Sentiment Analysis of Cricket Tweets: A Comparative Study of Machine Learning and Deep Learning Approaches

Problem Statement

The rapid proliferation of social media platforms, particularly Twitter, has resulted in the generation of vast amounts of textual data on a daily basis. Within this expansive sea of information lies valuable insights into public sentiment, opinions, and trends. In the realm of cricket, a sport of immense popularity worldwide, understanding the sentiments expressed by fans, analysts, and stakeholders can provide a competitive edge in various domains, including sports marketing, fan engagement, and predictive analytics.

However, the challenge arises from the sheer volume and diversity of cricket-related tweets, which encompass a wide range of emotions, opinions, and sentiments. Extracting meaningful insights from this unstructured textual data is a complex task, requiring advanced sentiment analysis techniques and machine learning models. *The problem at hand is to develop an accurate and robust sentiment analysis system specifically tailored to cricket-related tweets, capable of categorizing tweets into positive, negative, or neutral sentiments with a high degree of precision.* Furthermore, **the research aims to compare the effectiveness of various machine learning and deep learning algorithms in this context, ultimately providing valuable insights into public sentiment surrounding cricket events and topics.**

In summary, the problem statement revolves around the development of an efficient sentiment analysis model for cricket-related tweets, addressing the challenges posed by the diverse and dynamic nature of social media data.
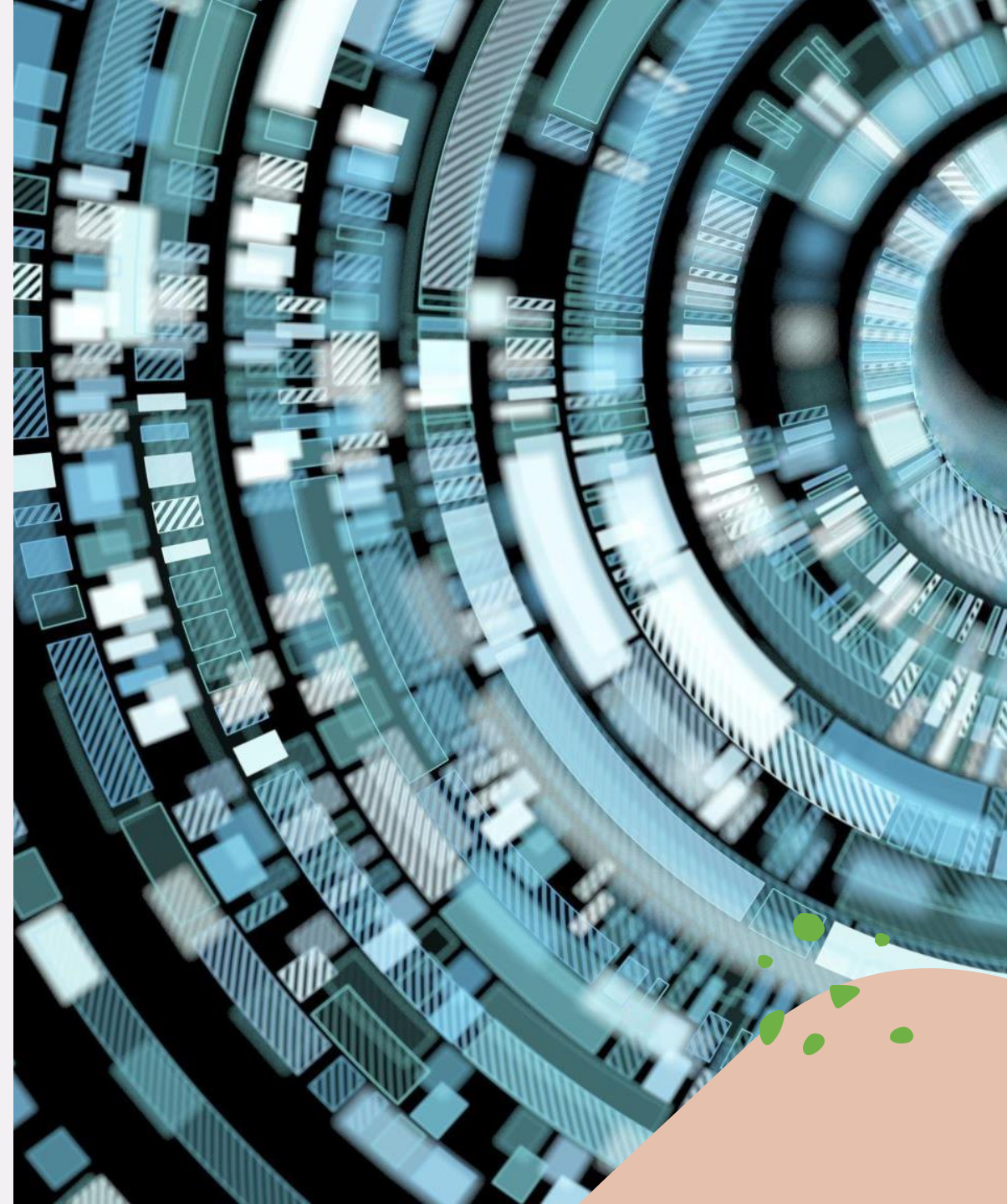
Abstract—The advent of social media platforms has ushered in an era where vast amounts of data are generated daily

Introduction

Twitter, a well-established web-based social networking platform, is renowned for its role in networking and microblogging, enabling users to send and receive messages known as tweets

