Sentiment Analysis of FIFA-Related Tweets: Integrating NLTK's VADER with BERT for Enhanced Classification Accuracy

Mirsad Hadžić
International Burch University
Francuske revolucije bb

Zerina Mašetić
International Burch University
Francuske revolucije bb

Fatima Mašić
International Burch University
Francuske revolucije bb

ABSTRACT
The purpose of this research is to perform sentiment analysis on Twitter data using Natural Language Processing (NLP) techniques, particularly leveraging the NLTK library in Python within a Jupyter notebook environment. The study aims to explore sentiment classification methods, evaluating the emotional tone of tweets and categorizing them as neutral, positive, or negative sentiments, utilizing NLTK's SentimentIntensityAnalyzer. The sample consists of Twitter data with columns like 'Tweet' and 'Sentiment' sourced from a CSV file. The methodology involves tokenizing and processing the text, grading sentiment, counting occurrences of the hashtag #fifa, and analyzing word frequencies [1].

In addition to the lexicon-based VADER approach, the study incorporates a transformer-based deep learning model—BERT (Bidirectional Encoder Representations from Transformers) -to enhance sentiment classification accuracy. BERT, pre-trained on large corpora and capable of understanding context and nuanced language, offers a state-of-the-art alternative to traditional models. This inclusion allows a comparative analysis between rule-based and deep learning approaches, highlighting BERT's effectiveness in handling complex tweet structures. Furthermore, the study investigates the impact of removing stopwords and explores the list of eliminated stopwords. The expected results include gaining insights into prevalent sentiments on Twitter regarding a specified topic, frequency of the hashtag #fifa, and a comprehensive understanding of word usage, visually depicted through wordclouds. Possible limitations include inherent subjectivity in sentiment analysis, potential variations in language use, reliance on hashtag frequency as an indicator of topic prevalence, and the effectiveness of stopword removal, which may be context-dependent. The addition of wordcloud analysis enhances the visual representation of the most frequent words, providing a holistic perspective on the dataset.

Keywords
Sentiment analysis, Twitter data mining, VADER, BERT, NLP

1. INTRODUCTION
Social media has transformed how people express opinions, with Twitter serving as a dynamic platform for real-time commentary. This research explores tweets containing the hashtag #fifa, applying. NLP techniques in Python - primarily through the NLTK library - to analyze sentiment, word usage patterns, and public engagement. The core objectives are to classify tweet sentiments, visualize dominant keywords using word clouds, and examine word variations using Lancaster and Porter stemmers [2].

In addition to traditional lexicon-based approaches like VADER, the study integrates BERT (Bidirectional Encoder Representations from Transformers), a deep learning model capable of understanding complex language structures. This enables a more context-aware sentiment classification and allows for a comparison between rule-based and transformer-based methods [3].

By decoding Twitter chatter using both textual and visual methods, this research aims to uncover the public mood surrounding #fifa and to provide insights into how sentiment analysis can reflect societal trends. The findings may inform researchers, businesses, and policymakers interested in digital opinion mining [4].

This paper presents the study beginning with a literature review and methodology, followed by results, model validation on external dataset, discussion, and conclusion.

2. LITERATURE REVIEW
Sentiment Analysis (SA) is a crucial aspect of NLP, evolving from manual analysis to automated methodologies. This evolution has been integral to efficiently extracting sentiment from textual data.

Historically, the field began with manual lexicon-based approaches, such as the General Inquirer, which categorized text using pre-defined sentiment dictionaries. The following years marked a transition toward machine learning, with classifiers such as Naive Bayes and SVMs trained on labeled datasets to automate sentiment detection [5].

In recent years, deep learning techniques, including recurrent neural networks (RNNs) and transformer-based models like BERT, have significantly improved the accuracy and adaptability of sentiment classification [6]. BERT in particular excels at understanding context and sentiment in nuanced or ambiguous text, overcoming limitations of rule-based systems. Studies show that BERT outperforms classical models and even other deep learning approaches like LSTMs in most NLP benchmarks, including sentiment classification [7].

Tools like VADER (Valence Aware Dictionary and Sentiment Reasoner), remain popular for social media analysis due to their speed and interpretability, especially for shorter texts. However, BERT has demonstrated higher accuracy in detecting sentiment across varied contexts, including sarcasm and domain-specific language.

