Sentiment Analysis of FIFA World Cup 2022 Twitter Reviews

Rijul Bilaiya
MIT World Peace University
Pune, India
1032201594@mitwpu.edu.in

Mrunal Dande
MIT World Peace University
Pune, India
1032202209@mitwpu.edu.in

Faisal
MIT World Peace University
Pune, India
1032201546@mitwpu.edu.in

Under the guidance of :
Suja Panicker
Assistant Professor
School of CET,
Dr Vishwanath Karad MIT WPU
Pune, India
Suja.panickar@mitwpu.edu.in

Abstract – Sentiment analysis is a natural language processing (NLP) technique used to determine the emotional tone behind a piece of text. It involves analyzing a text, such as a sentence or a paragraph, to understand whether the expressed sentiment is positive, negative, or neutral. Sentiment analysis, a pivotal domain in natural language processing, focuses on discerning emotional polarity within textual data. This computational technique involves parsing, tokenization, and employing diverse algorithms to categorize sentiments as positive, negative, or neutral. By assigning numerical scores or employing machine learning models, sentiment analysis aids in gauging public opinion, market trends, and customer feedback across various domains, enabling informed decision-making and tailored strategies for businesses and organizations.

***Key Words*:** Sentiment Analysis , Text Analysis , Parsing
Tokenization , Algorithms , Numerical Scores

1.INTRODUCTION

Sentiment analysis—also referred to as opinion mining—is an essential component of natural language processing (NLP), which is the study and interpretation of the attitudes, opinions, and feelings conveyed in textual data. It functions under the assumption that language carries sentiment by nature, which can be categorized as neutral, positive, or negative.

Sentiment analysis's main goal is to extract and assess the subjective information that is embedded in text, be it news stories, surveys, social media posts, or any other type of written communication. A fuller understanding of how people or groups feel about particular subjects, goods, services, occasions, or encounters is made possible by this approach.

Sentiment analysis is continually evolving with advancements in machine learning, deep learning, and the expansion of annotated datasets. Its applications span across social media monitoring, customer experience management, market research, brand reputation analysis, political analysis, and beyond. As the volume of textual data grows exponentially, sentiment analysis remains a pivotal tool in extracting valuable insights and understanding the collective sentiment of individuals or communities expressed through language.

2. LITERATURE REVIEW

A comprehensive literature review on sentiment analysis encompasses various facets of this field, incorporating historical developments, methodologies, applications, challenges, and recent advancements. Here's a breakdown of

what a literature review might cover:
[1]Historical Perspective: Start by exploring the historical evolution of sentiment analysis. Discuss seminal works, key milestones, and the evolution of methodologies from early rule-based systems to modern machine learning and deep learning approaches. Highlight how sentiment analysis has transitioned from basic polarity detection to more nuanced emotion analysis.

[2]Methodologies and Techniques: Dive into the methodologies and techniques used in sentiment analysis. Explore rule-based approaches utilizing lexicons, statistical methods, supervised learning algorithms (like Support Vector Machines, Naive Bayes), unsupervised learning (clustering, topic modeling), and more recent advancements such as deep learning architectures (like recurrent neural networks, convolutional neural networks, transformers) applied to sentiment analysis tasks.

[3]Datasets and Annotation: Discuss the role of datasets in sentiment analysis research. Highlight benchmark datasets (e.g., IMDb movie reviews, Twitter sentiment datasets), their characteristics, size, and annotation techniques. Elaborate on the importance of annotated datasets for training and evaluation of sentiment analysis models.

[4]Applications: Cover diverse applications of sentiment analysis across industries. Explore its role in social media monitoring, customer feedback analysis, market research, brand reputation management, political analysis, healthcare (patient sentiment analysis), and beyond. Discuss how sentiment analysis contributes to decision-making processes in these domains.

[5]Challenges and Limitations: Address the challenges and limitations faced in sentiment analysis. This might include handling sarcasm, irony, context-dependent sentiments, multilingual sentiment analysis, domain adaptation issues, and ethical considerations related to biases in data or models.