As of 2021, Twitter boasts an impressive user base of 199 million monetizable daily active users, with India contributing significantly, accounting for 17.5 millions of these users

Sentiment analysis, an integral facet of text mining, holds significant relevance in text categorization

Twitter serves as an intermediary for individuals to exchange their perspectives, thereby serving as a valuable source of information for the public

Introduction

Twitter tweets play an indispensable role in providing a comprehensive overview of public opinion regarding services or products in the market

Sentiment analysis, commonly referred to as opinion mining, is the process of systematically analyzing product reviews and online content to gauge the prevailing sentiment or attitude towards a particular service or product

Sentiment analysis operates as a powerful tool for understanding public opinion on social media platforms, effectively categorizing text into positive, negative, or neutral sentiments

Review of Previous Work

In recent decades, there has been a significant surge in research and development related to sentiment analysis

This surge has led to the exploration of various applications for sentiment analysis

One particularly challenging task within this domain is the analysis of sentiment in textual data, such as messages and tweets

The intricacies of sentiment analysis in such text-based formats make it a demanding and complex endeavor

Their research aimed to unravel the dynamics of Twitter discussions leading up to a movie's release and their correlation with box office earnings, both during the opening weekend and subsequent weekends

Review of Previous Work

The proposed system leveraged innovative machine learning techniques, including Naïve Bayes , Support Vector Machine , and Maximum Entropy

The study involved the extraction of linguistic features for the purpose of predicting three key aspects: the number of fan followers, the quantity of tweets, and score predictions derived from classifying tweets into positive, negative, and neutral categories
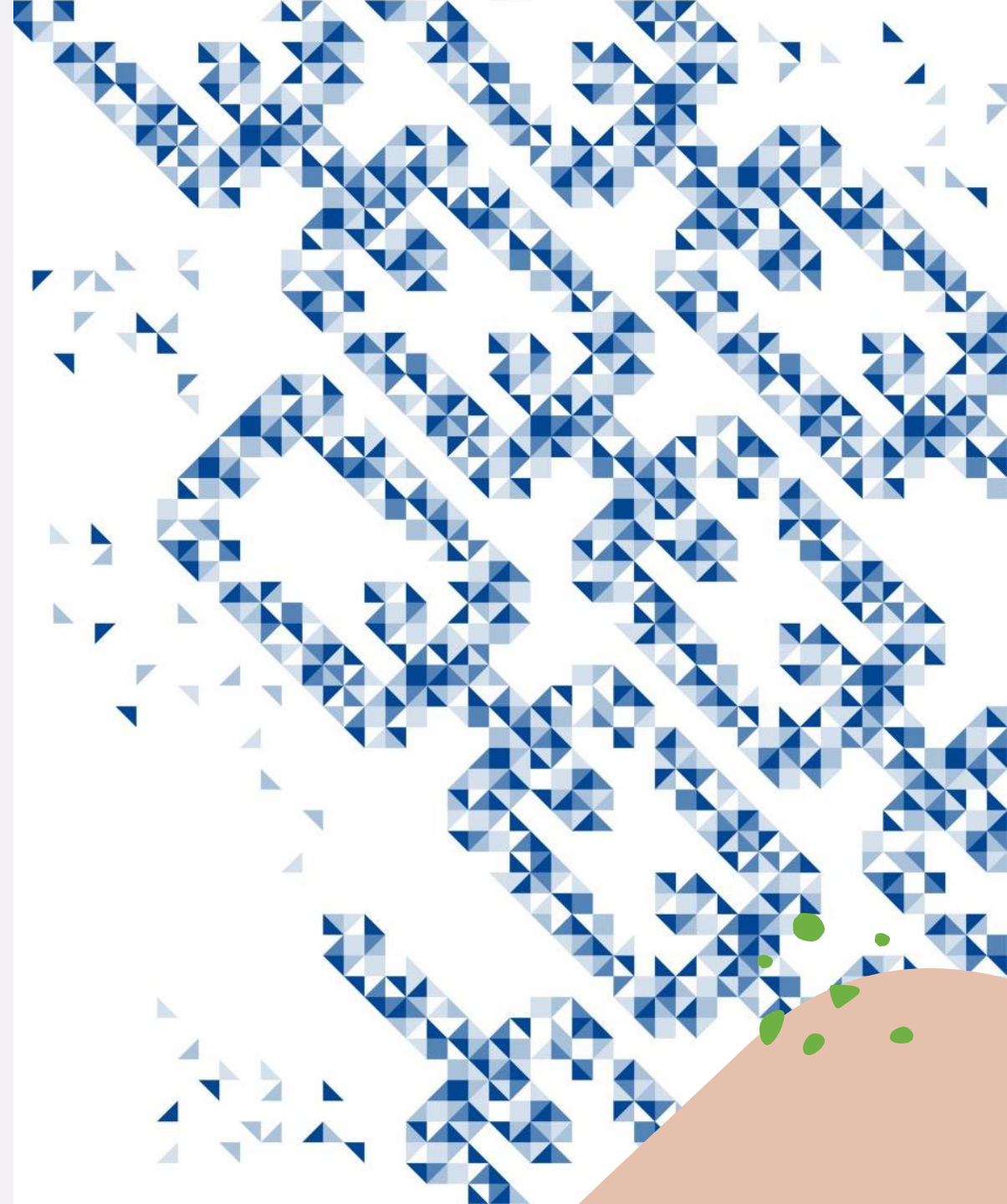
Data set description

