.id[1], adi.suryaputra@ciputra.ac.id[2]

Article Info

Article history:

Received 2026-01-23
Revised 2026-03-02
Accepted 2026-04-18

Keyword:

Community Forum,
Machine Learning,
Sentiment Analysis,
Support Vector Machine,
Tabletop Entertainment.

ABSTRACT

This research analyzes public sentiment in online tabletop entertainment communities using a hybrid approach that combines lexicon-based and machine-learning methods. Public opinion in the dataset was gathered from Reddit, given its growing reach among hobbyists, and categorized into three sentiment classes: positive, neutral, and negative. The pre-processing stage removes redundant entries, yielding 1494 unique posts. For the model evaluation phase, a relatively balanced testing dataset of 800 entries (500 actual posts and 300 constructed samples) was used, with sentiment distribution of 384 positive, 216 negative, and 200 neutral. Using BERT (bert-base-uncased), the final model achieved an accuracy of 0.8875 and a macro F1-score of 0.8712, with class-level performance as follows: negative (Precision 0.95, Recall 0.95, F1-score 0.95), neutral (Precision 0.97, Recall 0.62, F1-score 0.76), and positive (Precision 0.83, Recall 0.99, F1-score 0.90). The integration of lexicon-based (VADER) and transformer-based (BERT) methods for long-form tabletop community discussions represents the main contribution of this study and can be applied to various sentiment analysis applications. In the exploratory analysis of the actual dataset (1494 posts), sentiment distribution shows dominance of the positive category with 1151 posts (77.04%), followed by 306 neutral (20.48%) and 37 negative (2.48%), indicating real-world class imbalance despite strong balanced-test performance.

I. INTRODUCTION

Online forums have grown in the recent decade, becoming a major platform for discussion, information exchange, and community making across a wide range of interests. As more users rely on digital platforms to share opinions and experiences, these forums generate large volumes of text that reflect real-time public sentiment [1]. Communities centered around hobbies such as tabletop entertainment are particularly active, with discussions covering game reviews, player experiences, design opinions, and industry trends [4]. This increased activity creates both opportunities and challenges for understanding user attitudes, as the informal and highly varied language used in these discussions requires specialized analytical approaches [5].

Online discussion platforms have become a hotspot for users to express their opinions on different types of subject across different mediums [7]. Alongside this growth would be communities centered on tabletop entertainment. As discussion threads continue to grow rapidly alongside the increasing popularity of digital communication channels. However, high engagement within these communities doesn't always reflect user satisfaction, and the sentiment in long-form discussions often differs from surface level indicators such as post popularity or upvotes [6]. This creates inconsistencies between quantitative engagement metrics and the actual tone of user discourse, which can obscure community needs and expectations. Extracting public opinion using machine learning poses more challenges due to variations in writing style, slang, informal expressions, and specific terminology commonly found in hobbyist forums. As a result, sentiment analysis is required to process these large volumes of text and generate a more accurate understanding of user attitudes within the tabletop entertainment ecosystem [8].

Sentiment analysis is the process of extracting, processing, and interpreting data, usually in the form of text, and identifying the sentiment embedded within user-generated opinions. This algorithm functions by finding the patterns and hyperlanes that differentiate between positive sentiment and negative sentiment [9]. In the context of online tabletop entertainment forums, this process involves identifying how players express their experiences, preferences, and critiques regarding games or community interactions.

The VADER lexicon based model is used for sentiment analysis because it is specifically designed to handle informal, user-generated text and provides fast, rule-based classification across positive, neutral, and negative categories [10]. The algorithm calculates sentiment intensity using a combination of lexical scoring and heuristic rules that model punctuation, capitalization, degree modifiers, and contextual cues, allowing it to perform with high accuracy on social-media style datasets [11].

BERT classifiers is the ability to capture dependencies across entire sentences rather than assuming independence between features, which allows them to outperform traditional statistical methods in many text-analysis tasks [13].. Sentiment analysis using these hybrid approaches is widely applied in entertainment, marketing, and online-community research to analyze public opinion and patterns of discourse at scale. In this study, the Python programming language is used throughout the text mining process, from data preprocessing to sentiment labeling using both VADER and BERT classifier [14].

The combination of VADER and BERT is motivated by their complementary strengths in theory. VADER is effective in capturing explicit sentiment cues in informal social-media language, while BERT is capable of modeling deeper contextual relationships within long-form discussions. By integrating both approaches, the model might be able to balance lexical sensitivity with contextual understanding.

The weak-agreement mechanism is applied to reduce noise in automatic labeling by selecting instances where both models show consistent sentiment tendencies. This strategy is intended to improve labeling reliability compared to relying on a single model or simple majority voting, particularly in informal community-generated text.