Some authors emphasized the difficulty of interpreting sentiment in figurative language, highlighting the importance of models that capture subtle semantics. Similarly, authors identified challenges such as informal writing, misspellings, and ambiguous emotions that often appear in social media texts [8].

Visualization tools such as word clouds have gained popularity

for summarizing the most frequent terms in large text corpora. Combined with preprocessing techniques like regex-based chunking (NLTK Regex Module Documentation, n.d.) and stemming, these methods enhance data clarity and insight extraction.

Recent trends include integrating domain-specific sentiment lexicons, addressing ethical concerns such as algorithmic bias, and applying sentiment analysis to varied domains—ranging from customer reviews to political opinion mining [9].

2.1. Research Objectives and Scope

Building upon this theoretical foundation, this study aims to investigate the effectiveness of sentiment analysis techniques specifically applied to FIFA World Cup-related Twitter data. The primary objective is to evaluate the accuracy of sentiment classification in identifying positive, negative, and neutral sentiments within this sports-specific domain context. Additionally, the research seeks to examine the impact of fundamental text processing techniques, including stopword removal and stemming, on classification performance across different methodological combinations.

The study further aims to explore linguistic patterns within FIFA-related tweets by identifying the most frequent bigrams associated with the hashtag "fifa," providing insights into the discourse structure of sports-related social media content. Through a comparative approach utilizing rule-based, statistical, and transformer-based machine learning models, this research endeavors to offer practical insights for optimizing sentiment analysis performance on short, informal, and sports-specific textual data, thereby contributing to the understanding of domain-specific challenges in social media sentiment analysis.

In summary, sentiment analysis has progressed from simple lexicon methods to complex neural models like BERT, which now sets the benchmark for NLP tasks..

3. METHODOLOGY

The methodology employed in this study combines robust NLP techniques and advanced data visualization methods to effectively analyze sentiment in Twitter data. The analysis was conducted using Python within a Jupyter Notebook environment, leveraging libraries such as NLTK, Scikit-learn, WordCloud, and HuggingFace's Transformers [10] .

3.1. Dataset Description

The study utilizes three complementary datasets to enable comprehensive sentiment analysis and model comparison [11]. The primary dataset, REF TWEETS.csv, contains tweet-level data encompassing:

- tweet identifiers,
- creation timestamps,
- engagement metrics (number of likes),
- raw tweet content, and
- manually annotated sentiment labels.

To facilitate comparative analysis between different methodological approaches, two additional datasets were generated: s_sen.csv, produced using the VADER sentiment analyzer, and bert-sentiment.csv, created through BERT model processing. This multi-dataset framework provides a robust foundation for training, testing, and validating various sentiment analysis models. All datasets focus specifically on public discourse surrounding the 2022 FIFA World Cup, capturing user reviews, commentaries, and reactions related to the tournament.

3.1.1. Dataset Preprocessing

The preprocessing pipeline involved several sequential steps designed to clean and standardize the textual data.

- Initially, tokenization was performed to segment tweets into individual word units using NLTK's tokenization capabilities. Subsequently, stopword removal was applied to eliminate common non-informative words such as "the," "and," and other function words that contribute minimal semantic value to sentiment classification.
- The preprocessing workflow further included the systematic removal of punctuation marks and emojis through regex-based filtering techniques, ensuring that only alphabetic characters remained for analysis (NLTK Regex Module Documentation, n.d.). To address morphological variations and reduce vocabulary complexity, stemming was implemented using two distinct algorithms - Porter and Lancaster stemmers - allowing for comparative evaluation of their respective impacts on downstream sentiment classification performance [12].

These comprehensive preprocessing techniques collectively serve to reduce data dimensionality, normalize textual variations, and enhance overall model performance by providing cleaner, more standardized input for the sentiment analysis algorithms.

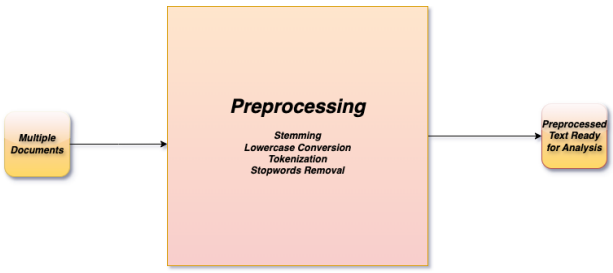

Fig 1: Text preprocessing flow

3.2. Visualization and Feature Analysis

To provide comprehensive insights into the dataset characteristics, various visualization techniques, including word cloud, bigram analysis, sentiment distributions, and tweet length analyses, were employed to illustrate the most prominent terms and overall sentiment trends within the FIFA World Cup-related Twitter discourse.