This dataset is used from Kaggle named as "Cricket Tweets" given by Gabriel Preda, Data Scientist at Endava, Bucharest, Bucharest, Romania

Dataset is collected by using the Twitter API which is publicly available

Main attributes which I have used for model such as 'id', 'user_name', 'user_location', 'user_description', 'user_followers', 'date', 'text', 'hashtags', 'source', 'retweets', 'favorites'

Methodology

To commence our analysis, we will categorize tweets into distinct groups based on their content

To accomplish this task, we will employ a diverse range of algorithms, including Support Vector Machine , Logistic Regression, Random Forest, and Deep learning

The tweets, initially represented as text strings, will undergo transformation into a numerical format through the utilization of the Term Frequency-Inverse Document Frequency technique
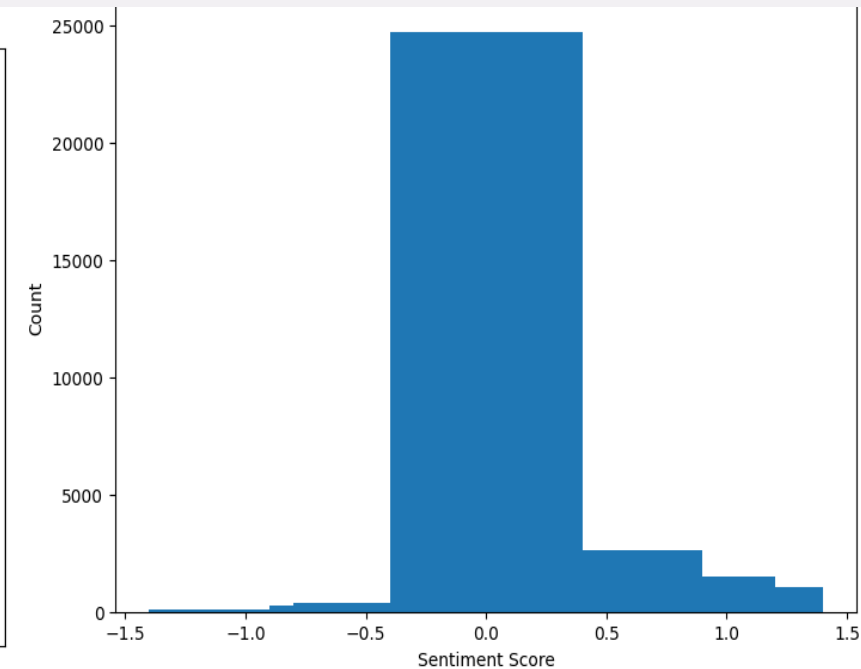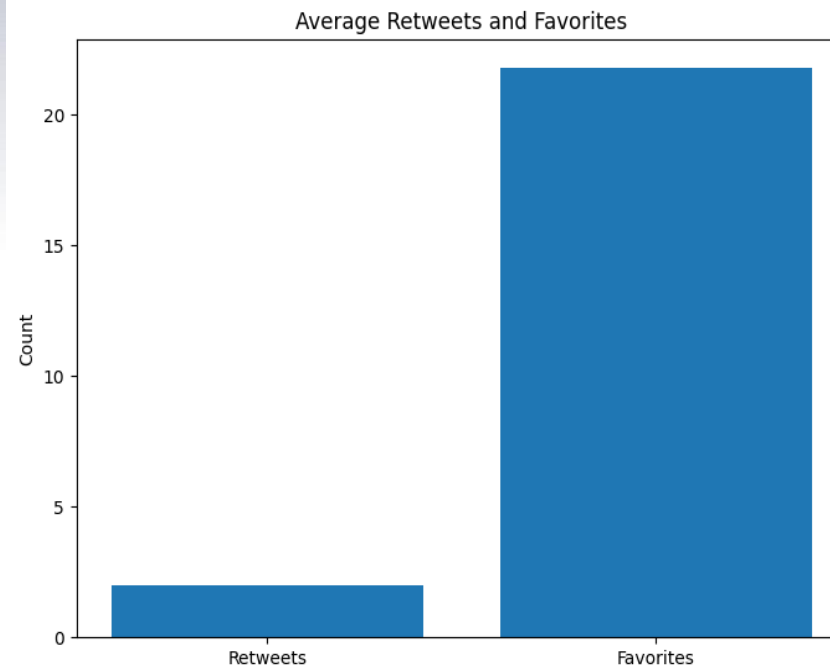
Logistic Regression

Model Building: For constructing our sentiment classifier, we employed logistic regression

Support Vector Machine

Support Vector Machine is a powerful machine learning algorithm with the primary objective of predicting a hyperplane within an N-dimensional space, where N represents the number of features in the dataset

SVM is rooted in computational learning theory, employing methodologies aimed at addressing complex structural problems