Studies using either or both methods of VADER and BERT have been done before. In the case of VADER, a similar study was conducted in order to analyze Covid-19 tweets from Twitter. The results from said study indicate a modest performance from the VADER model alone with an average accuracy rate of 62%, precision of 52%, recall of 64%, and f1-score of 55% [15]. The next study is a research on sentiment analysis using the BERT model to classify online consumer reviews. The findings show that BERT results with a precision of 88.09%, recall of 86.22%, F1-measure of 89.41%, and an accuracy of 88.48% [16].

This study was conducted using data from online discussions within tabletop entertainment communities on Reddit, specifically from the subreddits r/Boardgames, r/Tabletop, r/DnD, r/Warhammer, and r/rpg, using a hybrid sentiment-analysis method that combines VADER and BERT. Although previous studies have examined sentiment analysis using VADER or BERT in domains such as social media and consumer reviews, research specifically focusing on discussions regarding tabletop entertainment remains limited.

The goal of this study is to evaluate public sentiment in these communities to generate clearer and more accurate insights that can be used to understand community behavior, discussion trends, and attitudes toward tabletop games, systems, and related media.

By integrating lexicon-based and transformer-based approaches through a hybrid mechanism, this study aims to address the gap in domain-specific sentiment modeling for hobbyist online communities. Unlike previous studies that rely on a single-model approach, this research proposes a domain-specific hybrid framework tailored to long-form hobbyist discussions. Therefore, this study focuses on analyzing sentiment from 1494 Reddit posts across multiple tabletop entertainment subreddits over a period of 4 months. Specifically, the study classifies posts into positive, negative, and neutral categories to identify the types of discussions that dominate the community, highlight concerns and excitement points among users, and provide deeper insights into sentiment trends within the tabletop entertainment discussions.

This study contributes by providing one of the few sentiment analyses focused specifically on tabletop entertainment communities. It also offers empirical insights into class imbalance behavior and hybrid model performance within a niche but highly active online domain.

II. METODE

In carrying out this research, several stages were implemented to ensure the accuracy and reliability of the sentiment-analysis results. The first stage consists of collecting Reddit posts from multiple tabletop entertainment subreddits over a four-month period using the Reddit API. The second stage involves data preprocessing, including text normalization, removal of URLs, punctuation cleaning, lowercasing, and duplicate filtering.

Next is sentiment labeling using a hybrid approach that combines VADER and a BERT classifier, producing the sentiment assignments used for analysis. Model outputs are then validated through accuracy, precision, recall, and F1-score evaluations to measure the performance of each component in the hybrid system using manually labelled dataset in order to gauge performance. Finally, validated sentiment outcomes are processed and visualized to highlight overall sentiment trends, class distribution, and the dominant emotional tone within tabletop-entertainment discussions.

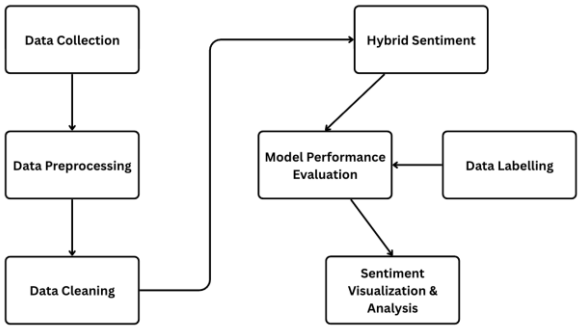

Figure 1. Research Method

A. Research Design

This research was conducted based on problems previously explained, namely to determine user sentiment regarding their experience in the tabletop community. This analysis will yield sentiment values of positive, negative, and neutral based on the user's post.

B. Data Collection & Data Preprocessing

The data preprocessing in this research starts with preparing Reddit posts for accurate sentiment classification using Jupyter Notebook. This stage begins with data scraping using the Reddit API and to process said data using text normalization, which includes converting all characters to lowercase, removing URLs, special characters, excessive whitespace, and non-informative tokens that may interfere with model interpretation.

Reddit was selected as the primary data source because it hosts active and topic focused hobby communities that allow users to share detailed experiences and opinions. Data were collected from the subreddits r/dnd, r/rpg, r/warhammer, r/tabletop, and r/boardgames using the official Reddit API. The collection period spanned from late September to late December. A total of 1,495 posts were gathered for exploratory analysis. No strict filtering criteria were applied during scraping, although the dataset predominantly consists of English language posts due to the nature of the selected communities.