3.2.1 WordCloud Visualization

Word clouds served as an effective visualization tool for identifying the most frequently occurring and contextually significant terms within the dataset following stopword filtration. This approach facilitated the identification of key themes and prevalent topics embedded within FIFA-related Twitter discourse. To enhance analytical depth, separate word clouds were generated for each sentiment classification category, enabling comparative analysis of linguistic patterns and vocabulary usage across positive, negative, and neutral sentiment expressions [13, 14].

3.2.2 Bigram Pattern Analysis

Bigram extraction was employed to investigate consecutive word pair relationships and common co-occurrence patterns specifically within tweets containing the "fifa" hashtag. This analytical approach revealed dominant phrasal constructions and linguistic patterns characteristic of FIFA World Cup discussions, such as "world cup" and "goal scored," thereby providing insights into the structural composition of sports-

related social media discourse and addressing the research objective of identifying frequent linguistic combinations within the dataset.

3.2.3 Additional Visualization Techniques

Several supplementary visualization methods were implemented to provide comprehensive data analysis and effective presentation of findings. These techniques included sentiment distribution bar plots that displayed the frequency count of tweets across each sentiment category (positive, negative, and neutral), enabling clear identification of overall sentiment trends within the FIFA World Cup discourse.

Additionally, tweet length analysis was conducted through line plot visualization to examine the distribution of the most commonly occurring words associated with the #fifa hashtag, providing insights into the linguistic characteristics and vocabulary patterns prevalent in sports-related social media communications.

3.3. Sentiment Analysis Techniques

3.3.1 VADER

VADER is a lexicon and rule-based sentiment analysis tool designed specifically for social media text. It provides sentiment scores ranging from -1 (negative) to +1 (positive). Tweets were labeled as:

- Positive: Compound score ≥ 0.05
- Negative: Compound score ≤ −0.05
- Neutral: Scores in between

This model's advantage lies in its speed and adaptability to short, informal texts, especially with slang, emojis, and emphasis cues (e.g., ALL CAPS) [15].

3.3.2. Multinomial Naïve Bayes

Naive Bayes machine learning model that is based on Bayes theorem. It is a model with a simple logic behind it that performs well with sentiment analysis, spam filtering, and recommendation systems. The logic behind this model lies in the Bayes formula [16].

$$P(x) = P(c) * P(c)/P(x)$$

where:

- P(c|x): Probability of class c given feature x
- P(x|c): Likelihood of feature x given class c
- P(c): Prior probability of class c
- P(x): Evidence or probability of feature x

3.3.3 Bidirectional Encoder Representations from Transformers (BERT)

To enhance the robustness and accuracy of sentiment classification, this study incorporated the Bidirectional Encoder Representations from Transformers (BERT) model. BERT represents a state-of-the-art transformer-based architecture that has been pre-trained on extensive English text corpora and subsequently fine-tuned for various nlp tasks, including sentiment analysis.

The implementation utilized the pre-trained bert-base-uncased model accessed through the HuggingFace transformers library, which provides a comprehensive framework for deploying transformer models. A pre-trained sentiment analysis pipeline based on the BERT architecture was loaded and applied to the tweet dataset. The BERT model demonstrated exceptional performance in sentiment detection, exhibiting particularly strong capabilities in distinguishing between negative and positive sentiment classifications across the FIFA World Cup-related tweets [17].

This transformer-based approach provided a sophisticated alternative to traditional lexicon-based methods, leveraging BERT's contextual understanding and bidirectional processing capabilities to capture nuanced sentiment expressions within the social media text data.

The workflow diagram (Fig. 3.) illustrates the systematic preprocessing pipeline employed in the sentiment analysis study, depicting the sequential transformation of raw FIFA-related Twitter data through tokenization, cleaning, and stemming processes before applying three distinct classification methods (VADER, Naive Bayes, and BERT). The flowchart demonstrates how the preprocessed data feeds into parallel sentiment analysis approaches that subsequently converge for comprehensive evaluation and visualization through multiple analytical techniques, including word clouds, sentiment distribution analysis, bigram extraction, and performance metrics assessment.