[6]Recent Advances and Future Directions: Highlight recent research trends and advancements in sentiment analysis. Discuss state-of-the-art models, novel techniques (such as transfer learning, multimodal sentiment analysis), and emerging areas like aspect-based sentiment analysis, emotion detection, and sentiment analysis in low-resource languages. Propose potential future directions and areas for further research. A literature review in sentiment analysis should synthesize and critically analyze existing research, providing a comprehensive understanding of the field's evolution, methodologies, applications, challenges, and avenues for future exploration and improvement.

3. RESEARCH METHODOLOGY

Research methodology in sentiment analysis involves a systematic approach to investigating, analyzing, and interpreting sentiments expressed within textual data. Here's a structured outline detailing the steps and components of a research methodology in sentiment analysis:Determine the research approach: whether it's exploratory, descriptive, experimental, or a combination of methodologies.Identify and gather relevant datasets suitable for the research objectives. Define criteria for data selection, considering factors like data source credibility, diversity, size, and annotation quality.Ensure ethical considerations, particularly when dealing with sensitive or personally identifiable information.

[1]Preprocessing and Data Preparation: Perform text preprocessing steps such as tokenization, stemming, lemmatization, removing stopwords, and handling special characters or punctuation.

Normalize the text data by converting it to a consistent format, addressing spelling variations, and encoding special characters or emojis.Conduct data cleaning to eliminate noise or irrelevant information that might affect sentiment analysis accuracy.

[2]Feature Extraction and Representation: Extract features or representations from the preprocessed text data suitable for sentiment analysis.

Utilize various techniques like Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), word embeddings (Word2Vec, GloVe), or contextual embeddings (BERT, GPT) to represent textual features.

[3]Sentiment Analysis Techniques: Select appropriate sentiment analysis methodologies based on the research objectives and the nature of the dataset. Experiment with rule-based approaches, lexicon-based sentiment analysis, machine learning algorithms (supervised or unsupervised), deep learning models, or hybrid methodologies to classify sentiments.

[4]Model Training and Evaluation: Train sentiment analysis models using appropriate algorithms or neural network architectures. Employ cross-validation or train-test splits to evaluate model performance, considering metrics like accuracy, precision, recall, F1 score, or area under the ROC curve (AUC). Fine-tune models and hyperparameters to optimize performance and generalize well to unseen data.

[5]Analysis and Interpretation: Analyze the sentiment analysis results, examining the distribution of sentiments, identifying patterns, and understanding the implications of the findings.

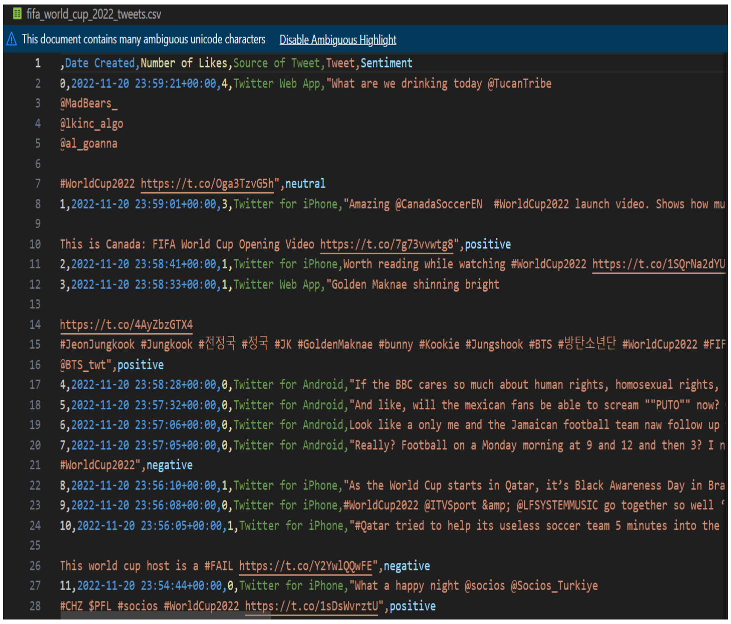
Interpret the insights obtained from sentiment analysis in the context of the research objectives, drawing conclusions and making recommendations based on the analysis.