Stopword removal is applied to eliminate common words that do not contribute meaningful sentiment signals. Lemmatization is then used to reduce words to their base forms, allowing models to recognize semantically similar expressions more effectively. Jupyter Notebook is a web-based interactive computing platform used for using Python in this study.

Additional preprocessing steps included replacing bracket symbols and markdown artifacts with normalized text representations to avoid tokenization inconsistencies. Slang expressions and community-specific terminology were intentionally preserved to maintain contextual authenticity. URLs, emojis, and non-textual elements were removed to reduce noise. No aggressive normalization was applied in order to retain natural linguistic patterns commonly found in online discussion forums.

C. Data Cleaning

Data cleaning focuses on removing duplicate entries, filtering out posts with missing or empty text fields, and eliminating noise such as links, emojis, markdown artifacts, and non-ASCII characters. Irrelevant posts that do not contain meaningful discussion are also discarded. By applying these cleaning steps, the dataset becomes more reliable and representative, allowing the sentiment models to operate on high-quality input.

D. Data Labelling & Hybrid Sentiment Labeling

TABLE I
SAMPLE OF DATA LABELING

No.	Post	Label
1	which game do you think would make the strongest ttrpg concept	pos
2	some games are so hard to find	neg
3	this was a goofy little project i worked on its fully open source and you can check it out here	pos
4	What do you think of the new expansion pack?	neu

Data labeling is done manually with sample data labeled as positive, negative, and neutral. Labeling is done with around 500 true labeled samples out of 1494 rows with also an extra 300 to even out the distribution of sentiment. Labeling is a process that can facilitate the model in classifying sentiment classification.

This stage is carried out to ensure that the sentiment output becomes more consistent and reliable before evaluation. In this process, the sentiment results produced by VADER and the BERT model are combined into a single final label. The steps performed include extracting polarity scores from VADER, generating contextual predictions from BERT, and applying a rule-based decision layer to merge both outputs into a unified positive, neutral, or negative sentiment category.

The hybrid mechanism follows a weak-agreement strategy in which BERT serves as the primary classifier and VADER acts as a supporting rule-based validator. When both models produce the same sentiment label, that label is directly assigned as the final output. In cases of disagreement, the decision prioritizes BERT predictions due to its contextual understanding capability. However, when BERT exhibits low confidence in its prediction, the VADER polarity score is used as a corrective signal. This rule-based decision layer aims to balance contextual deep-learning classification with lexicon-based sentiment scoring.

E. Model Evaluation

This stage is carried out to ensure that the sentiment output becomes more consistent and reliable before evaluation. In this process, the sentiment results produced by VADER and the BERT model are combined into a single final label.

The contextual classifier used in this study is based on the BERT base uncased architecture. The model was trained for 10 epochs using 800 data samples from real and fabricated comments, training data size is set to a 30/70 ratio. The learning rate was initialized using a standard fine-tuning configuration.

The labeled dataset consisted of 800 manually annotated samples. These samples are from real data carried from the scraped data with fabricated comments to balance out the testing phase, around 500 are from real posts while 300 are fabricated to ensure even distribution, with class distribution in the test set as follows: 384 positive, 216 negative, and 200 neutral posts. The model was used as the primary decision component within the hybrid framework.

The steps performed include extracting polarity scores from VADER, generating contextual predictions from BERT, and applying a rule-based decision layer to merge both outputs into a unified positive, neutral, or negative sentiment category. This stage also tests the accuracy of the model in order to determine the validity of the model.

To reduce bias caused by class imbalance, the manually labeled evaluation dataset was adjusted to achieve a relatively balanced distribution across positive, negative, and neutral categories. This balancing step ensures that performance metrics such as precision, recall, and F1-score provide a more reliable representation of model capability across all sentiment classes.

Accuracy is used to measure the proportion of correctly predicted sentiment labels and is calculated using the Accuracy, Precision, Recall, and f1-score formula

Accuracy is used to measure the proportion of correctly predicted sentiment labels

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision measures the accuracy of the model in identifying a specific sentiment class correctly:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity) measures the model's ability to retrieve all instances of a sentiment class

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

The F1-Score represents the harmonic mean of precision and recall, providing a balanced summary of model performance

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

F. Visualization

Visualization converts the sentiment results into simple graphical summaries, such as bar charts, to show the proportion of positive, neutral, and negative sentiments. This helps reveal overall trends in tabletop-entertainment discussions and supports easier interpretation of the analysis results.

III. Results and Discussions

This section focuses on the findings from the sentiment analysis of online tabletop entertainment discussions on Reddit using a hybrid VADER and BERT approach while also comparing it to standalone model performance. The section will be separated into 3 subjects which are the model's results such as performance testing using the 800 manually annotated samples, the sentiment analysis of the original 1494 posts, and the discussions around the use of this data and the discussions on the hybrid model itself.