At its core, the concept of a support vector machine revolves around the discovery of a decision boundary that maximizes the separation between two distinct classes within the data
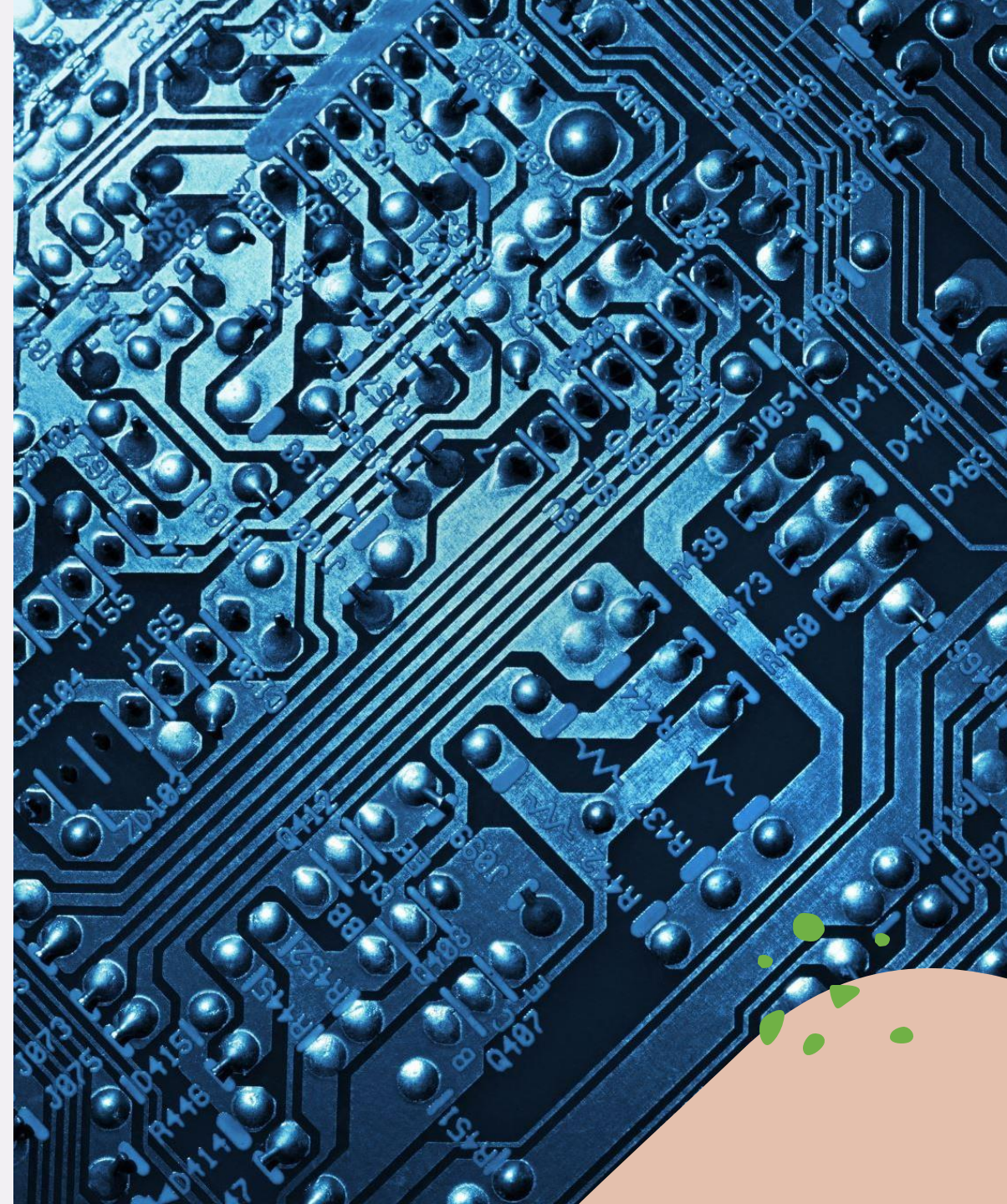
Random Forest Classifier

The random forest classifier is used for solving classification and regression problems

The versatility of the Random Forest algorithm extends its applicability into various facets of our daily lives

In the context of our Twitter sentiment analysis research, the Random Forest algorithm proves to be an indispensable tool, offering the potential to enhance the accuracy and reliability of sentiment classification in the dynamic and nuanced landscape of social media

E, Feature Extraction

Sentiment Scores: Calculating sentiment scores, which could be polarity scores or emotion-specific scores using sentiment analysis tools like TextBlob

Relevance for Decathlon

The problem statement of developing an accurate sentiment analysis model for cricket-related tweets can be related to Decathlon's mission, values, and business goals in the following ways:

1. Market Research and Customer Engagement: Decathlon, as a global sports retailer, is committed to understanding its customers' needs and preferences in the sports and fitness domain. By analyzing sentiments expressed on social media, especially related to popular sports like cricket, Decathlon can gain valuable insights into customer sentiment, preferences, and trends. This information can inform product development, marketing strategies, and inventory management to better cater to customer demands.

2. Customer-Centric Approach: Decathlon places a strong emphasis on a customer-centric approach. By monitoring sentiments on social media, including positive and negative feedback related to sports equipment and apparel, Decathlon can adapt its offerings and services to align with customer expectations. This proactive approach demonstrates a commitment to customer satisfaction and loyalty.

3. Market Expansion and Product Localization: Decathlon operates globally, and the popularity of sports varies across regions. By analyzing sentiment in cricket-related tweets, Decathlon can identify markets where cricket is highly influential and tailor its product offerings to cater to the specific needs and interests of cricket enthusiasts in those regions. This supports Decathlon's goal of localizing its products and services to meet regional demand effectively.

Relevance for Decathlon

4. Brand Reputation Management: Monitoring sentiment on social media allows Decathlon to proactively manage its brand reputation. By addressing customer concerns and feedback in real-time, Decathlon can demonstrate its commitment to quality, customer care, and responsiveness, aligning with its core values.