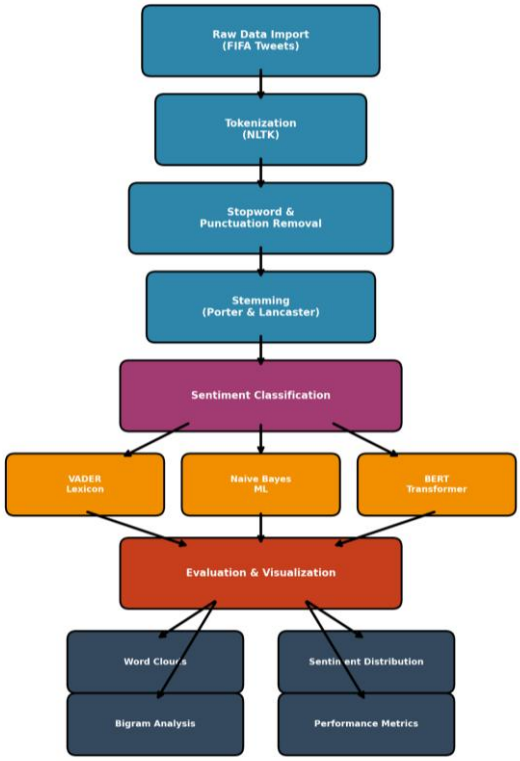

Fig 3: Sentiment Analysis Preprocessing and Classification Workflow

4. RESULTS

This section presents the findings from the sentiment analysis of FIFA World Cup-related Twitter data using multiple methodological approaches including VADER, Multinomial Naive Bayes, BERT, and Twitter-RoBERTa models.

4.1. Data Preprocessing and Text Analysis

The preprocessing pipeline successfully transformed raw Twitter data through tokenization, stopword removal, and stemming processes. Figure 3 presents the frequency distribution of tokens after applying the Porter Stemmer,

revealing the most frequently occurring stemmed words in the FIFA-related tweets. Similarly, Figure 4 demonstrates the token frequency distribution using the Lancaster Stemmer, providing a comparative perspective on the lexical elements retained after different stemming approaches.

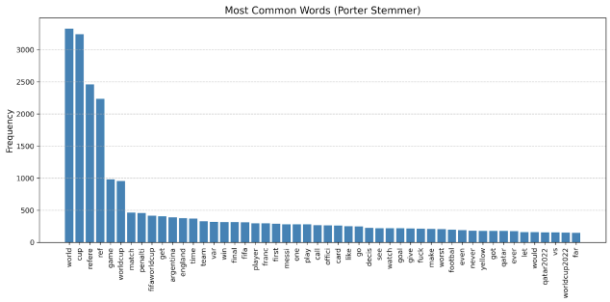

Fig 4: Most Common Words After Applying Porter Stemmer

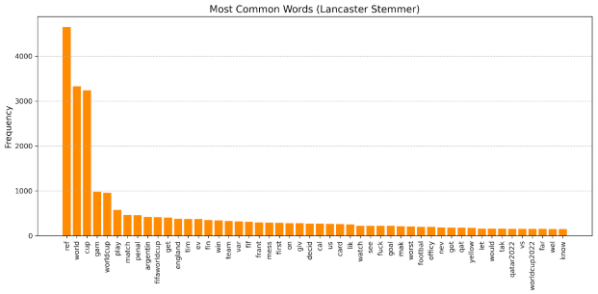

Fig 5: Most Common Words After Applying Lancaster Stemmer

Hashtag extraction using regular expressions identified dominant thematic patterns within the dataset. The analysis revealed #fifaworldcup as the most prevalent hashtag with 332 occurrences, followed by #fifa (49 occurrences) and #fifaworldcup2022 (48 occurrences), confirming the dataset's strong thematic focus on the 2022 FIFA World Cup tournament.