A. Model Results

This section begins with an explanation of the data sampling method before proceeding to model evaluation. The dataset consists of Reddit posts collected from selected tabletop entertainment subreddits over a defined period, followed by preprocessing to remove duplicates and irrelevant entries.

For model development and evaluation, a relatively balanced dataset was prepared to ensure fair performance assessment across positive, neutral, and negative classes. This separation between exploratory data distribution and balanced evaluation data allows for a clearer interpretation of model behavior under controlled and real-world conditions.

```
<class 'pandas.DataFrame'>
RangeIndex: 800 entries, 0 to 799
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      800 non-null    str
 1   text    800 non-null    str
 2   result  800 non-null    str
 3   label   800 non-null    int64
dtypes: int64(1), str(3)
memory usage: 495.0 KB
Unique labels: <ArrowStringArray>
['neu', 'pos', 'neg']
Length: 3, dtype: str
Number of unique labels: 3

Value counts:
result
pos    384
neg    216
neu    200
Name: count, dtype: int64
```

Figure 2. Test Dataset Used for Machine Learning

The testing dataset consists of 800 entries, including 500 real-world Reddit posts and 300 manually constructed samples. The real-world portion reflects the natural sentiment distribution observed in the community, resulting in a degree of class imbalance (384 positive, 216 negative, and 200 neutral). This imbalance is intentional, as the study aims to preserve real-world sentiment characteristics rather than artificially enforcing perfect class symmetry. The additional 300 manually labeled samples were introduced to improve minority-class representation and stabilize model evaluation, while still maintaining a distribution that approximates real community conditions.

This approach ensures that the model is evaluated under conditions that remain realistically skewed, allowing performance assessment to better reflect practical deployment scenarios.

Next, the methods used in this study are described. The approaches applied include the lexicon based VADER model, the BERT base uncased model, and a hybrid combination of both using weak agreement rules, as explained in the methodology section. All three classification approaches, VADER, BERT , and the proposed hybrid model, were evaluated using a 70:30 train test split. The models were trained for five epochs using standard training procedures without additional optimization or advanced tuning strategies.

This consistent configuration ensures a fair comparison across methods, allowing performance differences to reflect model capability rather than variations in the training setup.

```
=== VADER RESULTS ===
Accuracy: 0.625
F1 (macro): 0.5794612587552751
Confusion Matrix:
 [[91  4 20]
 [36 21  3]
 [11 16 38]]
              precision    recall  f1-score   support

         pos       0.66      0.79      0.72       115
         neu       0.51      0.35      0.42        60
         neg       0.62      0.58      0.60        65

    accuracy                           0.62       240
   macro avg       0.60      0.58      0.58       240
weighted avg       0.61      0.62      0.61       240
```

Figure 3. Results from VADER only

The VADER model achieved an overall accuracy of 0.625 with a macro F1-score of 0.579. These results show moderate classification performance across the three sentiment categories. While VADER is designed to handle informal and social-media-style text, its rule-based nature limits its ability to fully capture contextual nuance in longer discussion posts.

Based on the confusion matrix and class-level metrics, VADER performed strongest in the positive class, achieving a precision of 0.66, recall of 0.79, and F1-score of 0.72. Performance on the negative class was moderate, with an F1-score of 0.60, while the neutral class showed weaker detection capability, with an F1-score of 0.42.

The lower recall for neutral sentiment suggests difficulty in distinguishing informational or context-dependent statements from clearly positive or negative expressions, highlighting the limitations of purely lexicon-based approaches in complex discussion environments.

```
=== BERT (Rebuilt) ===
Accuracy: 0.8166666666666667
F1 Macro: 0.8083210046375667
Confusion Matrix:
 [[60  2  3]
 [ 2 38 20]
 [ 0 17 98]]

=== FULL CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

         neg     0.9677    0.9231    0.9449        65
         neu     0.6667    0.6333    0.6496        60
         pos     0.8099    0.8522    0.8305       115

    accuracy                         0.8167       240
   macro avg     0.8148    0.8029    0.8083       240
weighted avg     0.8168    0.8167    0.8163       240
```

Figure 4. Results from BERT only

The BERT base uncased model demonstrates significantly stronger performance compared to the lexicon-based approach, achieving an overall accuracy of 0.8167 and a macro F1-score of 0.8083. This improvement reflects BERT's ability to capture contextual relationships within longer Reddit discussions, allowing it to better distinguish between subtle sentiment variations across positive, neutral, and negative classes.