5. Competitive Advantage: Staying attuned to sentiment in the sports industry, including cricket, provides Decathlon with a competitive advantage. It allows the company to stay ahead of market trends, anticipate customer preferences, and make data-driven decisions that can position it as a leader in the sports retail sector.

In essence, the problem of sentiment analysis in the context of cricket-related tweets aligns with Decathlon's mission to make sports accessible to all and its values of customer satisfaction, innovation, and responsiveness. By addressing this problem, Decathlon can enhance its ability to serve cricket enthusiasts and sports lovers, contributing to its overall business success and growth.

Feasibility

The problem statement of developing an accurate sentiment analysis model for cricket-related tweets is both feasible and realistic for several reasons:

1. Availability of Data: There is a vast amount of cricket-related data, including tweets, available on social media platforms like Twitter. This data source is readily accessible, making it feasible to collect a substantial dataset for analysis.

2. Advancements in Natural Language Processing (NLP): Recent advancements in NLP and machine learning have significantly improved the accuracy and performance of sentiment analysis models. Various pre-trained language models and NLP libraries are readily available for use, making it realistic to build sophisticated sentiment analysis systems.

Feasibility

3. **Machine Learning and Deep Learning Tools:** The availability of machine learning and deep learning frameworks, such as TensorFlow, PyTorch, and scikit-learn, simplifies the implementation of sentiment analysis algorithms. These tools enable researchers and data scientists to develop, train, and evaluate models efficiently.

4. **Established Research in Sentiment Analysis:** Sentiment analysis is a well-established field of research with a vast body of literature. Researchers have developed proven methodologies and best practices for sentiment analysis, providing a solid foundation for tackling the problem.

5. **Business Value:** The problem aligns with the business goals of organizations, including market research, customer engagement, and brand management. Therefore, there is a clear incentive for businesses, such as sports retailers like Decathlon, to invest in sentiment analysis for sports-related topics.

6. **Scalability:** Sentiment analysis models can be scalable, allowing them to handle large volumes of data efficiently. This scalability ensures that the analysis can be applied to a wide range of tweets and topics related to cricket.

Feasibility

7. Ongoing Research and Innovation: The field of sentiment analysis is continually evolving, with ongoing research and innovation. This means that state-of-the-art techniques and models are available for improving the accuracy and performance of sentiment analysis systems.

8. Availability of Expertise: Many professionals and data scientists specialize in NLP and sentiment analysis, making it feasible for organizations to access the necessary expertise to tackle this problem effectively.

In summary, the problem of sentiment analysis for cricket-related tweets is feasible and realistic due to the availability of data, advances in NLP and machine learning, established research, scalability, and the potential for business value. Organizations, including sports retailers like Decathlon, can leverage sentiment analysis to gain valuable insights into customer sentiments and preferences in the sports domain, contributing to informed decision-making and enhanced customer engagement.

Impact

The potential impact of developing an accurate sentiment analysis solution for cricket-related tweets is significant and can benefit various stakeholders and the broader society in several ways:

1. Businesses and Retailers:

- Enhanced Customer Insights: Businesses, including sports retailers like Decathlon, can gain deeper insights into customer sentiments and preferences related to cricket and sports equipment. This information can inform product development, marketing strategies, and inventory management.

- Improved Customer Engagement: Understanding customer sentiment allows businesses to engage with their audience more effectively. Responding to positive sentiments can reinforce brand loyalty, while addressing negative sentiments can lead to improved customer satisfaction.

- Competitive Advantage: Implementing a robust sentiment analysis solution can provide a competitive advantage by enabling businesses to stay ahead of market trends and adapt to changing customer demands.

2. Sports Industry:

- Fan Engagement: Sports teams, leagues, and organizations can use sentiment analysis to gauge the sentiments of fans and adjust their strategies for fan engagement and marketing accordingly.

- Event Management: Sentiment analysis can assist in event planning and management by understanding public sentiments leading up to and during sporting events. This can help in making real-time adjustments to enhance the overall experience.

- Sponsorship Decisions: Companies looking to sponsor sports events or athletes can use sentiment analysis to assess the public perception of potential sponsorships, ensuring alignment with their brand values.

Impact

3. Media and Journalism:

   - Content Creation: Media outlets and journalists can use sentiment analysis to identify trending topics and sentiments in sports-related discussions, aiding in content creation and coverage.

   - Real-time Reporting: Journalists can incorporate sentiment analysis to provide real-time updates on how fans and the public are reacting to sports events and news.

4. Researchers and Academia:

- Academic Research: Researchers can use sentiment analysis as a tool to study the impact of sports on society, fan behavior, and the influence of sentiment on sports outcomes.

   - Innovation: Advancements in sentiment analysis techniques for sports-related data can contribute to innovation in the field of natural language processing and machine learning.

Impact

5. Society at Large:

- Informed Decision-Making: Sentiment analysis can contribute to informed decision-making in various domains, including sports, marketing, and entertainment.

- Fan Experience: Fans can benefit from improved fan engagement strategies and tailored experiences based on sentiment insights.

- Public Awareness: Sentiment analysis can highlight social issues and public sentiments related to sports, fostering awareness and discussion.

6. Social Media Platforms:

- Improved User Experience: Social media platforms can use sentiment analysis to improve user experience by suggesting relevant sports content, topics, and discussions to users based on their interests and sentiments.