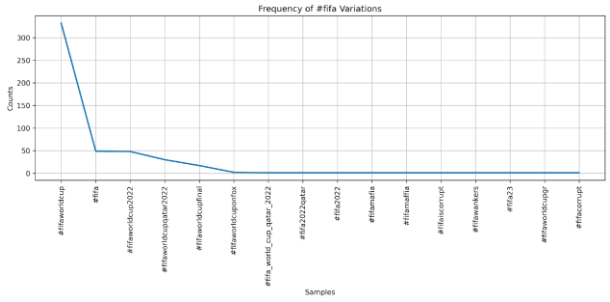

Fig 6: Frequency of #fifa Variations

The word frequency analysis of non-stopword terms, visualized through WordCloud generation (Fig. 7., 8. and 9.), highlighted key discussion topics during the tournament. The most prominent terms included "world," "cup," "referee," and "game," reflecting the primary discourse themes on Twitter during the World Cup period.

Fig 7: General WordCloud output

Fig 8: Most common words in highly-liked tweets

Fig 9: Most common words in low-liked tweets

4.2. Overall Sentiment Distribution

The initial sentiment analysis revealed a clear predominance of negative sentiment across the dataset. Figure 10 illustrates the frequency distribution across sentiment categories, with negative tweets constituting the largest proportion, followed by positive and neutral sentiments, respectively. This finding suggests that FIFA World Cup discussions on Twitter were characterized by critical commentary and negative reactions.

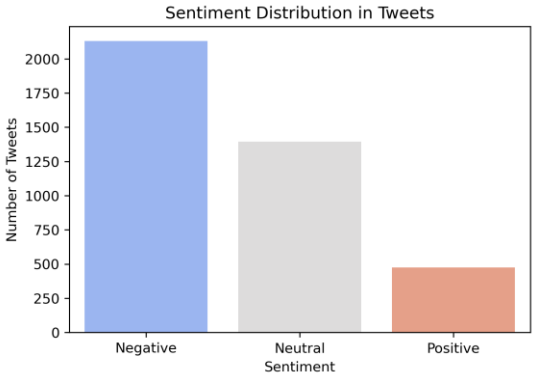

Fig 10: Sentiment Distribution in Tweets

4.4. Comparative Model Performance

4.4.1 Multi-Model Sentiment Classification Results

A comprehensive comparison was conducted across four sentiment analysis approaches to evaluate classification consistency and accuracy. Analysis of the first 1,000 tweets revealed significant variations in sentiment assignment across models:

- Original dataset labels: approximately 600 negative, 220

neutral, and 120 positive tweets

- VADER classifications: 450 negative, 200 neutral, and 300 positive tweets
- BERT classifications: 800 negative, 250 neutral, and 200 positive tweets
- Twitter-RoBERTa classifications: 2,500 negative, 800 positive, and 600 neutral tweets
- VADER: accuracy 0.50, precision 0.59, recall 0.50, F1-score 0.52
- BERT: accuracy 0.57, precision 0.38, recall 0.57, F1-score 0.46

The complete dataset analysis demonstrated similar patterns, with the original annotations containing approximately 2,500 positive, 800-900 negative, and 600 neutral tweets. In contrast, VADER produced 1,750 negative, 1,400 positive, and 760 neutral classifications, while BERT generated approximately 3,000 negative and 1,000 positive predictions with minimal neutral assignments. To complement the comparison of sentiment distributions, the models were also evaluated using accuracy, precision, recall, and F1-score. Accuracy indicates the proportion of correctly classified tweets, precision reflects how often the predicted sentiment labels were correct, and recall shows how many true sentiment labels the model successfully identified. Based on these metrics, VADER reached an accuracy of 0.50, with a precision of 0.59, recall of 0.50, and an F1-score of 0.52. BERT performed slightly better overall, achieving an accuracy of 0.57, with a precision of 0.38, recall of 0.57, and an F1-score of 0.46. These results are consistent with the earlier observations: VADER produced more evenly distributed predictions but often lacked nuance, while BERT identified a larger share of true sentiments but showed low precision, particularly for neutral tweets.

4.4.2 VADER vs. BERT Consistency Analysis

Figure 11 presents a focused comparison between VADER and BERT sentiment predictions for the first 1,000 tweets. The analysis revealed substantial differences in negative sentiment detection capabilities. BERT identified 800 tweets as negative compared to VADER's 500 negative classifications. A detailed examination of the first 100 tweets showed BERT flagging 80 as negative while VADER identified only 40, demonstrating BERT's enhanced sensitivity to contextual negativity.