Based on the confusion matrix and classification report, BERT performs particularly well in detecting negative sentiment, achieving a precision of 0.9677, recall of 0.9231, and an F1-score of 0.9449.

The positive class also shows strong performance with an F1-score of 0.8305. However, the neutral class remains comparatively weaker, with an F1-score of 0.6496, indicating ongoing difficulty in differentiating purely informational or context-dependent statements. Overall, the results suggest that transformer-based modeling substantially improves sentiment classification accuracy compared to rule-based methods, particularly in handling minority or nuanced sentiment expressions.

```
=== HYBRID (BERT + VADER) ===
Accuracy: 0.8916666666666667
F1 (macro): 0.877749185420929
Confusion Matrix:
 [[ 59   4   2]
 [  1  41  18]
 [  0   1 114]]
              precision    recall  f1-score   support

         neg       0.98      0.91      0.94        65
         neu       0.89      0.68      0.77        60
         pos       0.85      0.99      0.92       115

    accuracy                           0.89       240
   macro avg       0.91      0.86      0.88       240
weighted avg       0.90      0.89      0.89       240
```

Figure 5. Results from Hybrid

The hybrid model combining BERT and VADER achieves the strongest overall performance among the three approaches, with an accuracy of 0.8917 and a macro F1-score of 0.8777.

This improvement suggests that integrating contextual deep learning with lexicon based polarity validation enhances classification robustness. The results indicate that the hybrid mechanism successfully leverages the strengths of both models, leading to more balanced and consistent sentiment detection across classes.

Based on the confusion matrix and class-level metrics, the hybrid model performs exceptionally well in the negative class, achieving a precision of 0.98, recall of 0.91, and an F1-score of 0.94. The positive class also demonstrates strong performance with an F1-score of 0.92 and a high recall of 0.99, indicating effective detection of dominant community sentiment.Notably, performance in the neutral class improves compared to the individual models, reaching an F1-score of 0.77.

These findings demonstrate that the hybrid approach provides some improvement over standalone VADER and BERT models, particularly in achieving more stable performance across minority and majority sentiment categories.

B. Sentiment Analysis Results

This section presents the sentiment analysis results derived from the 1494 original Reddit posts that were processed using the latest hybrid model, focusing on overall distribution patterns and temporal dynamics. First, the graph of sentiment ratios illustrates the proportion of positive, neutral, and negative posts within the dataset, providing an overview of dominant community attitudes. This visualization allows for an immediate assessment of whether discussions are primarily supportive, critical, or informational in nature. By examining these ratios, broader sentiment tendencies within tabletop entertainment discussions can be identified before exploring temporal variation.

Next, the analysis extends to a time-based graph showing sentiment ratios across the full four-month observation period. This longitudinal view highlights fluctuations in community mood, revealing whether sentiment remains stable or shifts in response to specific events, releases, or discussions. Finally, a correlation matrix is introduced to examine the statistical relationships between sentiment categories over time, offering insight into how changes in one sentiment class may relate to increases or decreases in others. Together, these visualizations provide a structured understanding of both distributional and temporal sentiment behavior within the dataset.

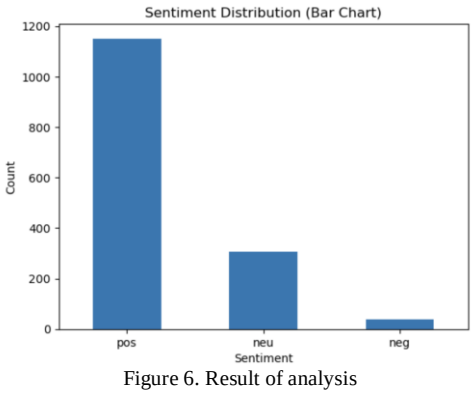

Figure 6. Result of analysis

This graph shows the overall sentiment ratio derived from the 1494 Reddit posts, where positive sentiment clearly dominates the discussion (1151 posts), followed by neutral (306 posts) and a relatively small proportion of negative content (37 posts). The distribution indicates that tabletop entertainment discussions within the observed subreddits are largely supportive and enthusiasm-driven, with criticism appearing only occasionally. The limited presence of negative sentiment suggests that community interactions tend to focus more on sharing experiences, recommendations, and excitement rather than sustained dissatisfaction.

Based on the manually annotated sampling data presented in the previous section, this ratio is consistent with the observed real world distribution, supporting the validity of the model's classification outcomes. The alignment between manual annotation patterns and the full dataset prediction strengthens confidence that the hybrid model accurately reflects the community's overall sentiment tendencies rather than artificially inflating positive classifications.