In conclusion, an accurate sentiment analysis solution for cricket-related tweets can have a far-reaching impact, benefiting businesses, the sports industry, media, academia, society at large, and even social media platforms. It empowers stakeholders with valuable insights into public sentiments, enabling them to make data-driven decisions, enhance engagement, and adapt to changing trends and preferences in the dynamic world of sports.
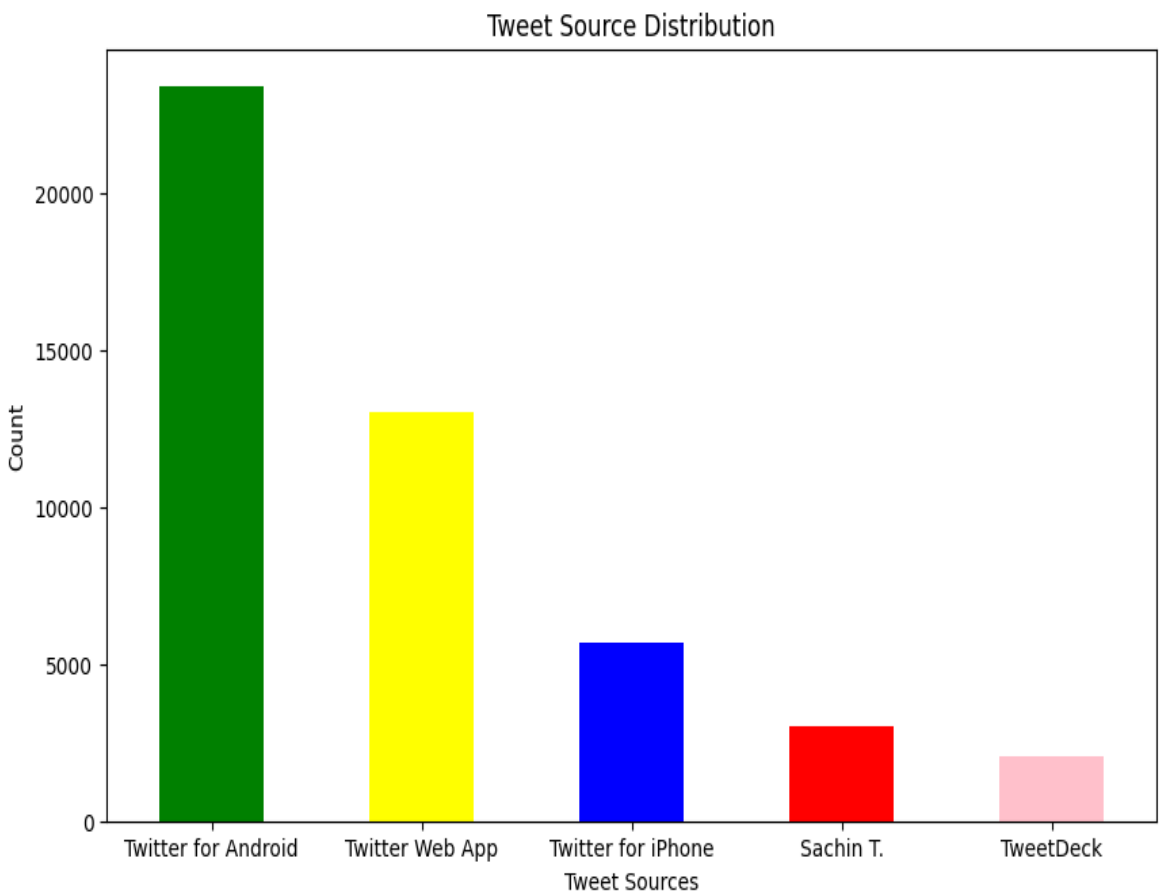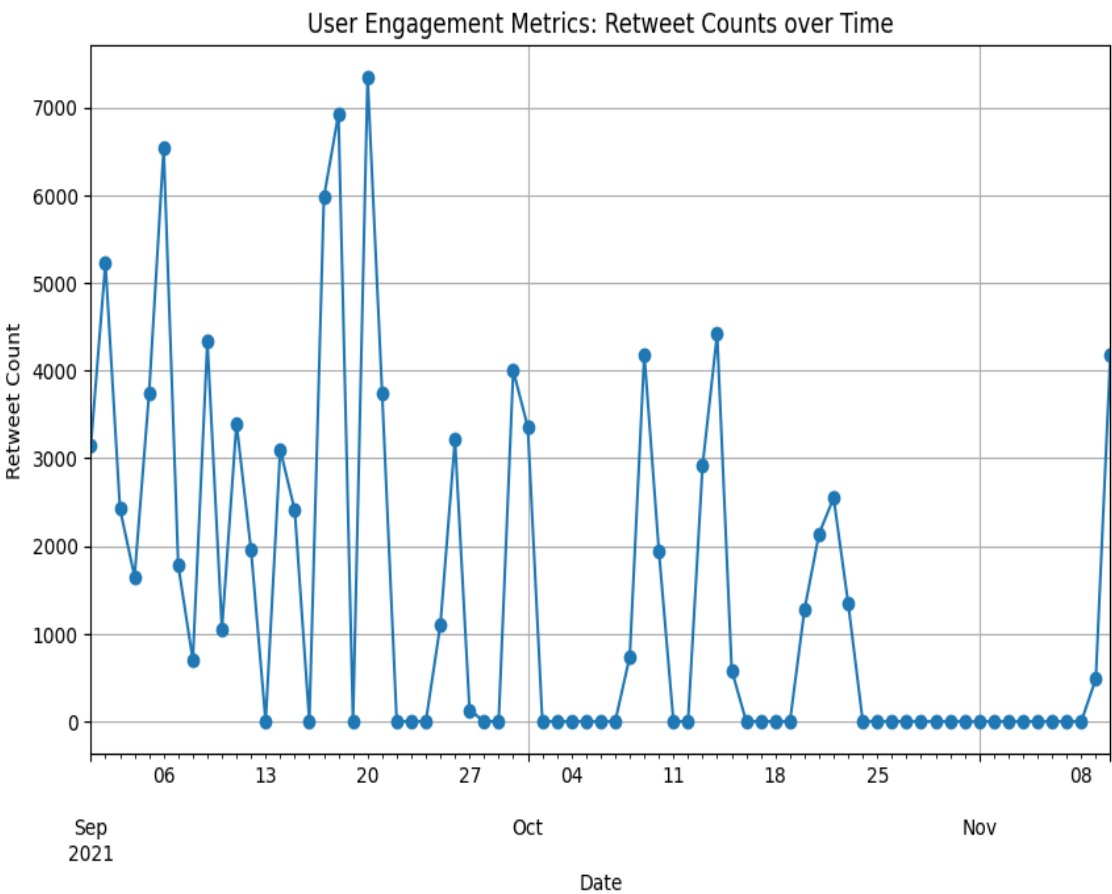
Results