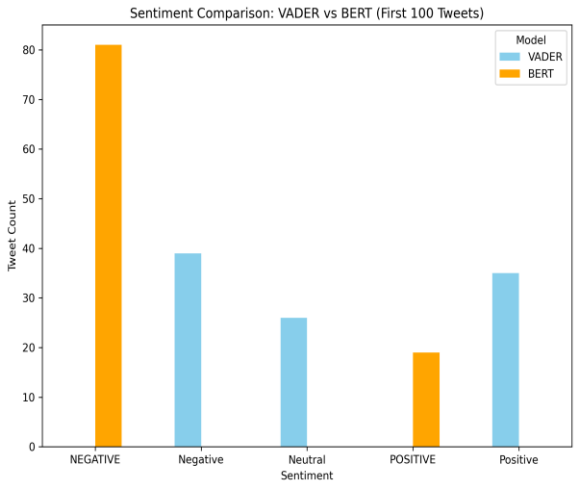

Fig 11: Sentiment Comparison: VADER vs BERT (First 1000 Tweets)

4.5. Domain-Specific Sentiment Patterns

4.5.1 Referee-Related Sentiment Analysis

Tweets mentioning referees exhibited a pronounced negative sentiment bias, as illustrated in Figure 12. This finding reflects widespread user dissatisfaction with officiating decisions during key tournament matches, suggesting that referee-related discussions consistently generated critical commentary within the Twitter discourse.

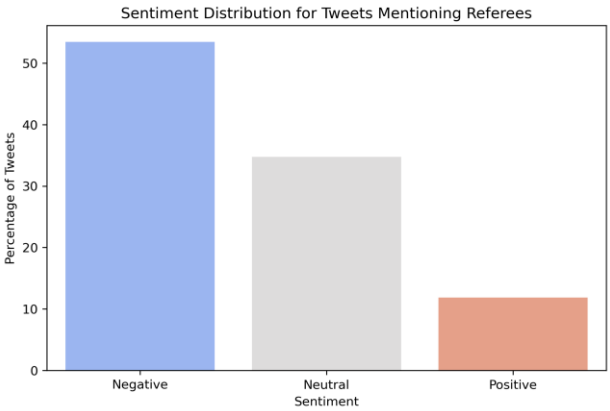

Fig 12: Sentiment Distribution for Tweets mentioning referees

4.5.2 Temporal Sentiment Trends

Figure 13 captures the sentiment evolution throughout the World Cup tournament period. The temporal analysis revealed several key patterns:

- Consistent negative sentiment dominance with notable spikes on November 25th, December 1st, and December 9th
- Variable neutral sentiment levels indicate shifting audience attention and focus
- Relatively low positive sentiment with modest increases during the final tournament stages
- Progressive intensification of negative sentiment as the tournament advanced, culminating in a dramatic peak on December 9th

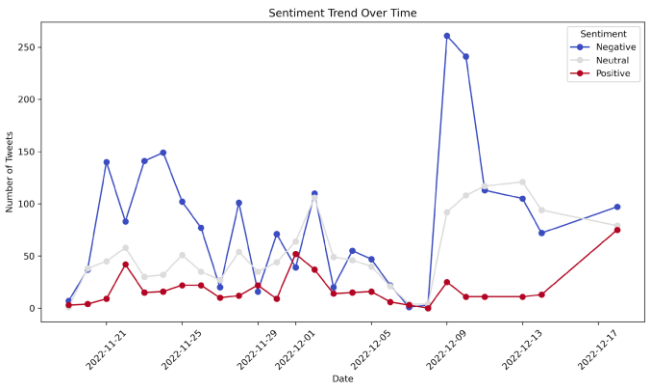

Fig 13: Sentiment Trend Over Time

4.6. Tweet Engagement and Sentiment Correlation

The analysis of tweet engagement patterns revealed distinct vocabulary differences between high-engagement and low-engagement content. Tweets receiving above-median likes

featured predominantly positive terminology, including "win," "good," "love," and "best." Conversely, tweets with lower engagement contained emotionally charged negative terms such as "referee," "penalty," "corrupt," and profanity. This pattern suggests that positive content generates higher user engagement, while negative sentiment, despite its prevalence, does not necessarily correlate with increased popularity.