Validation and Discussion: Validate the reliability and validity of the sentiment analysis results by discussing the limitations, potential biases, and uncertainties associated with the methodology and findings. Discuss the broader implications of the research outcomes, addressing how the insights contribute to existing knowledge or practical applications.

4. MODELLING APPROACH

A. Data Collection:

The dataset being used here consists of the FIFA World cup 2022.



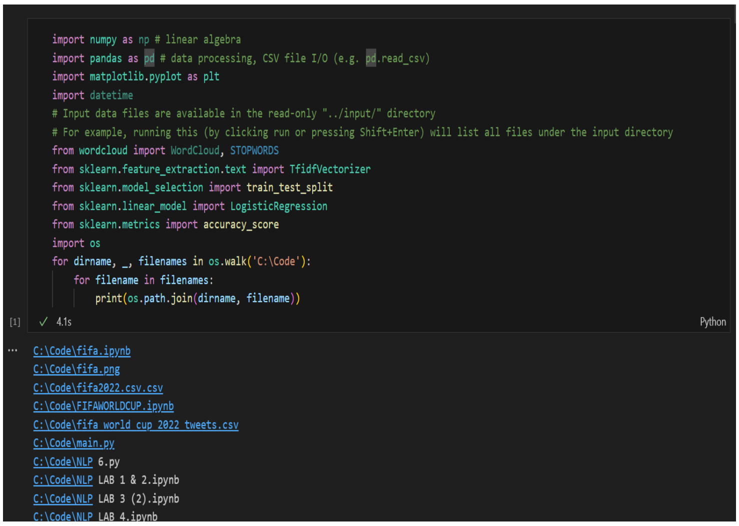
fifa_world_cup_2022_tweets.csv

This document contains many ambiguous unicode characters

```
,Date Created,Number of Likes,Source of Tweet,Tweet,Sentiment
0,2022-11-20 23:59:21+00:00,4,Twitter Web App,"What are we drinking today @TucanTribe
@MadBears_
@lkinc_algo
@al_goanna

#WorldCup2022 https://t.co/Oga3TzvG5h",neutral
1,2022-11-20 23:59:01+00:00,3,Twitter for iPhone,"Amazing @CanadaSoccerEN  #WorldCup2022 launch video. Shows how mu

This is Canada: FIFA World Cup Opening Video https://t.co/7g73vvwtg8",positive
2,2022-11-20 23:58:41+00:00,1,Twitter for iPhone,Worth reading while watching #WorldCup2022 https://t.co/1SQrNa2dYU
3,2022-11-20 23:58:33+00:00,1,Twitter Web App,"Golden Maknae shinning bright

https://t.co/4AyZbzGTX4
#JeonJungkook #Jungkook #전정국 #정국 #JK #GoldenMaknae #bunny #Kookie #Jungshook #BTS #방탄소년단 #WorldCup2022 #FIF
@BTS_twt",positive
4,2022-11-20 23:58:28+00:00,0,Twitter for Android,"If the BBC cares so much about human rights, homosexual rights,
5,2022-11-20 23:57:32+00:00,0,Twitter for Android,"And like, will the mexican fans be able to scream ""PUTO"" now?
6,2022-11-20 23:57:06+00:00,0,Twitter for Android,Look like a only me and the Jamaican football team naw follow up
7,2022-11-20 23:57:05+00:00,0,Twitter for Android,"Really? Football on a Monday morning at 9 and 12 and then 3? I n
#WorldCup2022",negative
8,2022-11-20 23:56:10+00:00,1,Twitter for iPhone,"As the World Cup starts in Qatar, it’s Black Awareness Day in Bra
9,2022-11-20 23:56:08+00:00,0,Twitter for iPhone,#WorldCup2022 @ITVSport &amp; @LFSYSTEMMUSIC go together so well
10,2022-11-20 23:56:05+00:00,1,Twitter for iPhone,"#Qatar tried to help its useless soccer team 5 minutes into the

This world cup host is a #FAIL https://t.co/Y2YwlQQwFE",negative
11,2022-11-20 23:54:44+00:00,0,Twitter for iPhone,"What a happy night @socios @Socios_Turkiye
#CHZ $PFL #socios #WorldCup2022 https://t.co/1sDsWvrztU",positive
```

