Information and Inference: A Journal of the IMA (2024) **01**, 1–37
https://doi.org/

Predicting Social Media Engagement through Feature Extraction with Large Language Models: An Analysis of Likes and Shares from User-Generated Content

PAVAN SAI VENKATA DURAGA PRASHANTH KORAPATI*
MS in Computer Science, Northumbria University, Ellison Pl, NE1 8ST, Newcastle upon Tyne, United Kingdom
*Corresponding author: kpsvdp@gmail.com

[Received on 23 May 2024; revised on 31 June 2024; accepted on 24 June 2024]

Abstract: Social media platforms have become essential spaces for public conversation, individual expression, and business promotion in this digital age. Understanding the factors that stimulate user interaction, such as likes and retweets, on these platforms is essential for individuals and companies looking to increase their online visibility. The pressing demand for a further in-depth understanding of the workings of social media interaction is addressed by this study. Although there is a lot of research on user engagement now available, there are gaps in the application of sophisticated natural language processing methods, especially Large Language Models (LLMs), for feature extraction from a variety of dynamic user-generated content. This research offers an innovative method that combines LLMs' ability to analyse and comprehend the complex, context-rich data that is present in social media posts. The goal of the project is to develop prediction algorithms that can precisely predict engagement metrics, likes and retweets, by extracting useful features like emotion. This study uses cutting-edge machine learning techniques, specifically Large Language Models (LLMs) for sentiment analysis and the H2O AutoML framework to generate and choose the best prediction models. This is further used to examine the effects of user emotion, reach, and other characteristics on social media engagement. This work delivers two contributions. Firstly, it aims to improve the theoretical knowledge of the connection between user involvement in social media spaces and their content attributes. In addition to the above, this research seeks to deliver practical insights for makers of content and marketers to help them optimise their plans for better digital engagement. This study has the potential to create an emerging standard in the analytical evaluation of social media content, providing a solid framework for both academic research and practical utilisation in the ever-changing environment of digital communication.

Keywords: Natural Language Processing (NLP); Large Language Models (LLMs); Sentiment Analysis; Machine Learning; H2OAutoML; Digital Engagement; Social Media Analytics; Prediction Algorithms; User-Generated Content

1. Introduction

The purpose of this study is to predict social media engagement by feature extraction from large language models. It is crucial to understand social media engagement for the sake of digital marketing, audience engagement tactics, and platform algorithms. This study looks at the text features that affect how users perceive content generated on social media and its effects on important engagement metrics like likes and retweets.

1.1. *Background*

The rapid expansion of social media platforms has significantly increased the volume of User-Generated Content (UGC). Extraction strategies and using massive language models to research text create large

gaps despite sizeable research on social media engagement prediction. The present literature has yet to completely elucidate the complicated relationship between language complexity and measures of consumption. This study aims to fill this gap by supplying deeper insights into how language complexity affects UGC sentiment [9]. The study also provides a bridge into subsequent engagement metrics, including likes and shares. The study seeks to enhance our understanding of social media dynamics by addressing such obstacles.

1.2. *Aim and Objectives*

1.2.1. **Aim**

Explore how language limitations have an effect on customer sentiment and social media engagement.

1.2.2. **Objectives**

- Explore the relationship between language complexity and user-generated content to discover patterns in social media interactions.
- Examine the impact of user-generated content to identify how it influences important communication characteristics like sharing and likability.
- Improve predictive modelling by using large-scale language models to predict social media interactions based on language complexity and sentiment analysis.

1.3. *Research questions*

- How does language complexity relate to the emotional content of consumer movements as well as reflecting patterns in social media interactions?
- What effect do manufacturers have on key communication functions, including likes and shares?
- How can predictive modelling be advanced using large-scale language models to forecast social media interactions that are primarily based on language complexity and sensitivity evaluation?

1.4. *Research Problem*

This research focuses on the need for deeper insights into the dynamics of social media engagement. Challenges encompass navigating changes in user language and changing user possibilities. Further, this makes it difficult to analyse product sentiment and expect engagement metrics. It is essential to conquer those challenges by tailoring content material techniques and platform algorithms to grow audience engagement and pleasure.

1.5. *Research Framework*

This research study proposes a thorough research strategy to investigate the factors impacting user interaction on social media platforms, emphasising likes and retweets as important engagement measures. The research starts with an introduction that emphasises the value of social media for promoting businesses, fostering public discourse, and allowing for individual expression.

The literature review discusses existing user engagement research, the use of Natural Language Processing (NLP) to analyse user-generated information, and the potential applications of Large Language Models (LLMs) in sentiment analysis.

The methodology part describes the procedures for collecting data from social media sources, cleaning and preprocessing, extracting emotions using LLMs, and performing sentiment analysis. The process of creating and choosing the best prediction models using the H2O AutoML framework is explained, along with the metrics used to assess the model's performance.

The analysis portion looks at the relationship between user emotions and engagement metrics, how reach and other factors affect engagement, and uses statistical tests to validate the models.

The performance of the prediction models and the significance of different features in predicting engagement are presented in the results and discussion sections.

The main conclusions are outlined in the conclusion, along with the contributions to academic knowledge and real-world applications and potential directions for further research.

2. Literature Review

2.1. *Introduction*

It is essential to understand social media engagement and sentiment analysis in today's virtual environment. It helps the organisational groups measure customer options, put together advertising and marketing techniques, and finally, and most importantly, helps to make a good brand image. Through sentiment analysis, organisations can interpret user sentiment, identify trends, and tailor content to better communicate with target audiences and be compatible with all differences. Further, this helps to grow strong online relationships between customers and organisations.

The literature review covers a huge range of subjects in social media studies. This is done from purchaser engagement to sentiment evaluation and system learning programmes [11]. This further discusses discovering elements that are affecting social media interactions, sentiment evaluation techniques, and the effect of factors the user reports on platform dynamics [8]. This analysis of the literature makes use of loads of techniques, which include data mining, herbal language processing, and predictive modelling. Further, this helps to discover customer behaviour as well as content material interplay patterns and the effectiveness of diverse analytical procedures on comprehension and predictively on social media occasions.

2.2. *Analyzing Social Media Engagement*

2.2.1. **Factors influencing engagement**

The review of this literature highlights various elements that have an effect on engagement in social media structures. This is done especially for content control and branding on customer interactions. User-generated content (UGC) has a tendency to have excessive tiers of engagement, with users being more likely to proportion and engage with peers than manufacturers [12]. On the other hand, the contents generated by the brand may be quite visible in low-involvement situations. The nature of the content material itself also influences the behaviour of the customers. This also suggests that content material plays a crucial role in figuring out engagement.

Furthermore, consumer engagement is an important issue that influences engagement behaviour, with immoderate tiers of engagement in the usage of interplay and engagement with social media content material [6]. Understanding these tendencies is significant for content creators and entrepreneurs to meet the organisations' strategies and, most importantly, maximise engagement metrics [14]. Additionally,

the insight from this study causes predictive models and frameworks for greater accurate prediction of social media engagement. This is based on elements inclusive of content types and consumer engagement within particular organisations.

2.2.2. **User Behaviour and preferences**

The behaviour of these users and their preferences for various content is quite special, as this is highlighted in the literature. Research indicates that customers show off different styles when interacting with user-generated content (UGC) as opposed to brand-generated content (BGC) on the platforms of social media. UGC tends to evoke higher ranges of interaction as compared to BGC, which is particularly because of its perceived authenticity and relevance [6]. Users are much more likely to depend on user-generated content than on promotional materials generated by brands. This can further increase likes, shares and comments on UGC.

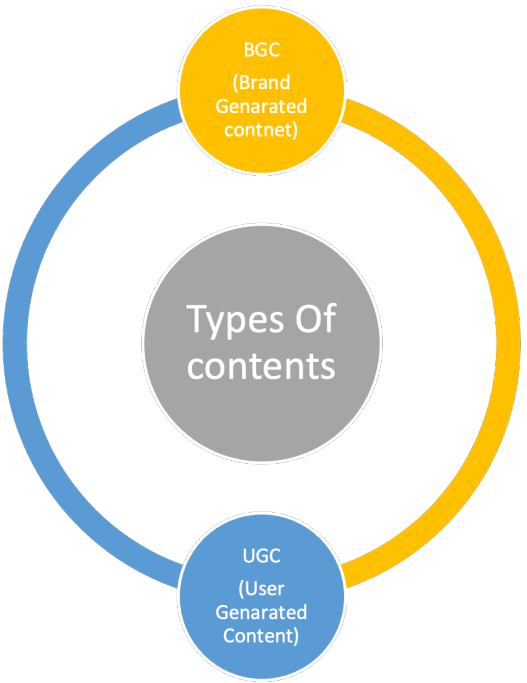

FIG. 1. Types of contents (Source: Self-developed)

Moreover, it is the type of content material that affects the behaviours of the customers. For instance, posts that evoke an emotional reaction or spark communication tend to get greater engagement than simply news or promotional posts of the brands. This approach requires creators to be cognizant not only of their message but also of the emotional impact it creates to maximise customer engagement.

Furthermore, the degree of customer involvement plays an important function in figuring out diversification among customers [4]. Branded content material can never the less entice interest as well, albeit to a lesser quantity in comparison to UGC in areas of low consumer engagement or

interest.However, UGC tends to outperform BGC in communicative aspects such as likes, shares and comments in some situations where the customers are relatively engaged or emotionally concerned.

Understanding these elements is important for marketers as well as content creators to expand effective engagement strategies. They are able to optimise engagement and improve the general effectiveness of their social media campaigns by striving to create content material in keeping with consumer preferences and behaviours [5]. Furthermore, insights from this study can lead to predictive models and algorithms. This is aimed at predicting social media engagement among customers. This is generally done based on the content material, the users involved, and information-pushed advertising techniques.

2.3. *Predictive Models and Machine Learning Techniques*

Different types of methods have been used within the reviewed research to take a look at social media engagement and sentiment [7]. These techniques consist of traditional statistical techniques, which include logistic regression as well as long short-term memory (LSTM) networks, and they are held in the domain of superior machine learning algorithms. Logistic regression is a classical statistical approach, and this has been utilised in a few studies to model the relationships among predictor variables such as linguistic features, sentiment scores, and engagement measures, which consist of likes and shares. This method gives some important insights into the importance and route of the institutions. This also affords a baseline framework for knowledgeable elements that influence social media. On the other hand, LSTM networks are a sort of recurrent neural network (RNN). This has won recognition because of its capability to seize series dependencies and durability in the data. These networks excel in chronological evaluation as well as making them ideally fitted for obligations. This is done together with forecasting and predictions of publicity over time based on past data.

Ensemble learning, which is the combination of multiple datasets, has additionally been used in lots of research to improve prediction performance. Ensemble strategies can increase the accuracy and robustness of predictive models, which are carried out in all social media interactions, by using the strengths of different algorithms and decreasing individual weaknesses. Overall, the variety of methodologies displayed the complexity of studying social media records and predicting engagement. Researchers are using an aggregate of traditional statistical techniques and the latest machine learning strategies to extract significant insights. This also increases predictive models that may better predict advertising strategies and decision-making tactics inside the domain of digital panorama.

2.3.1. **Accuracy rates and implications for understanding user behaviour**

The levels of precision achieved with the help of numerous techniques. This consists of logistic regression as well as LSTM networks and ensemble understanding [13]. This provides important insights into the effectiveness of predictive frameworks in the knowledge of user behaviour in social media. The excessive rates of accuracy indicate the reliability of this version in predicting consumer engagement metrics which include likes, shares, and sentiment analysis. Research can take advantage of this by getting deeper insights into the elements affecting social media engagement by describing usage behaviours [8]. This also enables organisations to optimise their content material management strategies and advertising and marketing campaigns and better match their audience. This can improve target market engagement as well as customer reputation and, ultimately, enterprise fulfilment.

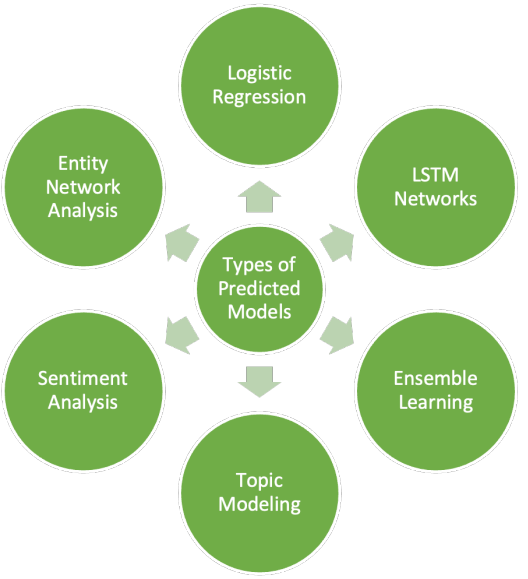

FIG. 2. Types of predicted models (Source: Self-developed)

2.4. *Sentiment Analysis in Social Media*

The studies reviewed used numerous methods to research social media data and understand consumer behaviour. Topic modelling is the approach of figuring out summary issues inside the accumulated papers using themes to display the content material of commonplace themes and discussions in social media. Research gained insights into subjects that engage customers and drive discussion on boards like Twitter and Reddit by figuring out topics of interest.