The 500 sampled data had roughly the same ratio as shown here with around 70% positives, this is way the previous work was shown to be imbalance and why we added the more balance version for the data sampling when testing the model

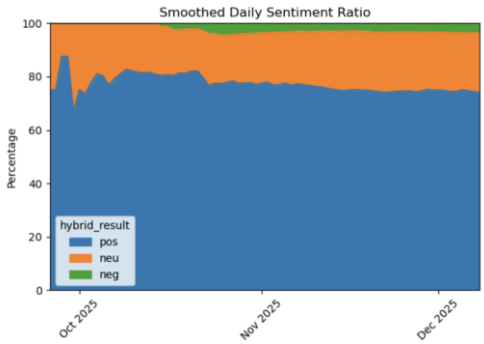

Figure 7. Ratio over the period of 4 months

This graph shows the smoothed daily sentiment ratio across the four-month observation period. Positive sentiment consistently dominates throughout the timeline, remaining relatively stable at around three-quarters of total daily discussions. Neutral sentiment occupies a smaller but steady portion, while negative sentiment remains minimal and only fluctuates slightly near the baseline. Although minor variations are visible, there are no extreme spikes or sharp reversals, suggesting that overall community mood remains consistently positive over time.

The gradual smoothing of the data highlights underlying trends rather than short-term noise, indicating that sentiment stability is not the result of isolated high activity days. The slight decline in positive ratio toward the later months is accompanied by a small increase in neutral proportion, rather than a surge in negativity. This pattern suggests that changes in discussion tone are more reflective of shifts toward informational or balanced conversations rather than increased dissatisfaction within the community. Although it might be due to the fact that earlier comments from that date might be

less then when it's in December due to spiking from communities around that time.

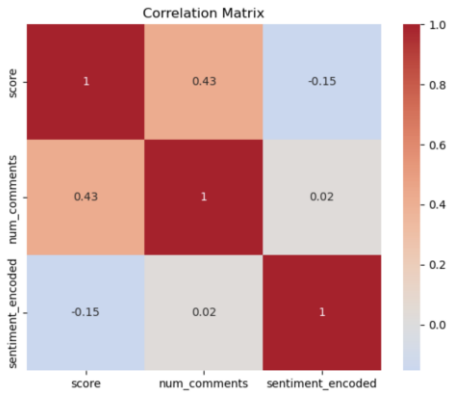

Figure 8. Correlation Matrix

This correlation matrix illustrates the relationships between sentiment score, number of comments, and encoded sentiment category. A moderate positive correlation (0.43) is observed between sentiment score and number of comments, suggesting that posts with stronger sentiment polarity whether positive or negative tend to generate higher engagement. This indicates that emotionally expressive discussions may encourage more interaction within the community.

In contrast, the correlation between encoded sentiment class and sentiment score is weakly negative (-0.15), while the relationship between encoded sentiment and number of comments is nearly negligible (0.02). These low values suggest that categorical sentiment labels alone do not strongly determine engagement levels. Overall, the matrix indicates that sentiment intensity appears more associated with interaction volume than the sentiment category itself, reinforcing the importance of considering both polarity strength and classification in community sentiment analysis.

C. Discussion

Regarding the Model, The results show that the hybrid VADER–BERT model achieves the highest overall performance among the evaluated approaches, with an accuracy of 0.8917 and a macro F1-score of 0.8777. This improvement over standalone VADER and BERT indicates that combining rule-based and transformer-based strategies enhances both general predictive power and class-level balance. The integration through weak-agreement rules helps reduce isolated misclassifications and stabilizes sentiment predictions across categories.

Methodologically, the hybrid framework mitigates the individual weaknesses of each model. VADER provides explicit polarity validation suited for informal Reddit language but lacks contextual depth, while BERT captures semantic nuance yet may overgeneralize in ambiguous cases. Their combination allows lexical polarity cues to complement contextual inference, leading to particularly strong performance in the negative class and noticeable improvement in neutral sentiment detection—an area traditionally difficult due to subtle or informational phrasing.

The improvement in macro F1-score further suggests that the hybrid model performs more consistently across majority and minority classes under naturally imbalanced conditions. Given that positive sentiment tends to dominate hobby-related Reddit communities, this balanced performance indicates greater robustness for real-world deployment. Overall, the findings support the idea that complementary modeling strategies are more effective than relying on either rule-based or deep learning methods alone for nuanced, community-generated discussions.

However, several limitations must be acknowledged. The dataset size is relatively modest, and the inclusion of manually constructed samples may introduce annotation bias. The use of a simple 70:30 split without extensive hyperparameter tuning may not reflect optimal model performance. Additionally, the weak-agreement integration relies on heuristic rules that may not generalize across domains, and the three-class labeling scheme may oversimplify complex expressions such as sarcasm or mixed sentiment.