B. Data Preparation and Preprocessing:

The first step in preparing and preprocessing the FIFA World Cup 2022 dataset was to download it from Kaggle in CSV format. The dataset contains 23000appro. rows and 6 columns. After that, we converted the CSV file to a JSON format using Python's Pandas library to make it easier to work with the data.



```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import datetime
# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory
from wordcloud import WordCloud, STOPWORDS
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import os
for dirname, _, filenames in os.walk('C:\Code'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
C:\Code\fifa.ipynb
C:\Code\fifa.png
C:\Code\fifa2022.csv.csv
C:\Code\FIFAWORLDCUP.ipynb
C:\Code\fifa world cup 2022 tweets.csv
C:\Code\main.py
C:\Code\NLP 6.py
C:\Code\NLP LAB 1 & 2.ipynb
C:\Code\NLP LAB 3 (2).ipynb
C:\Code\NLP LAB 4.ipynb
```

To further improve the dataset's quality, we manually reviewed and corrected any spelling errors and grammatical mistakes that were present. Finally, we split the preprocessed data into training and testing sets to evaluate the model's accuracy during the training process.

C. Algorithm

1. Lexicon-based Methods: One popular method is VADER (Valence Aware Dictionary and sEntiment Reasoner). This approach relies on a predefined dictionary containing words with associated sentiment scores. It assesses the overall sentiment by combining the scores of individual words in the text.

2. Machine Learning Algorithms: Algorithms like Naive Bayes, Support Vector Machines (SVM), and Random Forests or Decision Trees are frequently used. Naive Bayes relies on probabilities and word features to classify sentiments. SVMs create boundaries between different sentiment classes, while Random Forests use multiple decision trees to make predictions.

3. Neural Network Approaches: Neural networks like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models, as well as Convolutional Neural Networks (CNNs), are employed. RNNs and LSTMs analyze word sequences to capture context, while CNNs detect patterns in text sequences to identify sentiments.

4. **Hybrid Approaches**: Combining multiple methods, such as ensemble methods or rule-based systems, can improve sentiment classification accuracy. Ensemble methods merge predictions from different models, while rule-based systems use predefined rules or patterns in text to classify sentiments.

The choice of method often depends on factors like the dataset's size, the nature of the Twitter data being analyzed, available computational resources, and the specific goals of the sentiment analysis task. Experimentation and careful selection of the right method are crucial to effectively determine sentiments expressed in Twitter data.

D. Working

When exploring sentiments in tweets, there are various methods used to understand and classify these emotions. One common approach involves lexicon-based methods like VADER (Valence Aware Dictionary and sEntiment Reasoner), which relies on a pre-existing dictionary of words with associated sentiment scores. It looks at how these scores combine across words in a tweet to gauge the overall sentiment.

Machine learning techniques, such as Naive Bayes, Support Vector Machines (SVM), and Random Forests/Decision Trees, also play a significant role. Naive Bayes works with probabilities and word features to classify sentiments, while SVMs create boundaries between different emotional categories. On the other hand, Random Forests construct predictive models from multiple decision trees to infer sentiment.

Neural network approaches, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models, and Convolutional Neural Networks (CNNs), delve deep into the sequences of words to capture context. RNNs and LSTMs focus on understanding word relationships, while CNNs detect patterns within text sequences to identify sentiments.