On the other hand, sentimental categorization performed an essential function in analysing the emotional tone of the data. Research may want to determine how users experience precise subjects by classifying posts or comments as positive, bad, or impartial. The sentiment analysis particularly features topics, brands, or events [1]. This study contributed to knowledge of how sentiments on social media have an effect on engagement metrics and client behaviour.

Entity network evaluation focused on the relationships among the organisations mentioned in social media content material, along with individuals as well as the companies and products. Visualising these networks further helps the researchers visualise social networks around unique subjects and studies. This approach provided insights into internet community improvement and had an effect on influential users or businesses in shaping the discourse and engagement of consumers.

The use of sentiment analysis and topic modelling produced priceless insights into public opinion on a wide range of topics pertaining to social media architecture. Sentiment analysis techniques distinguish between high-quality, negative, or between factors expressed on particular topics or issues by reading user-generated content, consisting of posts, comments, and evaluations. These insights underline the public's perceptions as well as their attitudes and feelings on plenty of issues, from social problems to politics to purchasing products and the sentiments of brands.

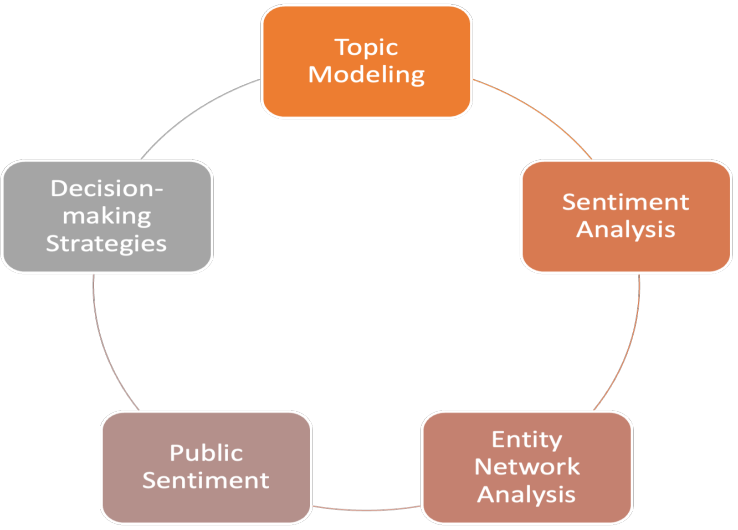

FIG. 3. Different components of social media data analysis (Source: Self-developed)

Understanding public sentiment allows companies and policymakers to gauge public opinion, select inclinations, and assess the effect of their actions or advertising and marketing campaigns. Furthermore, the insights can help with decision-making strategies. This includes providing appropriate responses, using focused strategies, and employing tactics that sustain the positive feelings of the target market. Businesses, governments, and different stakeholders can improve conversation strategies by using such insights. Further, this helps to indicate the concerns very effectively and promotes good engagement with the respective organisations' target audience.

2.5. *Impact of Social Media on Specific Domains*

Many studies have delved into numerous fields to examine the effect of social media on businesses and different segments of society. For example, research has tested the impact of social media on organisational finance. This is inclusive of analysing blockchain adoption, initial coin offerings (ICOs), and safety token services (STOs). In addition, the study tested the effect of social media sentiment on urban tour styles [2]. This is done mainly for remote work and the effect of the COVID-19 pandemic on travel behaviour.

Additionally, the researchers tested the role of social media in catastrophe response with the use of sentiment evaluation and real-time analytics [3]. This helps gauge public sentiment and predict flood severity. This study affords treasured insights into the intersection of social media and sectors. The study also gives insights into informing corporate practices, urban planning techniques, and catastrophe management efforts. Understanding the consequences of social media engagement for those sectors in terms of capacity to assist stakeholders to adapt to tendencies as well as mitigate risks to achieve their respective targets.

Different processes have been used to observe the impact of social media in particular industries. This consists of sentiment analysis as well as machine learning algorithms and data mining techniques. For example, in the study of urban commuting, researchers incorporated social media sentiment evaluation

into regression models to recognise the effect of far-off work on mobility styles. Similarly, actual-time analysis of social media data in catastrophe management has enabled the prediction of flood importance through the use of the LSTM network and ensemble learning methods. These strategies gained insights into public sentiment as well as their preferences and practices and supplied an informed framework for decision-making in infrastructure.

2.6. *Challenges*

A common task in social media studies is decreasing record bias and ensuring sample interpretability and generalizability. Future research guidelines may also acknowledge that, in order to get around these obstacles and boost the precision of social media analytics tactics and execution, it has been very beneficial to integrate multiple data sources, handle privacy concerns, and improve machine-learning knowledge of algorithms.

The literature overview presents valuable insights into the elements affecting social media engagement as well as techniques of sentiment evaluation. This literature review also helps to recognise user behaviour. The important findings include the types of content and the engagement of different machine learning techniques. Finally, these help to improve future research in this particular field.

2.7. *Introducing novel analytical framework*

This research addresses challenges in current literature by utilizing a novel approach by using latest transformer based models, which offers higher degree of attention to natural language processing and aids in better understanding of text emotions. Selecting an appropriate transformer based model is a challenge for this study. After numerous trial and error approaches on the dataset, it is decided to select a model pre-trained on multilingual text to accurately predict emotions from a wide variety of language texts or tweets. To facilitate more flexibility and ease of understanding, this study also incorporates H20 Auto ML framework which helps in automating tasks of running various models and perform post prediction statistical tests such as Variable importance, Partial Dependence Plots (PDP plots), SHAP summary and Individual Condition Expectation plots (ICE plots) which provides deeper understanding of the selected machine learning model.

2.8. *Summary*

This section sets the groundwork for identifying social media activity and evaluating sentiment. This also emphasises its importance for organisational success. A review of the literature on subjects like customer engagement and machine learning applications is included with this. Analysis of the drivers of social media engagement as well as user behaviour, predictive models, sentiment analysis methods, and challenges yields valuable insights for future research.

3. Research Methodology

3.1. *Philosophy*

The research methodology of this study follows a positivist approach. This approach prioritises empirical observations and scientific methods to understand social phenomena. In this study, quantitative data analysis is used to examine correlations among variables and make predictions based on mathematical models [10]. This method emphasises the importance of objective measurement and structured inquiry to uncover patterns and relationships within the research. The study aims to provide

insights based on empirical evidence through rigorous statistical analysis, contributing to a better understanding of the phenomena being investigated.

3.2. *Data Collection*

3.2.1. **Data Source**

The data for this evaluation includes 10,000 tweets, which are sourced from the Data World website. This is mainly generated from the ”Twitter data in sheets.xlsx” dataset. The use of this data set gives a wealthy archive of real world tweets.This dataset can be accessed here.

3.2.2. **Data Description**

Analyze the Twitter data To analyse the Twitter data, the necessary Python library, Pandas, was first imported. The dataset was then loaded from an Excel file which contains three separate sheets labelled ’Tweet’, ’Location’, and ’User’. Each sheet was read into its own DataFrame using Pandas’ read-excel function. The openpyxl engine was used to read the Excel file. The first few rows of each DataFrame were displayed to understand their structure and the type of data they contain. This initial exploration is shown in Fig. 4.

Twitterdatainsheets.xlsx

	TweetID	Weekday	Hour	Day	Lang	IsReshare	Reach	RetweetCount	Likes	Klout	Sentiment	text	LocationID	UserID
0	tw-682712873332805633	Thursday	17.0	31.0	en	0.0	44.0	0.0	0.0	35.0	0.0	We are hiring: Senior Software Engineer - Prot...	3751.0	tw-40932430
1	tw-682713045357998080	Thursday	17.0	31.0	en	1.0	1810.0	5.0	0.0	53.0	2.0	RT @CodeMineStatus: This is true Amazon Web Se...	3989.0	tw-3179389829
2	tw-682713219375476736	Thursday	17.0	31.0	en	0.0	282.0	0.0	0.0	47.0	0.0	Devops Engineer Aws Ansible Cassandra Mysql Ub...	3741.0	tw-4624808414
3	tw-682713436967579648	Thursday	17.0	31.0	en	0.0	2087.0	4.0	0.0	53.0	0.0	Happy New Year to all those AWS instances of o...	3753.0	tw-356447127

FIG. 4. Analyze the Twitter data

After loading the data, it was observed that some column names had leading or trailing spaces, which could cause issues when referencing these columns. The info() method was then used to display information about the ’Tweet’ DataFrame, including the number of entries, non-null counts, and data types for each column.

To clean up the column names, the str.strip() method was applied to the column names of each DataFrame to remove any extra spaces. This procedure is shown in Fig . 5.

In the Location sheet, the first five rows were displayed using sheet2.head(), revealing columns such as LocationID, Country, State, StateCode, and City. This data contains 6,291 entries with some missing values in the State and City columns. This sheet info is presented in Fig. 6.

Sheet3 has columns for UserID and Gender. This sheet contains 100,001 entries with two columns, both fully populated. Sheet 3 data is presented in Fig. 7.

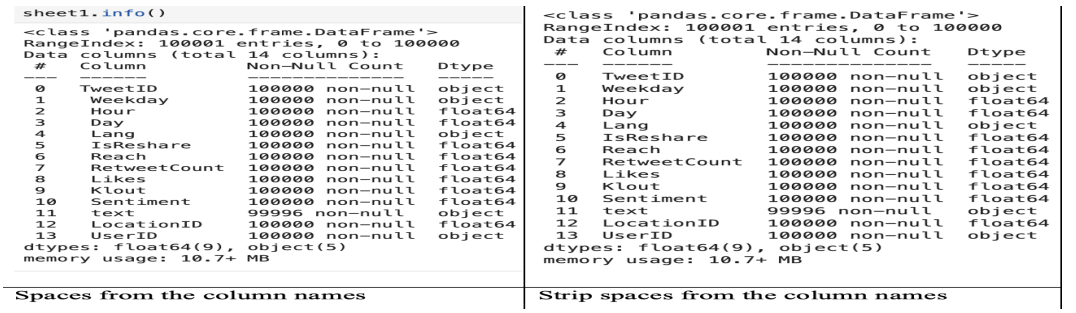

FIG. 5. sheet (columns names)

```
sheet2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6291 entries, 0 to 6290
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   LocationID  6289 non-null   float64
 1   Country     6290 non-null   object
 2   State       6174 non-null   object
 3   StateCode   6180 non-null   object
 4   City        6132 non-null   object
dtypes: float64(1), object(4)
memory usage: 245.9+ KB
```

FIG. 6. sheet2.info()

```
sheet3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100001 entries, 0 to 100000
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   UserID  100000 non-null  object
 1   Gender  100000 non-null  object
dtypes: object(2)
memory usage: 1.5+ MB
```

FIG. 7. sheet3.info()

To combine the data from all three sheets, the Tweet Sheet (sheet 1) was first merged with the Location Sheet (sheet 2) on the 'LocationID' column using an outer join, creating a DataFrame with both tweet and location details (fig. 8). Next, this result was merged with the User Sheet (sheet 3) on the 'UserID' column, again using an outer join, resulting in a final DataFrame that contains combined tweets, location, and user information (Fig. 9).

```
# Display the first few rows of the final merged DataFrame
final_df.head()
```

TweetID	Weekday	Hour	Day	Lang	IsReshare	Reach	RetweetCount	Likes	Klout	Sentiment
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tw-88907264	Tuesday	7.0	12.0	en	0.0	991.0	1.0	0.0	42.0	1.00
tw-323795713	Tuesday	7.0	12.0	en	0.0	991.0	1.0	0.0	42.0	1.00
tw-281409536	Monday	12.0	15.0	en	0.0	390.0	0.0	0.0	25.0	-2.00
tw-141334017	Wednesday	23.0	23.0	es	0.0	402.0	0.0	0.0	32.0	0.17

FIG. 8. Merge sheets DataFrame 1

text	LocationID	UserID	Country	State	StateCode	City	Gender
NaN	NaN	_x001A_nknow531394	NaN	NaN	NaN	NaN	Male
.@SailPoint chooses @Dynatrace @Ruxit for its ...	3811.0	tw-10000632	United States	Michigan	US-MI	Detroit	Female
.@SailPoint chooses @Dynatrace @Ruxit for easy...	3811.0	tw-10000632	United States	Michigan	US-MI	Detroit	Female
Validation for private cloud or the dominance ...	3766.0	tw-100012605	United States	Massachusetts	US-MA	Boston	Unisex
@Ruso_tv aws tu a mi corazon	1961.0	tw-1000137907	Mexico	Tamaulipas	MX	Nuevo Laredo	Female

FIG. 9. Merge sheets DataFrame 2

Data cleaning and preprocessing To clean and preprocess the data, the dataset was first checked for missing values and appropriate data types. Duplicate rows were removed. The 'Hour' and 'Day' columns were converted to numeric types, and 'IsReshare' was ensured to be a boolean. The text was cleaned up in the 'text' column by changing it to lowercase, removing URLs, non-whitespace characters that came after '@,' special characters, numbers, and extra spaces. The unnecessary columns 'TweetID', 'UserID', and 'LocationID' columns were dropped.