To strengthen the credibility of our proposed sentiment analysis approach for cricket-related Twitter data, we have conducted a comparative analysis

This analysis involves assessing the performance of our method against other established sentiment analysis algorithms. Traditional approaches indicate that the SVM classifier outperforms the other algorithms in terms of accuracy.
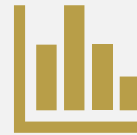

Top User Locations
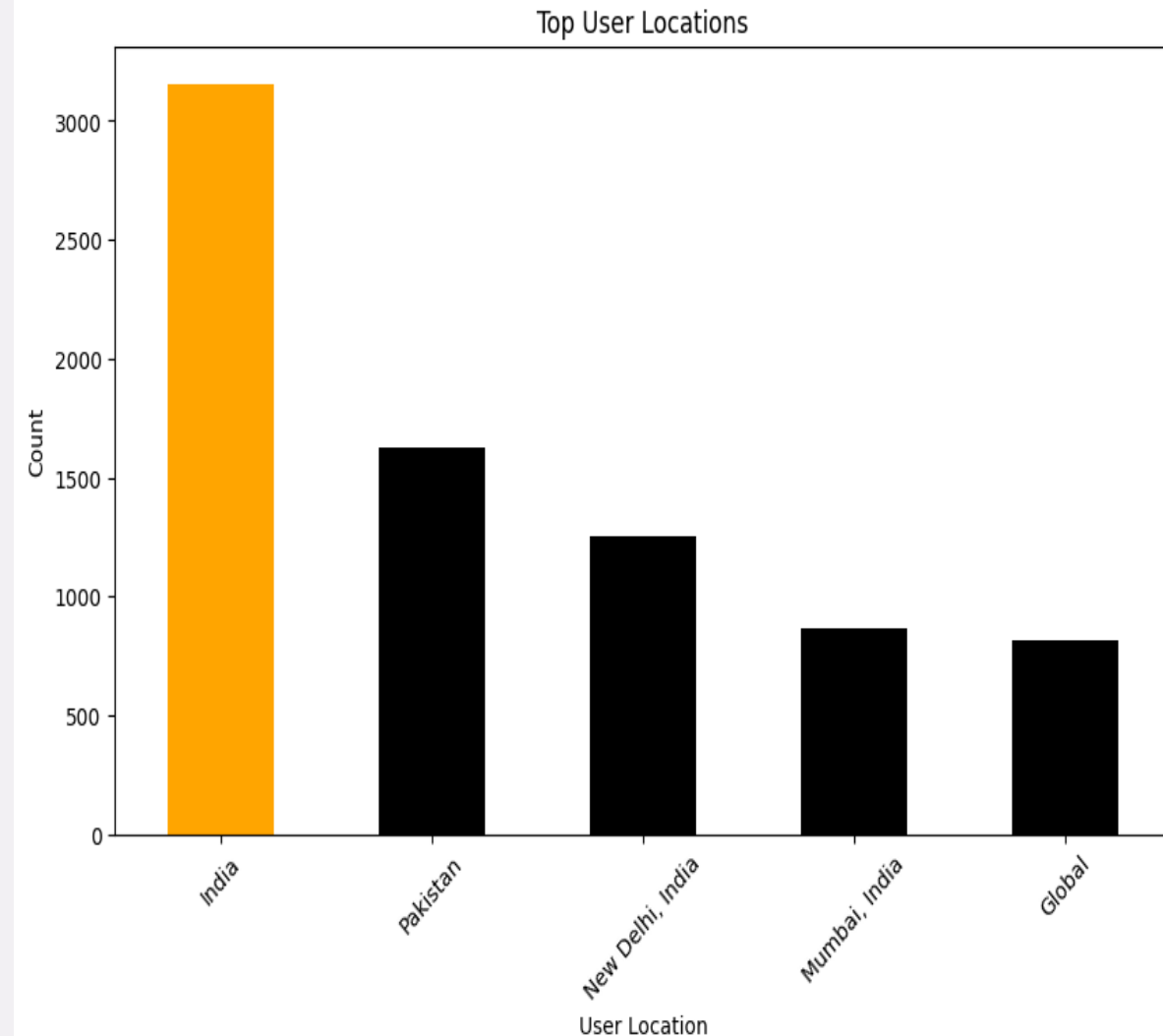
Results for User Engagement Analysis