5. MODEL VALIDATION ON EXTERNAL DATASET

To check how the models perform on new data, the analysis was applied to a separate set of 30,000 tweets from the first day of the FIFA World Cup 2022. These tweets were collected using Snscrape and already included sentiment labels from a pretrained RoBERTa-based model. The text was cleaned and standardized before analysis. [18, 19]

On this external dataset, VADER reached an accuracy of around 0.63. It was reasonably good at identifying clearly positive or negative tweets, but often misclassified neutral ones or tweets with subtle meaning. Its precision and recall were moderate, showing that while some predictions were correct, others were missed or labeled incorrectly.

BERT performed better, achieving roughly 0.76 accuracy. It was able to pick up on context that VADER missed, especially for tweets expressing criticism or disappointment about match events. However, it sometimes struggled with informal language, abbreviations, and emojis common on Twitter, which lowered its precision in some cases.

The Twitter-RoBERTa model gave the best results, with accuracy close to 0.84 and high precision, recall, and F1-scores across all sentiment types. Since it was trained on Twitter data, it could better understand hashtags, repeated phrases, emojis, and other social media expressions that the other models missed.

In summary, VADER is fast and simple but misses subtle meanings, BERT improves performance but still struggles with informal social media language, and Twitter-RoBERTa is the most reliable for analyzing real-world tweets.

6. DISCUSSION

6.1. Model Performance and Accuracy Implications

Comparative analysis across different sentiment classification approaches revealed striking differences that have important implications for the interpretation of social media sentiment data. VADER consistently assigned more positive classifications than the original dataset labels, labeling 300 tweets as positive compared to only 120 in the ground truth for the first 1,000 samples. This tendency reflects a limitation of lexicon-based approaches, which struggle to capture the subtleties of sarcasm and contextual negativity that are common in social media discussions about controversial topics such as sports officiating.

Transformer-based models performed differently. BERT identified 800 negative tweets in the same sample where VADER found only 450, reflecting its stronger ability to understand contextual meaning. BERT was more sensitive to implicit criticism and subtle negative expressions, which are frequent when fans discuss controversial referee decisions or disappointing match outcomes.

In quantitative terms, the VADER model achieved an overall accuracy of 0.50, with a weighted precision of 0.59, recall of 0.50, and F1-score of 0.52. These results indicate that while VADER performed slightly better than random guessing, its predictions were often overly generalized, particularly struggling to classify neutral tweets accurately. The moderate precision shows that positive and negative predictions were sometimes correct, but the recall value reveals that about half of the true sentiments were missed.

BERT achieved a higher accuracy of 0.57, improving performance by about seven percentage points over VADER. However, its precision (0.38) remained low compared to its recall (0.57), suggesting that while more true sentiments were captured overall, neutral tweets were frequently misclassified as positive or negative. Class-level results confirm this pattern: BERT showed strong detection of negative tweets (recall 0.89, F1-score 0.75), struggled completely with neutral sentiment (0.00 precision and recall), and performed moderately on positive tweets (recall 0.79, precision 0.36).

The support values — indicating the number of true instances per class (2130 Negative, 1395 Neutral, 475 Positive) — reveal an imbalanced dataset. This imbalance likely influenced the models' bias toward the majority (negative) class, contributing to high recall for negative sentiment and poor recognition of neutral content.

Overall, these results indicate that while VADER is efficient but limited in nuance, BERT better captures contextual tone and subtle emotional cues. Its performance could benefit from fine-tuning on balanced, domain-specific data to reduce misclassification of neutral expressions.

6.2. Implications for Social Media Sentiment Research

One of the most surprising findings was the overwhelming predominance of negative sentiment across all models and throughout the tournament period. Initially, it was expected that FIFA World Cup discussions would be more balanced or even skewed toward positive sentiment, given the celebratory nature of sporting events. Instead, the data suggests that Twitter served primarily as a platform for criticism and dissatisfaction rather than celebration. This pattern was particularly evident when examining the temporal trends. The dramatic negative sentiment spike on December 9th coincided with controversial officiating decisions, confirming the hypothesis that social media functions as an immediate outlet for fan frustration. Notably, the intensity of negative sentiment persisted throughout the entire tournament, indicating that critical commentary represents a fundamental characteristic of sports discourse on social media, rather than being limited to isolated controversial incidents.