Missing Values in Dataset Missing values were checked. Rows with missing 'text' column values were removed. Missing values in the 'State', 'StateCode', 'City', and 'Gender' columns were replaced with "Unknown". The cleaned dataset was saved to a new CSV file. This process ensured no critical

data was missing, improving the dataset's quality for further analysis. The process is presented in Fig. 10, showing the steps taken to ensure data completeness and accuracy.

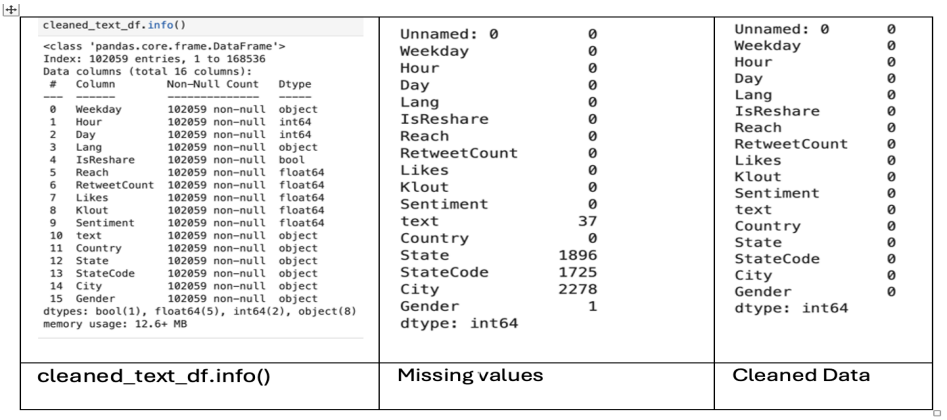

FIG. 10. Missing Values in Dataset

3.3. *Data visualisation*

3.3.1. **Word cloud of tweets**

To create a word cloud for visualising the most frequently used words in the tweets, the text was cleaned by removing URLs, special characters, and extra spaces and converting it to lowercase. Using the nltk library, stopwords for multiple languages were downloaded and added to a set of stopwords. All the cleaned tweet text was then combined into a single string. The word cloud was generated using the WordCloud library and subsequently displayed to visualise the most common words in the tweets (Fig. 11).

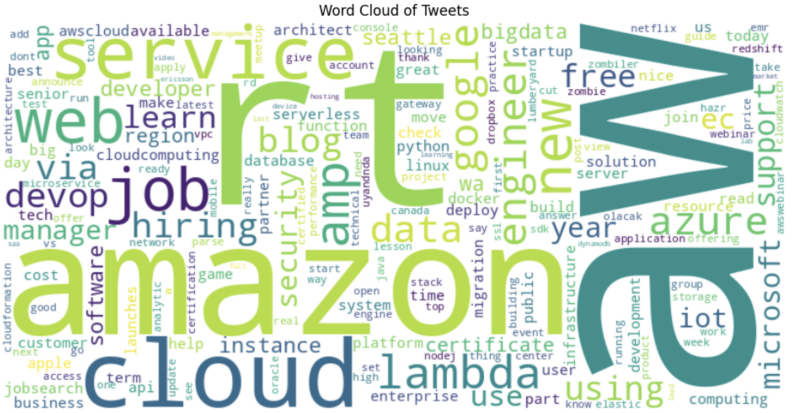

FIG. 11. Word cloud of tweets

3.3.2. **Tweets by Weekday**

To analyse tweet activity by weekday, a bar chart was created using the seaborn library. The resulting chart illustrates the distribution of tweets across the week, showing that tweet activity peaks on Tuesday and Wednesday and is lowest during the weekend. This visualisation, presented in Fig. 12, helps identify patterns in user engagement and can be used to optimise the timing of tweets for better reach and interaction.

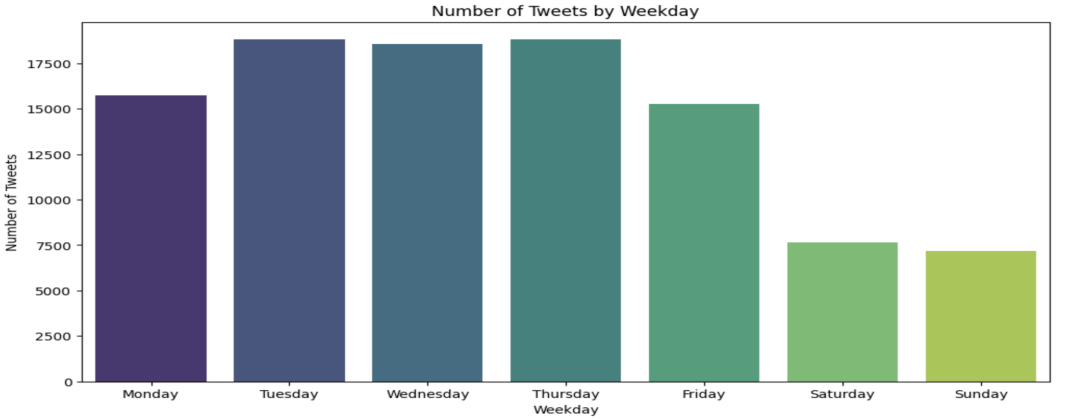

FIG. 12. Number of Tweets by Weekday

3.3.3. **Tweets by Hour**

A bar chart was made to examine Twitter activity by hour of the day. The distribution of tweets over a 24-hour period is displayed in the resulting chart, as shown in Fig. 13. This enables the discovery of peak tweeting hours. This graphic is essential for figuring out when users are most engaged, which helps with tweet scheduling to maximise engagement and reach.

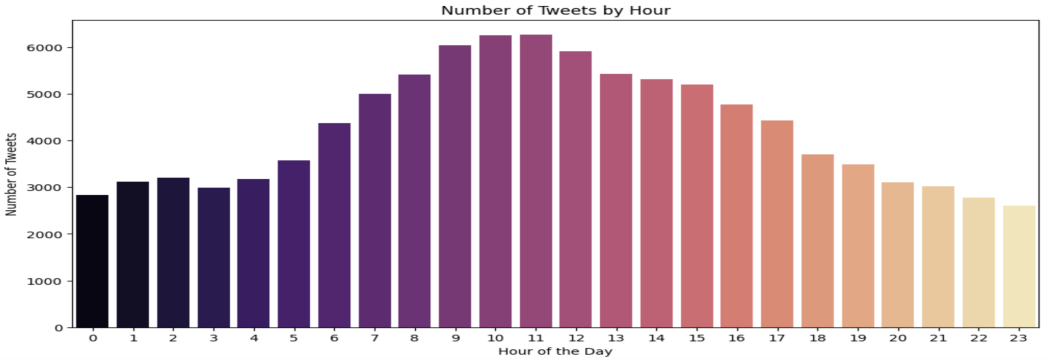

FIG. 13. Number of Tweets by hour

3.3.4. **Gender Distribution**

A bar chart was used to assess the gender distribution of users in the dataset. The code produced a count plot showing the number of users for each gender category. The distribution of genders among the users is displayed in the resulting chart, which is shown in Fig. 14, emphasising the representation of various genders in the dataset. Understanding the demographic makeup of the user base is essential for targeting content and interaction strategies to the audience.

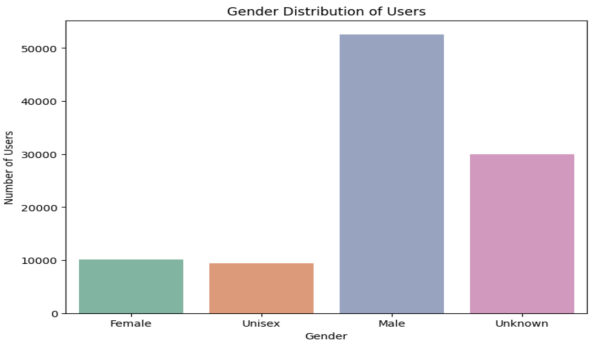

FIG. 14. Gender Distribution of users

3.3.5. **Heatmap of tweet activity**

A heatmap showing tweet counts with annotations was plotted. As shown in Fig. 15, this visualisation aids in identifying peak activity times and days by displaying the density of tweets at various hours during the week.

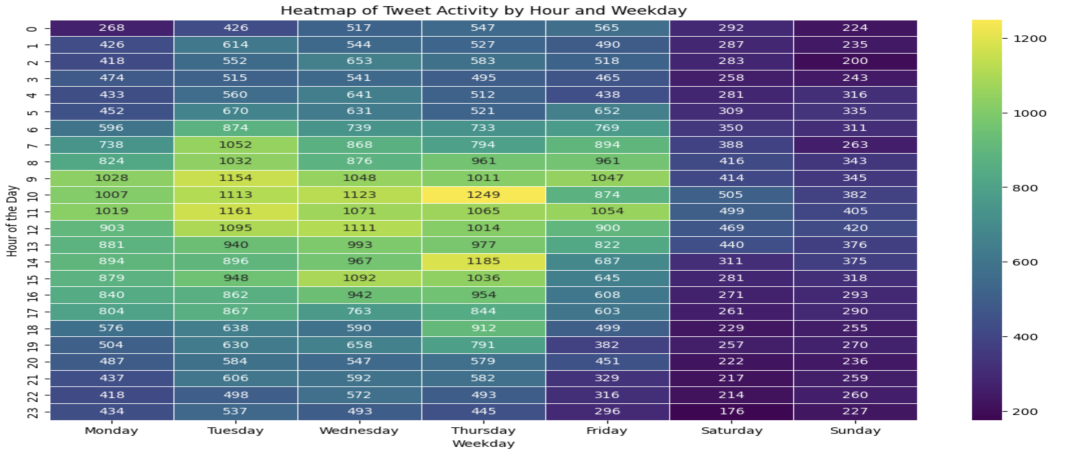

FIG. 15. Heatmap of Tweet Activity by Hour and Weekday

3.3.6. **Correlation matrix**

The correlation matrix for the selected numerical columns was computed, measuring the strength and direction of the relationships between these variables. This visualisation, presented in Fig. 16, helps identify which variables are strongly correlated, providing insights into how different aspects of tweet performance and user interaction are related

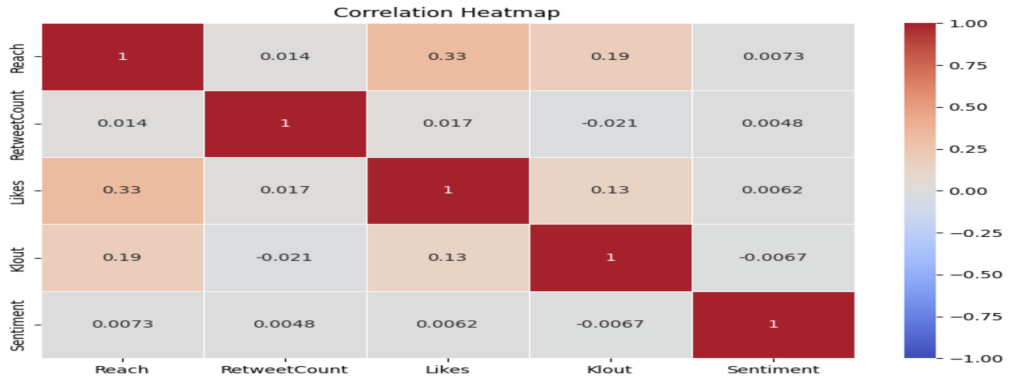

FIG. 16. Correlation Heatmap

3.3.7. **Average sentiment by country**

A bar chart was created by first sorting data by ”country,” and the average sentiment score was determined for each nation. The average sentiment scores for each country were plotted. The resulting chart, shown in Fig. 17, displays the variation in sentiment across different countries, highlighting which countries have the most positive or negative tweet sentiments. This analysis helps understand regional differences in tweet sentiment, which is valuable for tailoring content and strategies to different audiences.