Results for Logistic Regression

Logistic Regression implementation is carried out from the dataset collected from Kaggle

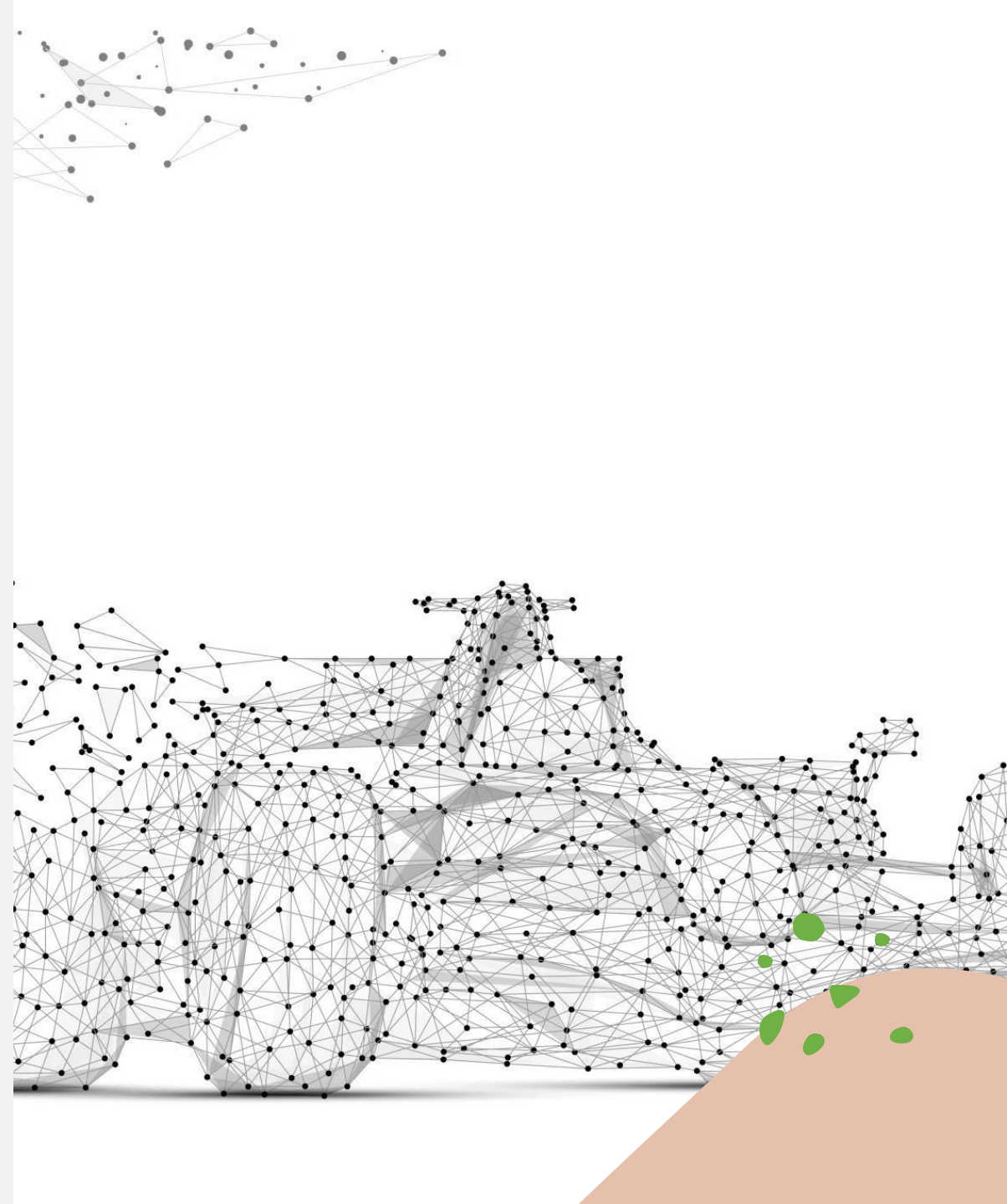The dataset collected about 51862 tweets from the Twitter application

The method's accuracy was estimated as 98.0%

Results for Support Vector Machine

The accuracy for Support vector machine is 98.1%

This provides the highest accuracy in comparison to other machine learning algorithms like Random Forest tree and logistic regression for the given dataset

Results for Random Forest Classifier

The implementation of the Random Forest Classifier is carried out on a dataset consisting of around 25000 tweets

The accuracy for the random forest classifier is 65.9%

The report consisting of accuracy, f-score and precision is given below

Results for Deep Learning

The accuracy for the model developed using the TensorFlow, CNN gives the accuracy as 94.2% which is more accurate than the random forest classifier

Our CNN model's superior performance underscores the effectiveness of deep learning in handling complex sentiment analysis tasks, especially when dealing with unstructured data like tweets

This paper contributes to the field by evaluating the performance of multiple sentiment classification algorithms and proposing a method to enhance classification accuracy

THANK YOU