Furthermore, the overall scale of the collected Reddit posts (1494 entries) and the sample evaluation set may limit the statistical representativeness of the findings. While sufficient for comparative modeling, a larger corpus could provide stronger generalization and reduce sampling variance. The selected tabletop entertainment subreddits also represent only a portion of the broader tabletop community, and may reflect specific cultural norms, moderation styles, or audience demographics. As such, the results may not fully capture sentiment patterns across other tabletop communities, platforms, or offline discussions. Future research should expand dataset scale, diversify community sources, and explore cross-domain validation or more fine-grained sentiment frameworks to strengthen robustness and generalizability.

In regards to the sentiment analysis part, overall the sentiment analysis results indicate that online tabletop entertainment discussions on Reddit are predominantly positive, with neutral content forming a stable secondary portion and negative sentiment appearing only minimally. The consistency of this distribution across both the aggregate ratio graph and the four-month temporal trend suggests that the community maintains a generally supportive and enthusiasm-driven atmosphere. Minor fluctuations over time do not indicate structural shifts in sentiment, but rather natural variations in discussion topics. Additionally, the correlation analysis shows that engagement is more closely associated with sentiment intensity than with sentiment category alone, implying that emotionally expressive posts tend to generate more interaction regardless of whether they are positive or negative.

From a practical perspective, these findings have several real-world applications. For community managers and moderators, understanding dominant sentiment patterns can

help monitor community health and detect early signs of dissatisfaction or conflict.

For tabletop publishers, designers, and marketers, sentiment trends can provide insight into audience reception of products, rule changes, or new releases, supporting data-driven decision-making. More broadly, the hybrid sentiment modeling approach demonstrated in this study can be applied to other hobby-based or niche online communities to track audience perception, evaluate engagement dynamics, and support strategic communication planning. By combining contextual and lexicon-based methods, organizations can obtain more reliable sentiment insights in informal, user-generated discussion environments.

It's noted that the current work represents a revised version of the study, in which the implementation code was restructured and optimized to improve overall efficiency and clarity. Several components of the hybrid modeling pipeline were refined, including data preprocessing flow, model integration logic, and evaluation procedures. These adjustments were made to reduce redundancy, improve computational performance, and ensure that the sentiment classification process operates in a more streamlined and reproducible manner.

In addition to technical optimization, the revision also focused on improving methodological consistency between data labeling, model validation, and visualization stages. By refining the weak-agreement mechanism and organizing the workflow more systematically, the updated implementation provides a clearer structure for combining lexicon-based and transformer-based approaches within a single analytical framework.

For researchers interested in applying this hybrid method, it is recommended to carefully adapt the integration rules and preprocessing steps to the linguistic characteristics of their specific domain. While the VADER–BERT combination can improve robustness, optimal performance depends on dataset size, class distribution, and contextual variability. Future researchers are encouraged to conduct cross-domain validation, experiment with hyperparameter tuning, and consider expanding beyond three-class sentiment schemes to enhance generalizability and analytical depth.

IV. CONCLUSION

This study examined sentiment within online tabletop entertainment communities on Reddit using a hybrid sentiment-analysis framework that integrates VADER and BERT. By analyzing 1,494 posts collected over a four-month period, the research aimed to identify dominant sentiment patterns and evaluate the effectiveness of combining lexicon based and transformer based models. The hybrid approach was proposed to address the limitations of relying on a single-model method in informal, long-form discussion environments.

The model evaluation results demonstrate that the hybrid VADER-BERT framework outperforms standalone approaches, achieving the highest accuracy and macro F1-score among the tested models. The integration through weak-agreement rules improves classification stability and enhances balance across positive, neutral, and negative classes. In particular, improvements in neutral sentiment detection highlight the benefit of combining contextual understanding with polarity validation.

The sentiment analysis results reveal that discussions within the selected tabletop subreddits are predominantly positive, with neutral posts forming a steady secondary portion and negative sentiment appearing minimally. Temporal analysis across the four-month period indicates that this distribution remains relatively stable, suggesting a consistently supportive and enthusiasm-driven community atmosphere. Correlation analysis further shows that engagement levels are more closely associated with sentiment intensity than sentiment category alone.

Despite these promising findings, several limitations should be considered. The dataset size and evaluation sample remain relatively modest, and the inclusion of manually constructed samples may introduce bias. Additionally, the selected subreddits may not fully represent the broader tabletop community, limiting generalizability across other platforms or demographic groups. The reliance on heuristic integration rules and a three-class labeling scheme may also oversimplify complex expressions such as sarcasm or mixed sentiment.