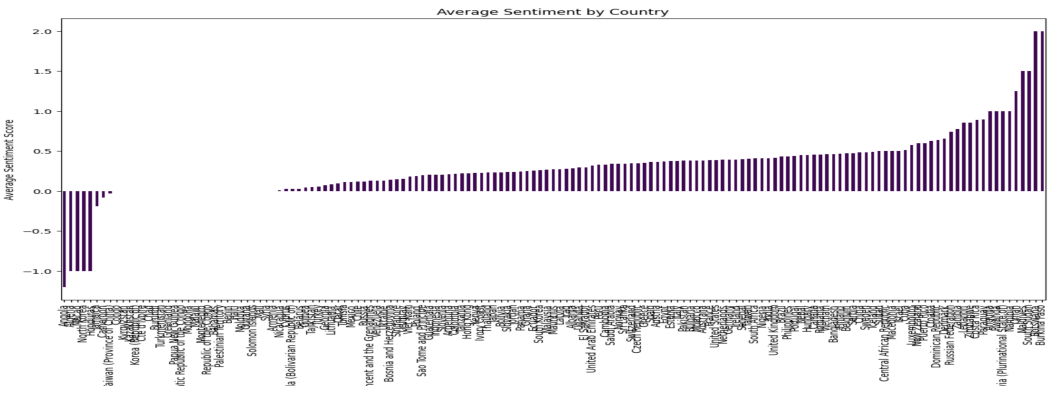

FIG. 17. Average Sentiment by Country

3.3.8. **Average likes and retweets by gender**

The initial chart (Fig. 18) shows the average number of retweets by gender, with men having the highest average number, followed by unisex, female, and unknown genders. The average like count for the unknown gender group is significantly higher than the other categories, as seen in the second figure (Fig. 18), which shows average likes by gender. This method helps in visualising how gender differs in social media involvement.

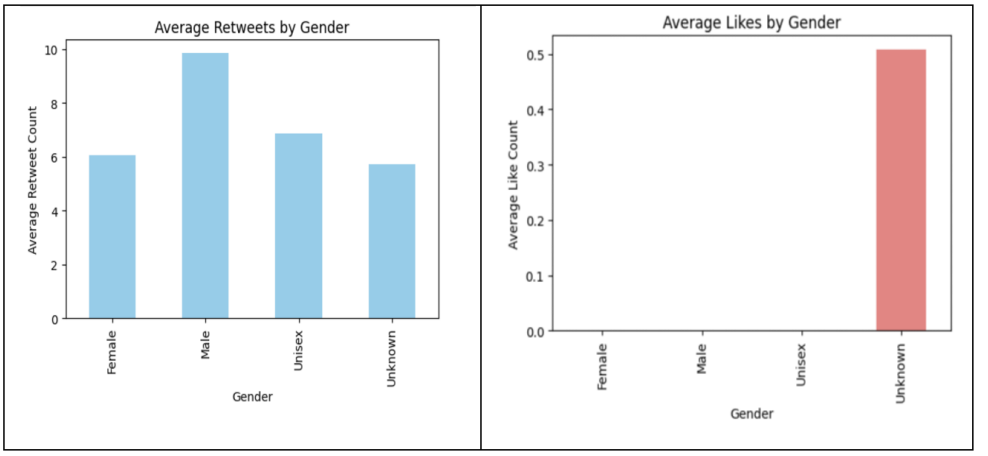

FIG. 18. Average likes and retweets by gender

3.3.9. **Distribution of tweets by language**

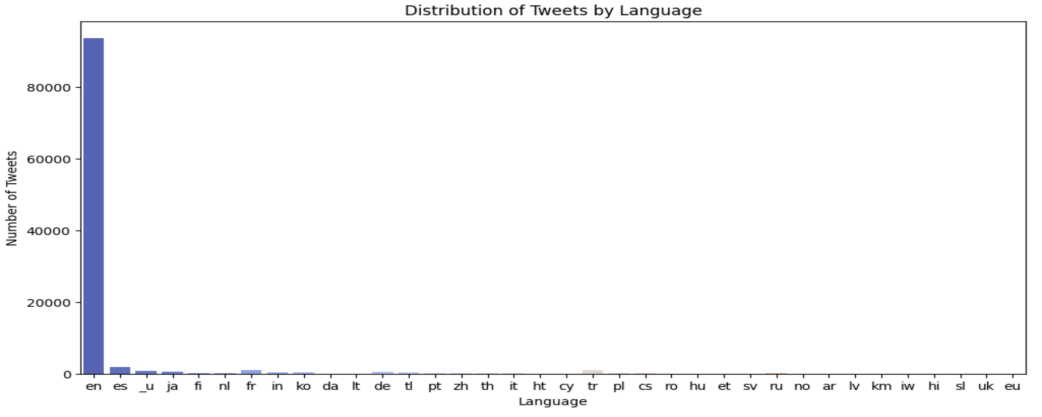

FIG. 19. Distribution of Tweets by Language

A bar chart plot (Fig. 19) highlights that English tweets dominate, with significantly fewer tweets in other languages. This visual representation helps understand the prominence of English in the dataset.

3.3.10. **Geographical Distribution of tweets by country**

A count plot was used to visualise this geographical distribution of tweets and analyse it by country. The countries were arranged according to tweet frequency. This graphic (Fig. 20) clarifies how tweets are distributed throughout various nations.

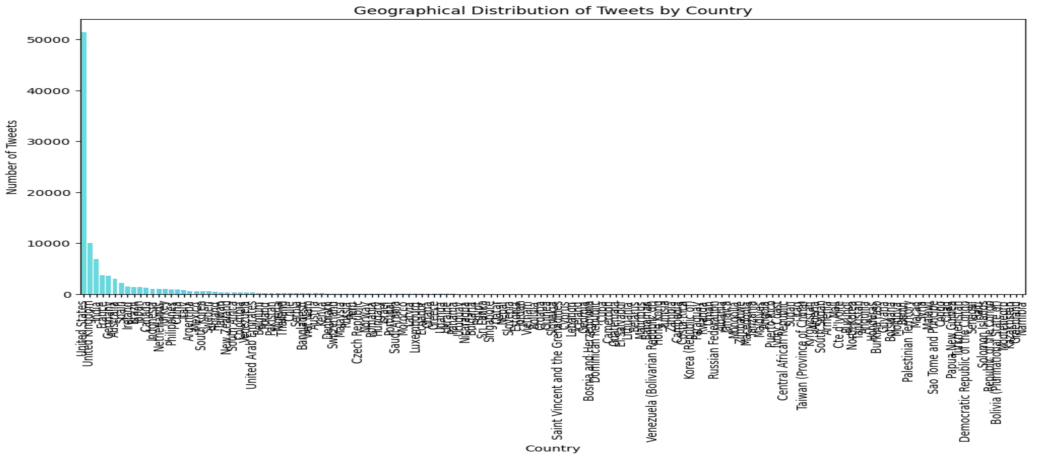

FIG. 20. GeographicalDistribution of tweets by country

3.4. ***Sentiment Analysis***

3.4.1. **Pre-trained model for emotion detection**

For emotion detection, pre-trained libraries AutoModelForSequenceClassification and AutoTokenizer were used. The ”cardiffnlp/twitter-roberta-base-emotion” model, designed for emotion detection in multilingual text, was loaded with the pre-trained method. Default emotion labels were retrieved from the model’s configuration.

3.4.2. **Experimenting on sample text**

Using a transformer library emotion detection model that has already been trained was part of the technique. To ensure that the ”cardiffnlp/twitter-roberta-base-emotion” model for text classification returned all scores, it was loaded using the pipeline function. From the cleaned dataset, for initial analysis, sample texts in the following languages in random were chosen: English, Spanish, French, German, and Italian. These texts were then verified for emotion classification. This method illustrated the model’s adaptability and efficacy in multilingual settings by demonstrating how well it could categorise emotions across several languages.

3.4.3. **Applying transformer model on full dataset**

The transformers library and the tqdm library for progress tracking were utilised. The emotion detection pipeline was loaded with the specified model and tokenizer. A function to classify emotions was created to process texts in batches, classify emotions, and determine the most probable emotion category with associated probabilities. This function updated a results dictionary and identified the best emotion for each text. Emotion probabilities and the best emotion categories were appended to the

dataset. The updated DataFrame was displayed and saved to a new CSV file (Fig. 21), ensuring comprehensive emotion analysis.

```
# cleaned_data.to_csv('twitter_roberta_based_emotion_sentiment_updated_data.csv')  -- uncomment this before running again
```

lout	...	Country	State	StateCode	City	Gender	joy_probability	optimism_probability	anger_probability	sadness_probability	best_emotion
42.0	...	United States	Michigan	US-MI	Detroit	Female	0.054265	0.699100	0.195411	0.051224	optimism
42.0	...	United States	Michigan	US-MI	Detroit	Female	0.067866	0.274897	0.578567	0.078670	anger
25.0	...	United States	Massachusetts	US-MA	Boston	Unisex	0.983336	0.002837	0.003984	0.009843	joy
32.0	...	Mexico	Tamaulipas	MX	Nuevo Laredo	Female	0.150150	0.572824	0.129507	0.147518	optimism
32.0	...	Mexico	Tamaulipas	MX	Nuevo Laredo	Female	0.263906	0.326902	0.151249	0.257943	optimism
...	...	...	...	...	...	...	...	...	...	...	...
55.0	...	Netherlands	North Holland	NL	Amsterdam	Unknown	0.029394	0.802525	0.144538	0.023543	optimism
0.0	...	Switzerland	Zurich	CH	Winterthur	Male	0.524965	0.278535	0.082156	0.114344	joy
43.0	...	United States	Illinois	US-IL	Hampshire	Unisex	0.229822	0.456503	0.282737	0.030937	optimism
20.0	...	Germany	Bavaria	DE	Munich	Male	0.391068	0.317381	0.133483	0.158068	joy
38.0	...	United States	New York	US-NY	New York City	Unknown	0.112041	0.290365	0.551295	0.046299	anger

FIG. 21. pre-trained emotion recognition model)

3.5. *Data training using H20 Auto ML*

The Pandas library is used to load the data from the saved csv file. Unnecessary columns are eliminated to get the data ready for training. 'Unnamed: 0', 'StateCode','Hour','Day','joy probability','optimism probability','anger probability','sadness probability', 'IsReshare','text', 'Country', 'State', 'City' are the specific columns that are removed. By keeping only the relevant columns, this cleaning method produces a dataset that is better suited for training machine learning models in H20. After the dataset has been cleaned, it is prepared for additional analysis and model training.

The cleaned data was then converted into an H2OFrame. Predictors and response columns were set, and response columns were removed from the list of predictors. The data was split into training and validation sets with an 80:20 ratio.

H2O AutoML was used to train models for predicting "Likes" and "RetweetCount" separately. AutoML was configured to build a maximum of ten models, and the training process involved various algorithms, including XGBoost, GLM, GBM, and Stacked Ensembles. The performance of the models was monitored, and the best models were identified based on their RMSE (Root Mean Squared Error) values.

	Weekday	Lang	Reach	RetweetCount	Likes	Klout	Sentiment	Gender	best_emotion
0	Tuesday	en	991.0	1.0	0.0	42.0	1.00	Female	optimism
1	Tuesday	en	991.0	1.0	0.0	42.0	1.00	Female	anger
2	Monday	en	390.0	0.0	0.0	25.0	-2.00	Unisex	joy
3	Wednesday	es	402.0	0.0	0.0	32.0	0.17	Female	optimism
4	Thursday	es	403.0	0.0	0.0	32.0	0.20	Female	optimism
...	...	...	...	...	...	...	...	...	...
102017	Monday	en	5910.0	1.0	0.0	55.0	0.00	Unknown	optimism
102018	Wednesday	en	716.0	24.0	0.0	0.0	0.00	Male	joy
102019	Thursday	en	399.0	16.0	0.0	43.0	1.00	Unisex	optimism
102020	Wednesday	en	87.0	154.0	0.0	20.0	0.00	Male	joy
102021	Saturday	en	633.0	0.0	0.0	38.0	0.00	Unknown	anger

102022 rows × 9 columns

FIG. 22. pre-trained emotion recognition model

4. Results and Discussion

4.1. *Model Explainability*

4.1.1. Leaderboard for Likes

The leaderboard data for predicting likes is retrieved and the top entries are displayed. The leaderboard is printed, and the initial entries are reviewed. This analysis is conducted to identify and evaluate the top-performing entries based on likes, as shown in Fig. 23.

For ”Likes,” the best model was GBM with an RMSE of 1.7696. Variable importance was analysed, revealing that ”Reach” and ”Klout” were the most significant predictors.

Leaderboard for Likes:

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
GBM_2_AutoML_25_20240522_205219	1.76958	3.13142	0.136578	nan	3.13142
GBM_1_AutoML_25_20240522_205219	1.78519	3.18689	0.151287	nan	3.18689
GBM_3_AutoML_25_20240522_205219	1.79068	3.20653	0.139331	nan	3.20653
StackedEnsemble_BestOfFamily_1_AutoML_25_20240522_205219	1.79094	3.20745	0.147811	0.167233	3.20745
StackedEnsemble_AllModels_1_AutoML_25_20240522_205219	1.79352	3.2167	0.142643	0.16834	3.2167
GBM_4_AutoML_25_20240522_205219	1.79637	3.22696	0.135363	nan	3.22696
XGBoost_3_AutoML_25_20240522_205219	1.81142	3.28126	0.138993	nan	3.28126
DRF_1_AutoML_25_20240522_205219	1.81937	3.31011	0.133536	0.169287	3.31011
XGBoost_1_AutoML_25_20240522_205219	1.83743	3.37615	0.148607	nan	3.37615
XRT_1_AutoML_25_20240522_205219	1.85839	3.45363	0.138827	0.16404	3.45363

[10 rows x 6 columns]

FIG. 23. Leaderboard for Likes

4.1.2. **Leaderboard for Retweet**

The leaderboard for predicting retweets was retrieved and displayed the top entries. This action is conducted to analyse the top-performing entries based on retweet, as shown in Fig. 24.