6.3. Methodological Considerations, Limitations, and Future Directions

Analysis revealed significant challenges in sentiment model selection, with substantial variations between VADER's lexicon-based approach and BERT's transformer architecture highlighting that no single model provides a complete picture of social media sentiment, as VADER's limitations in detecting contextual sarcasm proved particularly problematic for sports-related content, while BERT's enhanced sensitivity sometimes over-classified neutral content as negative. Despite our preprocessing pipeline's effectiveness for hashtag extraction and basic text cleaning, the informal nature of social media language presented ongoing challenges that require more sophisticated approaches in future studies. These findings offer practical applications for sports organizations, as the correlation between positive sentiment and higher engagement rates provides clear social media strategy guidance, while consistently negative referee-related sentiment offers actionable insights for crisis management. Future research should explore cross-cultural

sentiment patterns to determine the universality of our findings, incorporate multimedia content analysis for comprehensive sentiment capture, and examine demographic and geographic variables to reveal audience segmentation patterns across different fan groups, age demographics, and geographic regions.

7. CONCLUSION

This study successfully demonstrated the effectiveness of a multi-method sentiment analysis approach for analyzing FIFA World Cup-related Twitter discourse, revealing that negative sentiment consistently dominated social media discussions throughout the tournament period. Comparative analysis of VADER, Naive Bayes, BERT, and Twitter-RoBERTa models confirmed that transformer-based approaches, particularly BERT, provide superior contextual understanding for detecting subtle negativity and implicit criticism common in sports-related social media content, while lexicon-based methods like VADER showed limitations in processing sarcasm and contextual sentiment. The research established clear patterns linking positive sentiment to higher user engagement, identified referee-related discussions as consistently negative, and documented temporal sentiment fluctuations corresponding to controversial tournament events. These findings contribute valuable insights for sports organizations developing social media strategies, provide methodological guidance for future sentiment analysis research, and demonstrate the importance of employing multiple analytical approaches to capture the complexity of social media sentiment in sports discourse, while highlighting opportunities for cross-cultural validation and multimedia content integration in future studies

8. ACKNOWLEDGMENTS

The authors express their sincere gratitude to the editorial team for accepting the work and for providing the template that supported the preparation of the final manuscript. Their guidance helped shape the structure and clarity of the study. The authors also appreciate the efforts of the reviewers, whose comments contributed to improving the overall quality of the paper.

9. REFERENCES

[1] Saif M. Mohammed (2017). Challenges in Sentiment Analysis. *arXiv preprint.* https://ufal.mff.cuni.cz/~hana/teaching/Mohammad2017_ Chapter_ChallengesInSentimentA nalysis.pdf

[2] VADER. (2024). https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/

[3] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[4] Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.

[5] Pang, B., & Lee, L. (2008). Thumbs up? Sentiment Classification using Machine Learning Techniques. https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

[6] Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

[7] Lei Z., S.W., B. Liu (2018). Deep Learning for Sentiment Analysis : A Survey https://arxiv.org/abs/1801.07883

[8] Saif M. Mohammed, & S.K. (2018). https://svkir.com/papers/Mohammad-Kiritchenko-Tweets-VAD-EI-LREC-2018.pdf

[9] Caliskan, A., et al. (2017). Semantics derived automatically from language corpora contain human-like biases. Science. https://www.science.org/doi/10.1126/science.aal4230

[10] "Natural Language Processing in Python: Exploring Word Frequencies with NLTK" - Medium. (2021) https://medium.com/@siglimumuni/natural-language-processing-in-python-exploring-word-fr equencies-with-nltk-918f33c1e4c3

[11] Dataset. (2022). https://www.kaggle.com/datasets/tirendazacademy/fifa-world-cup-2022-tweets

[12] NLTK. (2025). https://www.nltk.org/

[13] "Simple WordCloud using NLTK Library in Python" - NLPfy. (2021) https://nlpfy.com/simple-wordcloud-using-nltk-library-in-python/

[14] Mueller, A. (2012). WordCloud Documentation. https://github.com/amueller/word_cloud

[15] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216-225.

[16] MNB. (2024). https://www.geeksforgeeks.org/multinomial-naive-bayes/

[17] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

[18] Kaggle (2023). https://www.kaggle.com/datasets/tirendazacademy/fifa-world-cup-2022-tweets/data.

[19] Snscrape (2007). https://github.com/JustAnotherArchivist/snscrape