Overall, this study contributes to sentiment analysis research by providing domain-specific insights into tabletop entertainment discussions and demonstrating the practical value of hybrid modeling strategies. The proposed framework shows potential for broader application in hobby-based online communities, supporting community monitoring, audience analysis, and data-driven decision-making. Future research should expand dataset scale, diversify data sources, and explore more advanced or fine-grained sentiment modeling techniques to strengthen robustness and applicability.

REFERENCE

[1] Nenny Anggraini, Syopiansyah Jaya Putra, Luh Kesuma Wardhani, Farid, Nashrul Hakiem, and Imam Marzuki Shofi, "A Comparative Analysis of Random Forest, XGBoost, and LightGBM Algorithms for Emotion Classification in Reddit Comments," Jurnal Teknik Informatika, vol. 17, no. 1, pp. 88–97, May 2024, doi: https://doi.org/10.15408/jti.v17i1.38651.

[2] R. D. . Kurniawan, A. . Yohannis, and W. T. . Atmojo, "Sentiment Analysis of Getcontact Application Reviews on Google Play Store Using Naive Bayes Algorithm", J. Tek. Inform. (JUTIF), vol. 6, no. 4, pp. 2848–2858, Sep. 2025.

[3] D. C. Youvan, "Understanding Sentiment Analysis with VADER: A Comprehensive Overview and Application," ResearchGate, Jun. 2024, doi: https://doi.org/10.13140/RG.2.2.33567.98726.

[4] S. Naghikhani, "Beyond The Box," 2024. Available: https://openresearch.ocadu.ca/id/eprint/4419/1/Naghikhani_Sourena_2024_MDES_SFI_MRP.pdf

[5] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A Survey on Sentiment Analysis Methods, Applications, and Challenges," Artificial Intelligence Review, vol. 55, no. 55, Feb. 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[6] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," Artificial

Intelligence Review, vol. 56, Mar. 2023, doi: https://doi.org/10.1007/s10462-023-10442-2.
[7] Y. Wang, M. Zhang, N. Luo, and L. Guo, "Understanding how participating behaviours influenced by individual motives affect continued generating behaviours in product-experience-shared communities," Behaviour & Information Technology, pp. 1–21, Sep. 2021, doi: https://doi.org/10.1080/0144929x.2021.1970807.
[8] J. Donald, J. M. Banner, R. Satria, Winema Tania, and W. James, "Sentiment analysis of user-generated content," Oct. 20, 2024. https://www.researchgate.net/publication/385084925_Sentiment_analysis_of_user-generated_content
[9] "School Manager by Family Zone," Google.co.id, 2025. https://books.google.co.id/books?hl=en&lr=&id=xYhyEAAAQBAJ&oi=fnd&pg=PP1&dq=Sentiment+analysis+is+the+process+of+extracting (accessed Dec. 07, 2025).
[10] J. Al-Garaady and M. M. Albuhairy, "Public Sentiment Analysis in Social Media on the SARS-CoV-2 Vaccination Using VADER Lexicon Polarity," Ssrn.com, Mar. 2022. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3647564 (accessed Dec. 11, 2025).
[11] "Sentimental Analysis for Political Polarization Using VADER Sentiment Lexicon," Journal of Xidian University, vol. 15, no. 5, May 2021, doi: https://doi.org/10.37896/jxu15.5/014.
[12] [12]R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," Electronics, vol. 11, no. 3, p. 374, Jan. 2022, doi: https://doi.org/10.3390/electronics11030374.
[13] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based BERT model in sentiment analysis," Applied and Computational Engineering, vol. 71, no. 1, pp. 14–20, May 2024, doi: https://doi.org/10.54254/2755-2721/71/2024ma.
[14] A. J. Dhruv, R. Patel, and N. Doshi, "Python: The Most Advanced Programming Language for Computer Science Applications," Proceedings of the International Conference on Culture Heritage, Education, Sustainable Tourism, and Innovation Technologies, vol. 1, 2020, doi: https://doi.org/10.5220/0010307902920299.
[15] V. Arya, Amit Kumar Mishra, and A. González-Briones, "Sentiments Analysis of Covid-19 Vaccine Tweets Using Machine Learning and Vader Lexicon Method," Advances in distributed computing and artificial intelligence journal, vol. 11, no. 4, pp. 507–518, Jun. 2023, doi: https://doi.org/10.14201/adcaij.27349.
[16] M. P. Geetha and D. Karthika Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model," International Journal of Intelligent Networks, vol. 2, pp. 64–69, 2021, doi: https://doi.org/10.1016/j.ijin.2021.06.005.