For ”RetweetCount,” the best model was XGBoost with an RMSE of 206.5388. Variable importance was analysed, revealing that ”Reach” and ”Klout” were the most significant predictors.

Leaderboard for RetweetCount:

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
XGBoost_1_AutoML_26_20240522_205250	206.539	42658.3	13.9179	nan	42658.3
StackedEnsemble_BestOfFamily_1_AutoML_26_20240522_205250	206.655	42706.5	12.0695	nan	42706.5
StackedEnsemble_AllModels_1_AutoML_26_20240522_205250	206.659	42708	12.0652	nan	42708
GBM_4_AutoML_26_20240522_205250	206.667	42711.1	12.2476	nan	42711.1
XGBoost_3_AutoML_26_20240522_205250	206.736	42740	12.3814	nan	42740
GBM_3_AutoML_26_20240522_205250	206.756	42748	12.3161	nan	42748
GBM_2_AutoML_26_20240522_205250	206.767	42752.5	12.3722	nan	42752.5
XGBoost_2_AutoML_26_20240522_205250	206.769	42753.6	12.8299	nan	42753.6
GBM_1_AutoML_26_20240522_205250	206.781	42758.4	12.3521	nan	42758.4
XRT_1_AutoML_26_20240522_205250	206.789	42761.6	11.7831	1.55301	42761.6

[10 rows x 6 columns]

FIG. 24. Leaderboard for Retweet

The explain method is then called on these models with the validation dataset, rendering the visual explanations inline to the code. These explanations help interpret the models’ performance and predictive factors. This process is conducted to understand the most effective models for likes and retweet predictions.

4.2. *Plot Analysis for Likes*

4.2.1. **Residual Analysis**

The residual analysis plot for the model ”GBM2” was interpreted. The residuals scatter around zero, suggesting generally accurate predictions. However, increasing residual variance with higher fitted values indicates heteroscedasticity. Several outliers were observed, revealing instances of significant prediction errors. A slight downward trend in the mean residual line suggests minor systematic bias, with overpredictions at lower fitted values and underpredictions at higher ones.(Fig. 25)

4.2.2. **Learning Curve Plot**

The learning curve for ”GBM2” shows the mean squared error (mse) decreasing for both training and validation sets as the number of trees increases. Initially, the mSE drops sharply, indicating a significant improvement in model performance. After around 30 trees, the validation mse stabilises, suggesting optimal model complexity is reached. Beyond this point, further trees do not reduce the validation error, indicating potential overfitting as the training error continues to decrease. The selected number of trees, marked by a green line, is appropriate, balancing training and validation errors effectively.(Fig. 26)

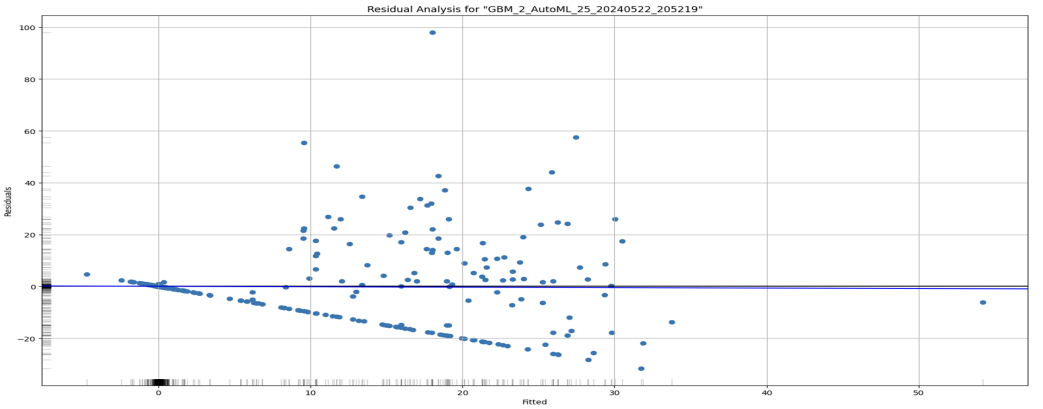

FIG. 25. Residual Analysis

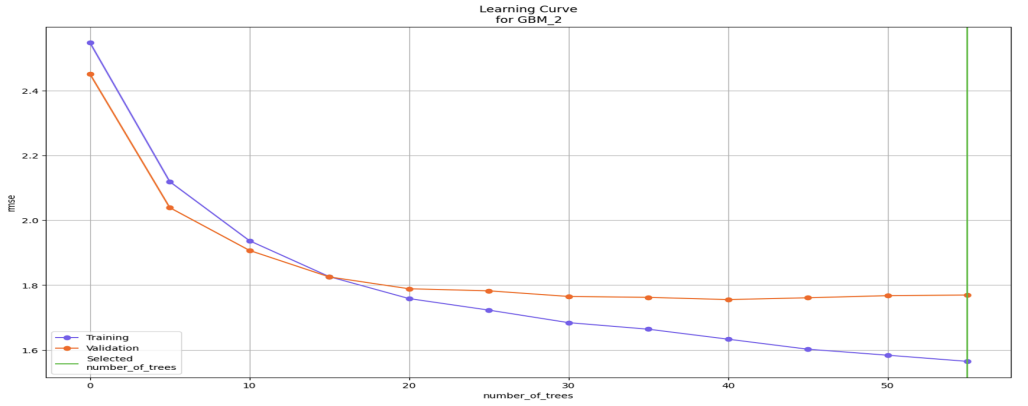

FIG. 26. Learning Curve Plot

4.2.3. **Variable Importance**

The variable importance plot for ”GBM2” was interpreted. The most significant variable is ”reach,” indicating it has the greatest influence on model predictions. ”Klout” follows with notable importance. ”Weekday,” ”Best emotion,” and ”Sentiment” have moderate importance, contributing less significantly. ”Gender” and ”Lang” show minimal importance, suggesting a limited impact on the model’s predictions. This analysis highlights the critical role of ”reach” and ”klout” in the model, guiding future focus on these variables for improving prediction accuracy.(Fig. 27)

4.2.4. **SHAP Summary**

The SHAP summary plot for ”GBM 2” was analysed. It shows the influence of each feature on the model’s predictions. ”Reach” has the highest impact, with significant SHAP values indicating strong influence. ”Klout” follows, also showing substantial impact. ”Gender,” ”Weekday,” ”Best emotion,” ”Sentiment,” and ”Lang” have smaller impacts. The colour gradient reveals how feature values affect predictions, with red for high and blue for low values. This analysis confirms ”reach” and ”klout” as

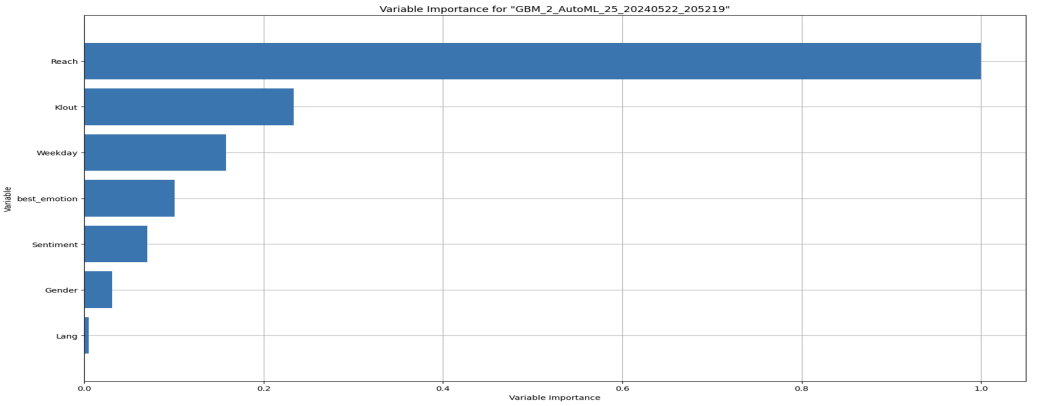

FIG. 27. Variable Importance

critical features, while other variables contribute less significantly. This insight aligns with previous variable importance-findings.(Fig. 28)

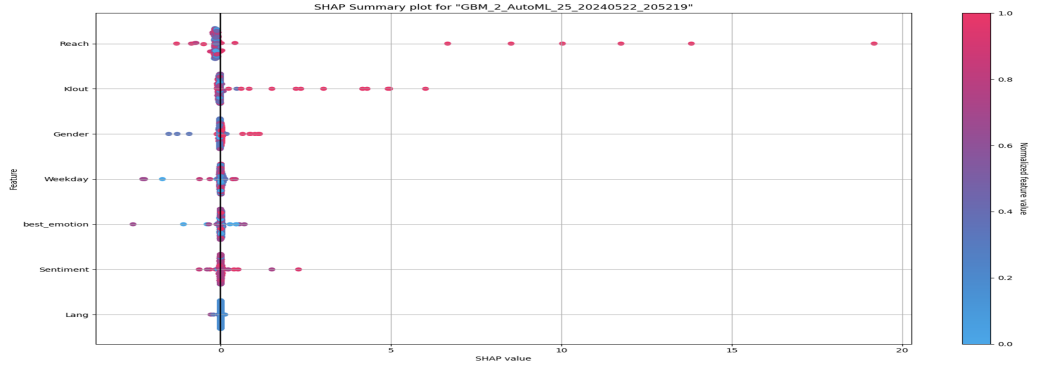

FIG. 28. SHAP Summary

4.2.5. **Partial Dependence Plots**

The ”GBM 2” partial dependence plots were examined in order to determine how different features affected the model’s predictions. Each plot provides insights into how changes in a specific feature impact the mean response, offering a detailed view of the model’s behaviour.

Partial Dependence Plot for Reach The plot for Reach shows a sharp increase in the mean response at low values, followed by a stabilisation at higher values. This indicates that Reach has a significant impact on the model’s predictions, particularly at lower levels. As Reach increases from 0 to about 0.05, the mean response rises steeply, suggesting that small increases in Reach have a substantial effect on the predictions. Beyond this point, it can be seen the impact plateaus, indicating that further increases in Reach do not significantly change the mean response. This pattern suggests that while

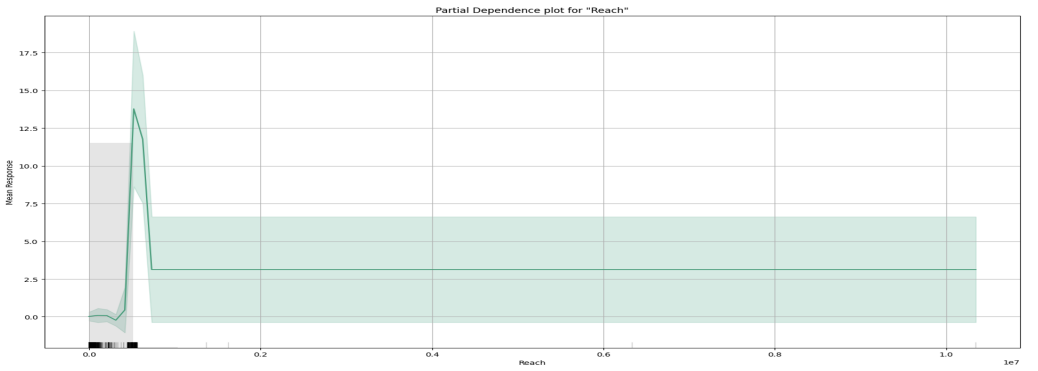

FIG. 29. Partial Dependence Plot for Reach

Reach is a crucial factor in driving predictions, its influence diminishes after a certain threshold.(Fig. 29)

Partial Dependence Plot for Klout The Klout plot shows a more complex relationship with the mean response. Initially, the response remains relatively stable, with minor fluctuations. However, there are noticeable spikes and drops at certain Klout values, indicating non-linear effects. A Klout score of 60 to 70 causes the mean response to spike sharply before declining. This implies that Klout has varying impacts on the model's predictions depending on its value. The variability and non-linearity imply that Klout interacts with other features in the model, leading to a more intricate influence on the predictions.(Fig. 30)