Hybrid methods, combining various techniques like ensemble methods or rule-based systems, strive to improve accuracy. Ensemble methods merge predictions from diverse models, while rule-based systems rely on predefined rules or patterns in text for sentiment analysis.

The choice of method depends on factors like the dataset's size, the nature of the Twitter data being studied, available resources, and the specific goals of the sentiment analysis task. Careful consideration and experimentation with these methods are crucial in effectively interpreting sentiments expressed in Twitter data.

5. RESULT

After code completion, The presentation and format of sentiment analysis results may vary based on the specific tools, techniques, or platforms used for analysis. Whether it's a dashboard, a report, or an interactive interface, the sentiment analysis results aim to convey the emotional nuances present in the textual data, aiding decision-making processes in various domains such as marketing, customer feedback analysis, social media monitoring, and more.

6. CONCLUSION AND FUTURE SCOPE

In the conclusion of a study on sentiment analysis, it's essential to summarize the key findings, implications, and limitations while considering the future scope and potential advancements in the field. Here's an outline of the conclusion and future scope of sentiment analysis:

Conclusion:

1. Summary of Findings: Recapitulate the main findings of the sentiment analysis study, emphasizing the identified patterns, sentiment distributions, or significant insights obtained.

2. Implications and Applications: Discuss the practical implications of the findings in various domains (e.g., marketing, customer service, public opinion analysis).

Highlight how the insights derived from sentiment analysis can inform decision-making processes or strategies in relevant industries.

3.Contributions toKnowledge:Outline the contributions of the study to the existing body of knowledge in sentiment analysis.

Discuss how the research fills gaps or extends understanding in specific aspects of sentiment analysis methodologies or applications.

4. Limitations and Challenges: Acknowledge the limitations or constraints faced during the study, such as data biases, annotation challenges, or model limitations.

Discuss how these limitations might have impacted the results or interpretations.

Future Scope:

1. Advanced Methodologies:Propose advancements in sentiment analysis methodologies, such as exploring hybrid models that combine rule-based approaches with deep learning techniques. Discuss the potential of leveraging advanced NLP models like transformers for more accurate sentiment analysis.

2. Contextual Understanding: Emphasize the importance of context-aware sentiment analysis to better capture nuances, sarcasm, and cultural differences in text. Explore methods to incorporate contextual information into sentiment analysis models.

3. Multimodal Sentiment Analysis: Investigate the integration of multiple data modalities (text, images, audio) for a more comprehensive sentiment analysis approach. Explore the challenges and opportunities in analyzing sentiments from diverse data sources.

4. Ethical Considerations: Highlight the need for ethical considerations in sentiment analysis, addressing issues of bias, fairness, and privacy in data and models.

Discuss approaches to mitigate biases and ensure ethical practices in sentiment analysis research.

5. Domain-Specific Analysis: Focus on domain-specific sentiment analysis tailored to industries like healthcare, finance, or politics. Explore specialized sentiment analysis techniques catering to unique requirements in these domains.

6. Real-Time Analysis and Dynamic Adaptation: Discuss the potential for real-time sentiment analysis systems and adaptive models that continuously learn from new data streams.

7. REFERENCES

[1] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval.

[2] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies.

[3] Socher, R. et al. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Conference on Empirical Methods in Natural Language Processing (EMNLP).

[4] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Conference on Empirical Methods in Natural Language Processing (EMNLP).

[5] Hutto, C.J. & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International AAAI Conference on Weblogs and Social Media (ICWSM).

[6] Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-Based LSTM for Aspect-Level Sentiment Classification. Conference on Empirical Methods in Natural Language Processing (EMNLP)

[7] Severyn, A., & Moschitti, A. (2015). Twitter Sentiment Analysis with Deep Convolutional Neural Networks. Conference on Empirical Methods in Natural Language Processing (EMNLP).

[8]