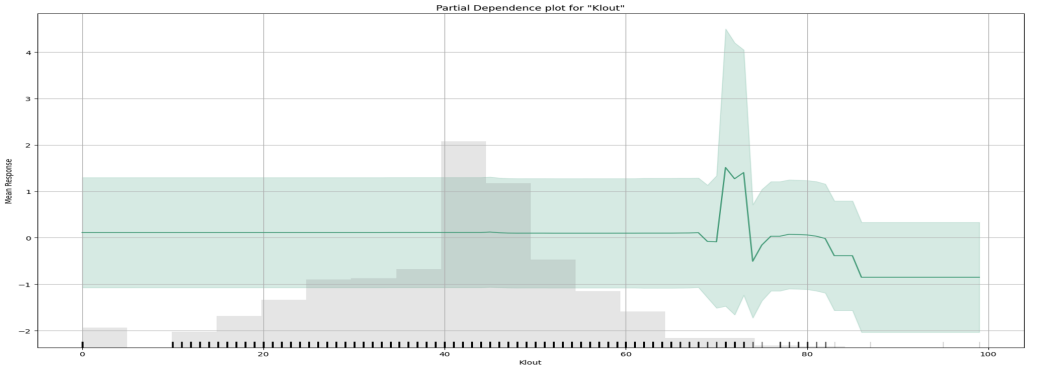

FIG. 30. Partial Dependence Plot for klout

Partial Dependence Plot for Weekday The weekday plot illustrates the mean response across different days of the week. The response varies slightly, indicating that the day of the week has a moderate impact on the model's predictions. For instance, weekends (Saturday and Sunday) show higher mean responses compared to weekdays. This pattern indicates that the model accounts for temporal

variations, with predictions slightly influenced by the day of the week. However, the differences are not drastic, indicating that the weekday is a less critical feature compared to reach and klout.(Fig. 31)

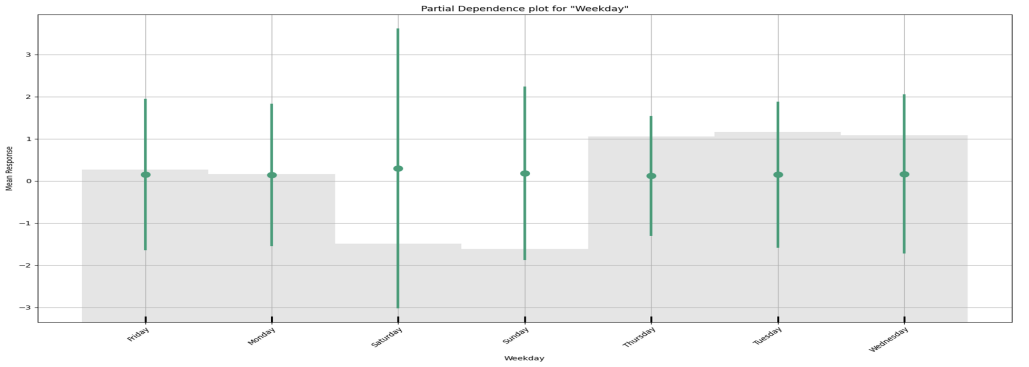

FIG. 31. Partial Dependence Plot for Weekday

Partial Dependence Plot for Best Emotion The plot for Best Emotion shows how different emotional states affect the mean response. The emotions are categorised as anger, joy, optimism, and sadness. The plot reveals that optimism and sadness are associated with higher mean responses, while anger and joy show lower mean responses. This indicates that the emotional tone of the data influences the model's predictions, with certain emotions having a more pronounced impact. The variation across different emotions suggests that emotional context is an important factor in the model's decision-making process.(Fig. 32)

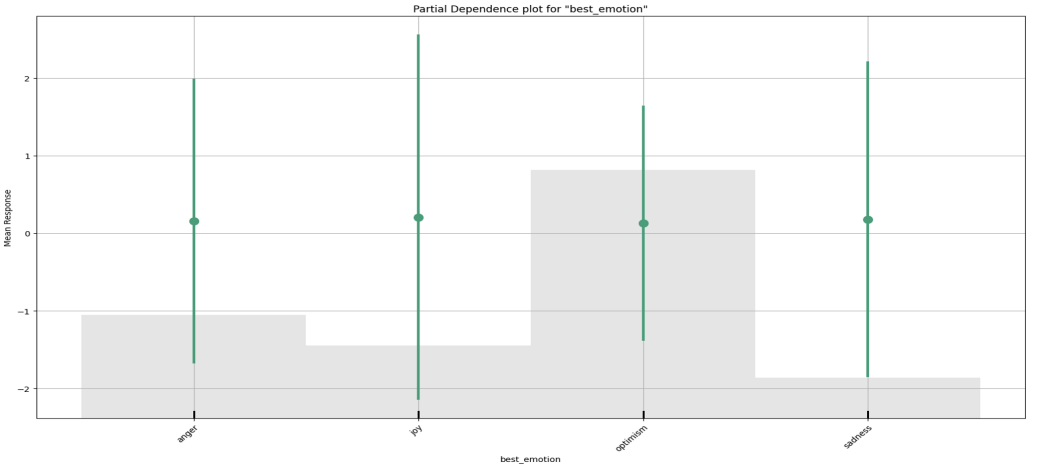

FIG. 32. Partial Dependence Plot for Best Emotion

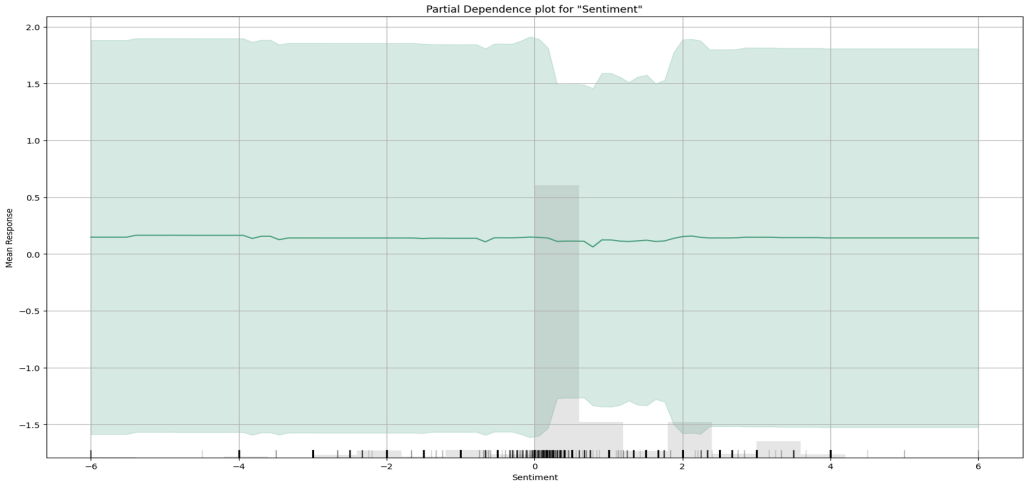

FIG. 33. Partial Dependence Plot for Sentiment

Partial Dependence Plot for Sentiment The sentiment plot displays the relationship between sentiment scores and the mean response. The plot shows that the response remains relatively flat across the sentiment spectrum, indicating that sentiment has a consistent effect on the predictions. There are minor fluctuations, but overall, the sentiment does not significantly alter the mean response. This suggests that while sentiment is considered in the model, its impact is more stable and less variable compared to other features like reach and klout.(Fig. 33)

4.2.6. **Individual Conditional expectations**

The model ”GBM” was inspected for Individual Conditional Expectation (ICE) plots to see how different percentiles of the data points for each feature affected the model’s predictions. These plots provide a detailed view of the variation in model response across different values of the features.

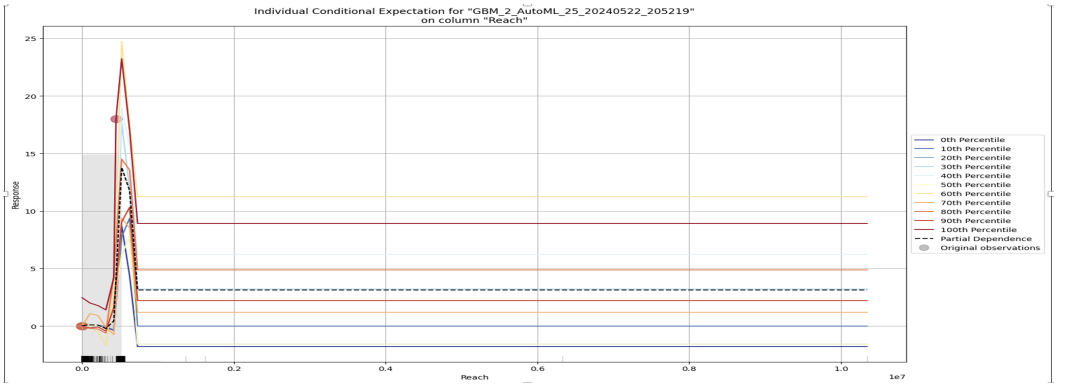

FIG. 34. ICE Plot for Reach

ICE Plot for Reach The ICE plot for Reach shows a distinct pattern where the response increases sharply at very low values of Reach and then stabilises. This pattern is consistent across different percentiles, suggesting that small increases in reach at lower levels significantly impact the predictions. Higher values stabilise the response, suggesting that additional increases in Reach do not significantly alter it. This behaviour suggests that reach is a critical factor for the model, particularly at lower values where it has a significant impact on the predictions.(Fig. 34)

ICE Plot for Klout The ICE plot for Klout exhibits a more complex relationship with the response. The plot shows that at low to moderate values of Klout, the response remains relatively stable. However, at higher values (around 60 to 80), there is a noticeable spike in the response, particularly for the higher percentiles (e.g., 90th and 100th percentiles). This spike suggests that Klout has a significant non-linear impact on the model's predictions, with higher values leading to sharp increases in response. This pattern indicates that Klout's influence on predictions is more pronounced at specific ranges, reflecting interactions with other features.(Fig. 35)

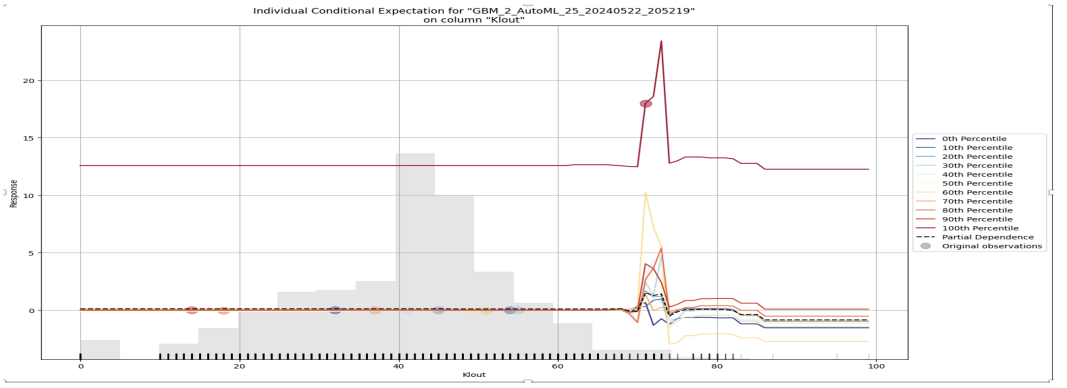

FIG. 35. SICE Plot for Klout

ICE Plot for Weekday The ICE plot for Weekday indicates that the model's response varies across different days of the week. The response is higher on weekends (Saturday and Sunday), suggesting that these days have a positive impact on the predictions. The response for weekdays is generally lower, indicating that the model accounts for temporal variations. The consistent pattern across different percentiles suggests that the impact of the day of the week is uniformly significant across the data. This behaviour highlights the importance of temporal factors in the model's predictions.(Fig. 36)

ICE Plot for Best Emotion The ICE plot for Best Emotion shows the impact of different emotional states on the response. Emotions such as optimism and sadness result in higher responses, particularly for the higher percentiles. Anger and joy, on the other hand, lead to lower responses. This pattern suggests that the model differentiates between various emotional states, with certain emotions like optimism and sadness having a more positive influence on predictions. The consistency across percentiles indicates that these emotional impacts are robust across different data points. (Fig. 37)

ICE Plot for Sentiment The ICE plot for Sentiment reveals that the response remains relatively stable across different sentiment values, with minor fluctuations. There is a noticeable dip in response for the higher percentiles at certain sentiment values, but, the sentiment does not significantly alter the

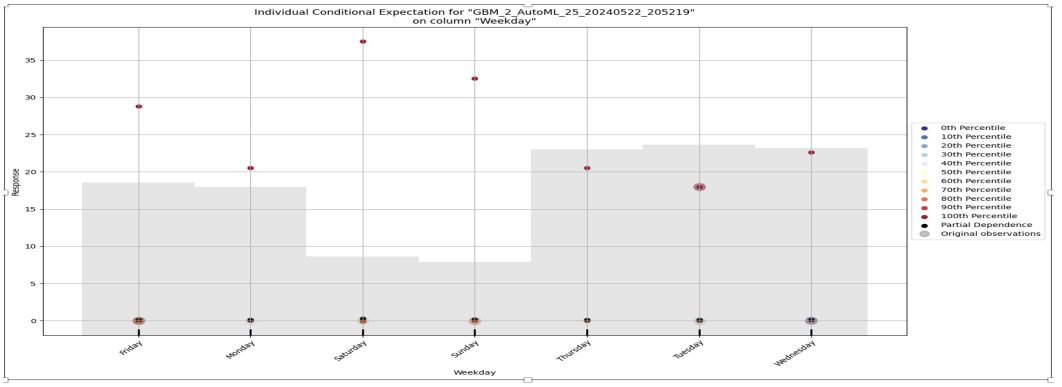

FIG. 36. ICE Plot for Weekday

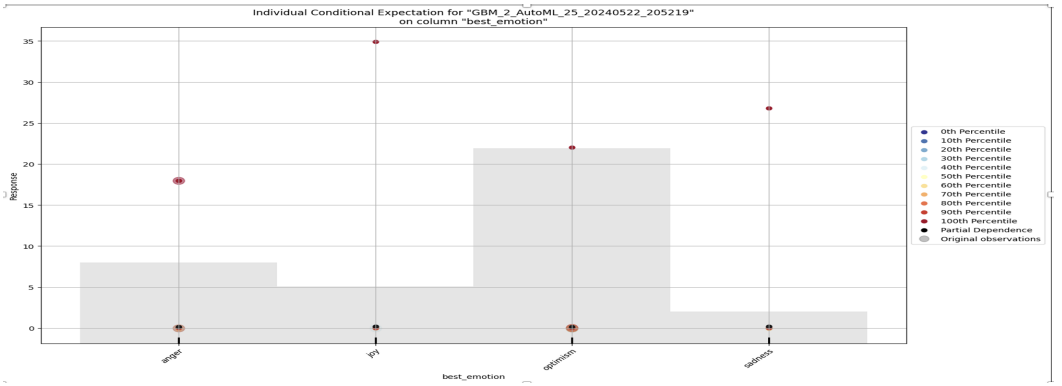

FIG. 37. ICE Plot for Best Emotion

response. This pattern suggests that while sentiment is considered in the model, its impact is relatively uniform and less variable compared to other features like Reach and Klout. The stability across different percentiles indicates that sentiment's effect on predictions is consistent.(Fig. 38)

4.3. *Retweets Analysis Plot*

4.3.1. **Residual Analysis**

The residual plot for the XGBoost model indicates a generally good fit for the majority of data points, with most residuals clustering close to the zero line. However, there are significant outliers visible, particularly one residual above 25,000 and another near 15,000. These outliers suggest that for some data points, the model's predictions are far from the actual values. This could imply the presence of anomalies or a need for further model refinement. The horizontal spread of residuals also shows that prediction errors do not vary much with fitted values, which suggests homoscedasticity.(Fig. 39)

4.3.2. **Learning Curve Plot**

The learning curve for the XGBoost model shows that the training error (mse) decreases rapidly with the addition of trees, stabilising around 25. However, the validation error remains consistently high, at

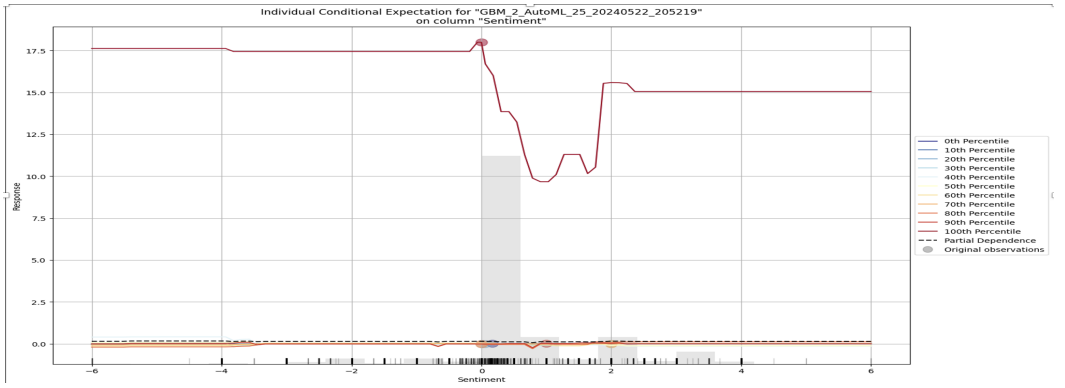

FIG. 38. ICE Plot for Sentiment

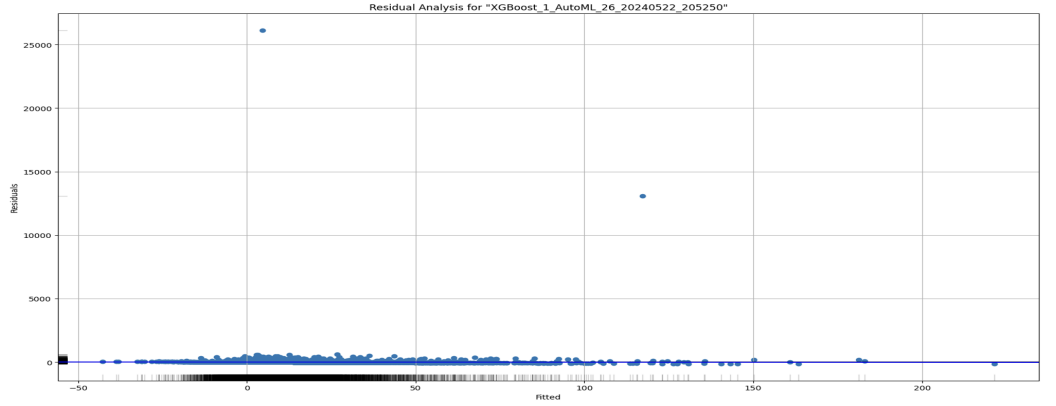

FIG. 39. Residual Analysis

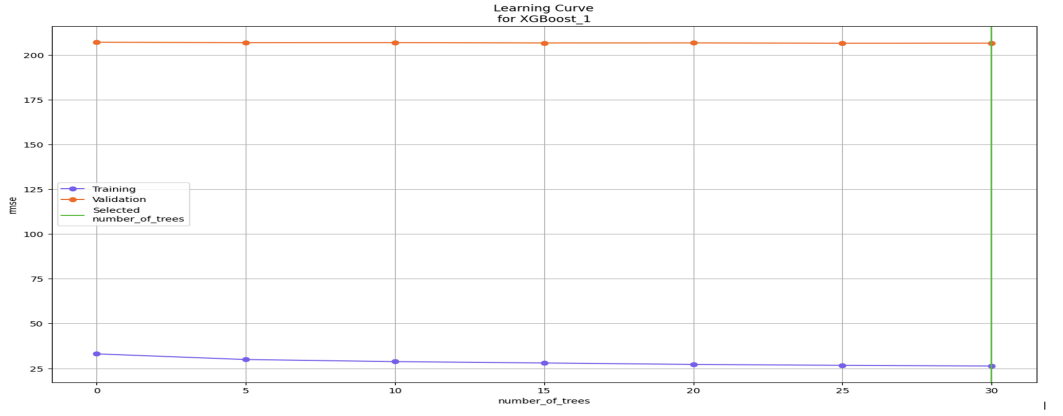

FIG. 40. Learning Curve Plot

around 200, indicating overfitting. Despite increasing the number of trees, the model fails to generalise well to the validation set. This discrepancy suggests that while the model performs well on training data, it does not translate to unseen data, highlighting the need for better regularisation or more representative training data to improve validation performance and prevent overfitting.(Fig. 40)

4.3.3. **Variable Importance**

The variable importance plot for the XGBoost model reveals that ”reach” and ”klout” are the most influential features, significantly contributing to the model’s predictions. ”Sentiment” also plays a notable role, but to a lesser extent. Other variables like ”best emotion, optimism,” specific weekdays, and ”gender, male” show minor importance. This indicates that the model relies heavily on reach and influence metrics to make predictions, while emotional tones and temporal factors have a limited impact. This insight can guide further feature engineering and model refinement efforts, focusing on the most impactful variables for better predictive performance. (Fig. 41)

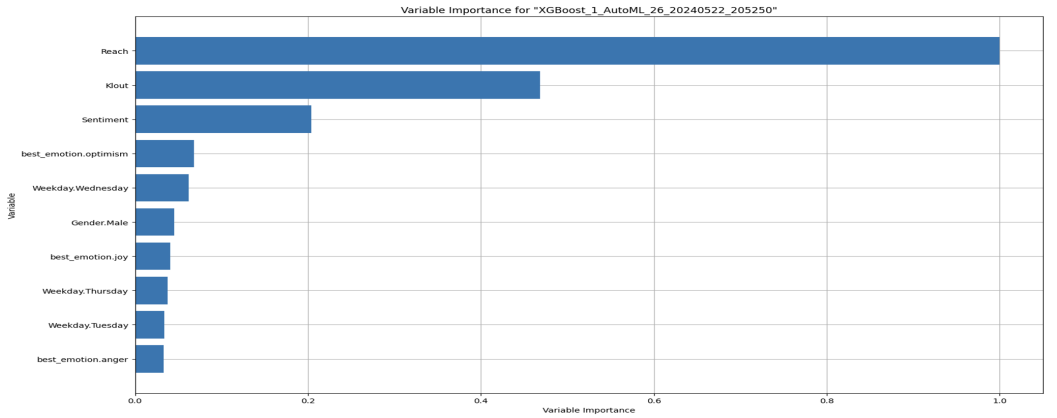

FIG. 41. Variable Importance

4.3.4. **SHAP Summary**

The SHAP summary plot for the XGBoost model illustrates the impact of various features on the model’s predictions. ”Klout” and ”reach” are identified as the most influential features, with high SHAP values indicating a strong effect on the output. Positive SHAP values (in red) correspond to high feature values, significantly influencing predictions, while negative SHAP values (in blue) indicate low feature values. ”Best emotion,” ”Weekday,” and ”Sentiment” also show notable impacts but are less significant. The plot highlights the importance of these features, suggesting that influence metrics and sentiment analysis are crucial for accurate predictions in this model.(Fig. 42)

4.3.5. **Partial Dependence Plots**

Partial Dependence Plot for ”Reach” The partial dependence plot for ”Reach” demonstrates how changes in this feature affect the model’s predictions. Initially, there is a sharp increase in the predicted response when ”Reach” is very low, indicating high sensitivity in this range. As ”Reach” values increase beyond this initial spike, the predicted response stabilises, showing a consistent effect with less variation. This suggests that while the initial reach significantly impacts predictions, higher values

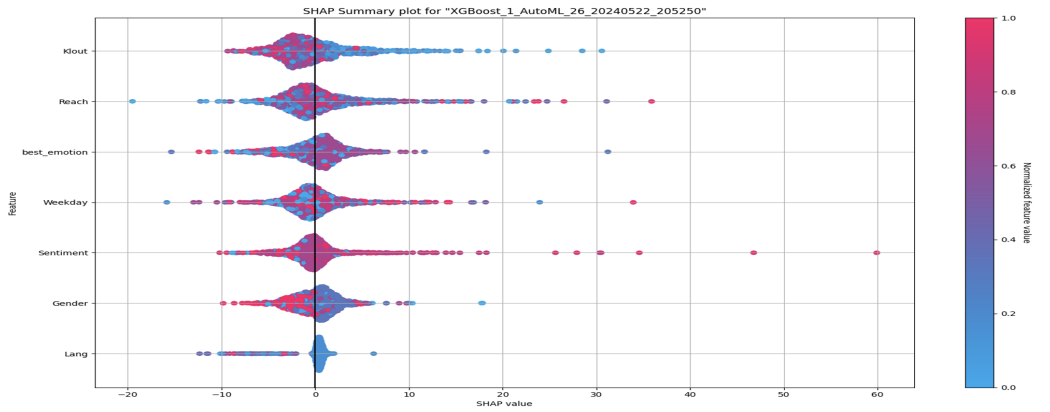

FIG. 42. SHAP Summary

have a more uniform effect. The shaded area indicates the confidence interval, showing high variability at low reach values, which stabilises as reach increases.(Fig. 43)

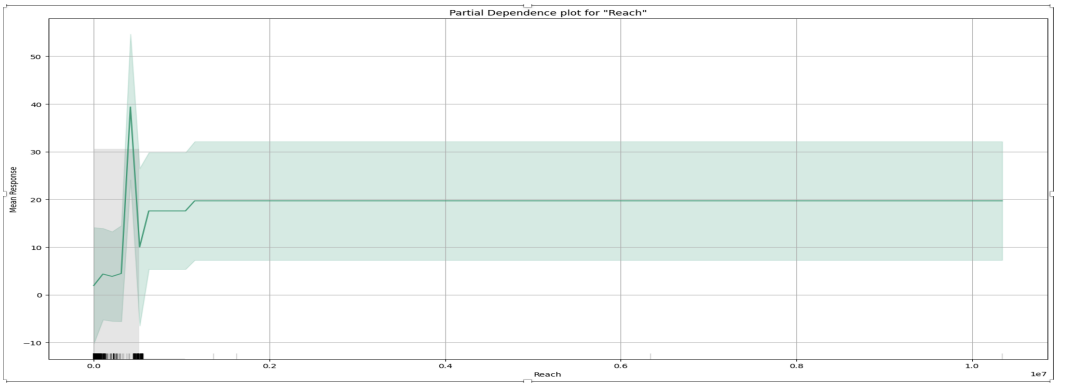

FIG. 43. Partial Dependence Plot for ”Reach”

Partial Dependence Plot for ”Klout” The plot for ”Klout” shows a complex relationship with the predicted response. Initially, as ”Klout” increases, the predicted response decreases, suggesting that higher ”Klout” scores are associated with lower predictions. This relationship continues with some fluctuations, indicating that ”Klout” influences predictions in a non-linear fashion. The confidence intervals are wider at lower ”Klout” values, indicating greater uncertainty in predictions in this range. As ”Klout” increases, the confidence interval narrows, showing more stable predictions. This plot suggests that ”Klout” has a significant but complex impact on the model’s predictions, with varying degrees of influence across different values.(Fig. 44)

Partial Dependence Plot for ”Weekday” The ”Weekday” plot shows the average predicted response for each day of the week. It reveals that certain days, such as Friday, Monday, and Sunday, have higher average predicted responses compared to others. The vertical lines indicate the range of responses for

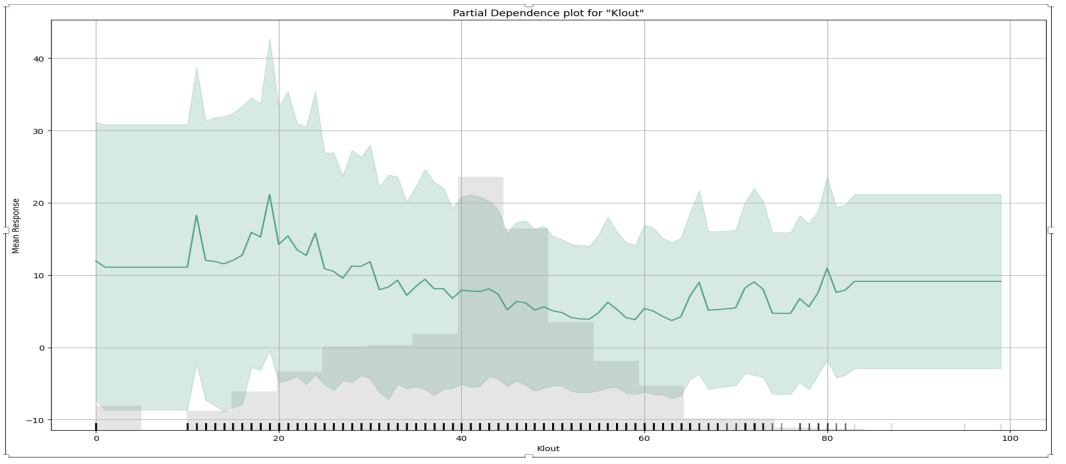

FIG. 44. Partial Dependence Plot for ”Klout”

each day, with some days showing greater variability. This suggests that the day of the week influences the model’s predictions, potentially due to varying patterns in data collected on different days. The effect of weekdays appears to be significant, with specific days standing out as more impactful on the model’s output.(Fig. 45)

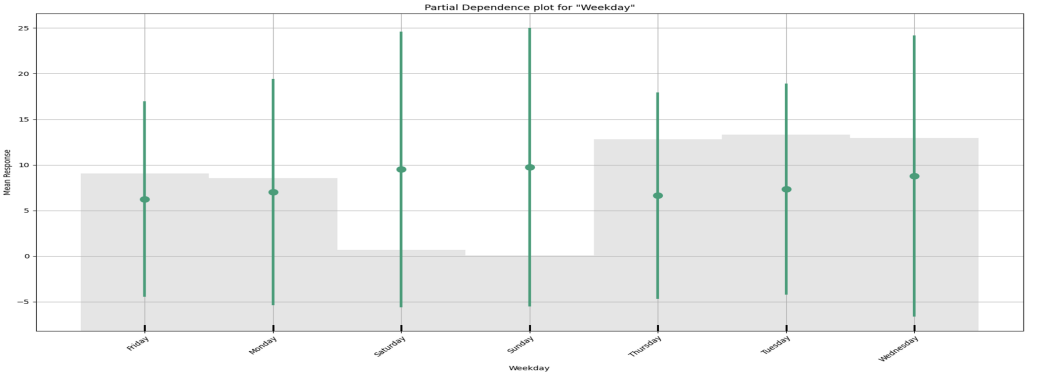

FIG. 45. Partial Dependence Plot for ”Weekday”

Partial Dependence Plot for ”Sentiment” The plot for ”Sentiment” shows that the predicted response remains relatively stable across a range of sentiment values, with a significant spike at a sentiment value of around 2. This indicates that positive sentiments around this value have a notable impact on predictions, while other values, both positive and negative, show less influence. The confidence intervals are wider around the spike, indicating more variability and less certainty in predictions at this point. This suggests that sentiment has a specific, targeted impact on the model’s predictions, particularly for strongly positive sentiments. (Fig. 46)

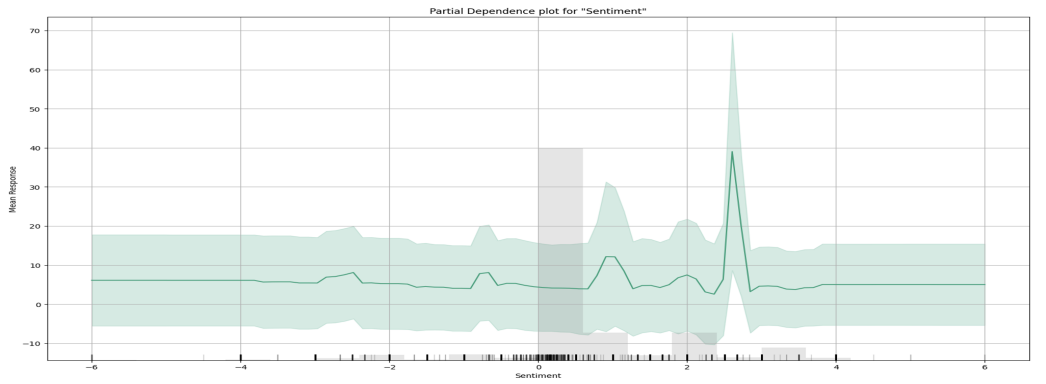

FIG. 46. Partial Dependence Plot for ”Sentiment”

Partial Dependence Plot for ”best emotion” The ”best emotion” plot indicates the effect of different emotions on the predicted response. Optimism appears to have the highest average predicted response, suggesting that optimistic emotions significantly boost predictions. Other emotions, such as joy and sadness, have varying impacts, with sadness showing a lower predicted response. The vertical lines indicate the range of responses, with optimism showing the highest variability. This plot highlights that emotional context, particularly optimism, plays a crucial role in influencing the model’s predictions, with different emotions contributing to different extents. (Fig. 47)

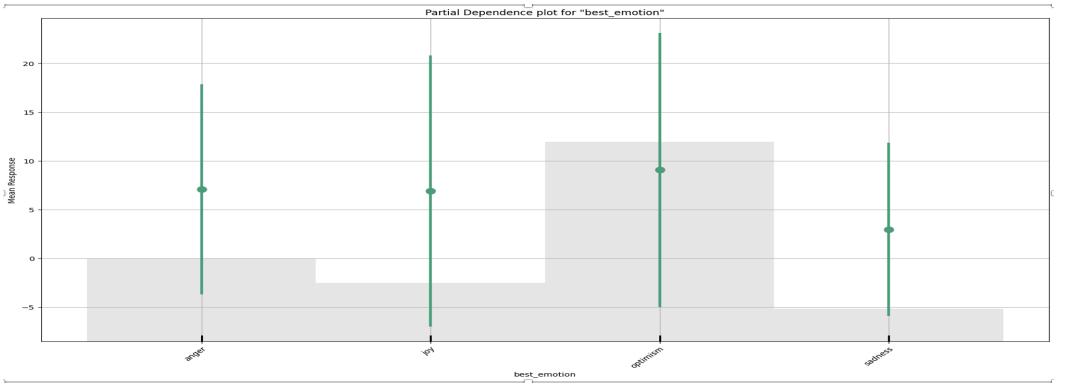

FIG. 47. Partial Dependence Plot for ”best emotion”

4.3.6. **Individual Conditional expectations**

Individual Conditional Expectation (ICE) Plot for ”Reach” The ICE plot for ”Reach” reveals that at very low values of reach, the responses vary significantly among different percentiles. As reach increases slightly, there is a sharp spike in the response, particularly for higher percentiles, indicating high sensitivity to small changes in reach at low levels. Beyond this spike, the response becomes relatively stable, with different percentiles showing parallel but distinct levels. This suggests that

”Reach” has a crucial impact on the model’s predictions at low values, but its influence stabilises as reach increases.(Fig. 48)

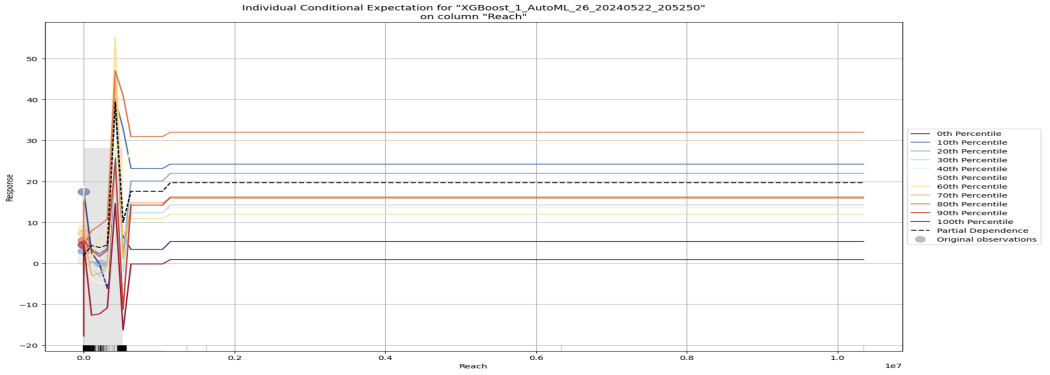

FIG. 48. Individual Conditional Expectation (ICE) Plot for ”Reach”

Individual Conditional Expectation (ICE) Plot for ”Klout” The ICE plot for ”Klout” shows a complex and highly variable relationship between ”Klout” and the model’s response across different percentiles. There are multiple fluctuations and spikes, indicating that the model’s predictions are sensitive to changes in ”Klout” at various levels. The wide range of responses across percentiles suggests that ”Klout” interacts with other features in a non-linear manner, leading to diverse prediction patterns. This complexity implies that ”Klout” has a significant but unpredictable effect on the model’s outcomes, highlighting the need for further analysis to understand its interactions. (Fig. 49)

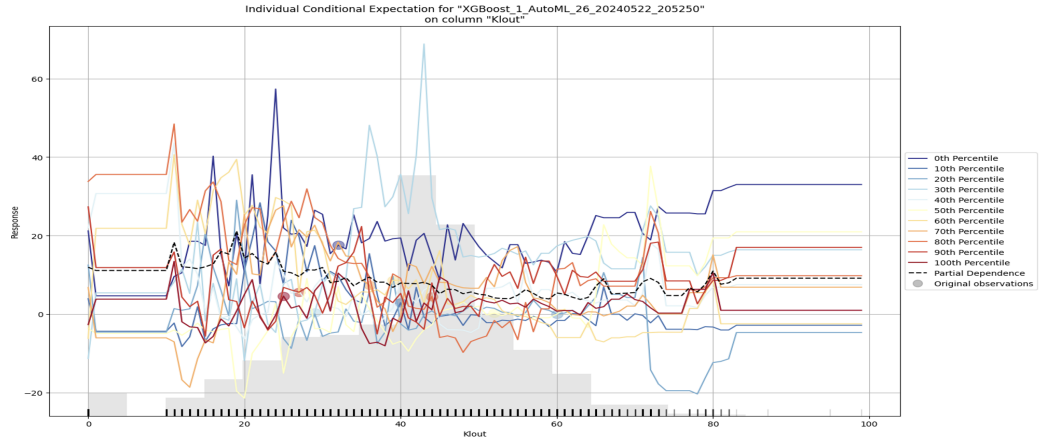

FIG. 49. Individual Conditional Expectation (ICE) Plot for ”Klout”

Individual Conditional Expectation (ICE) Plot for "Weekday" The ICE plot for "Weekday" shows how the model's response varies across different days of the week. Each point represents the average response for a specific percentile on a particular weekday. It is observed that certain days, such as Friday, Sunday, and Wednesday, have higher response levels, especially at the higher percentiles. The plot indicates that the day of the week significantly influences the model's predictions, with specific days showing distinct impacts. This suggests that temporal patterns play a crucial role in the model's predictive behaviour. (Fig. 50)

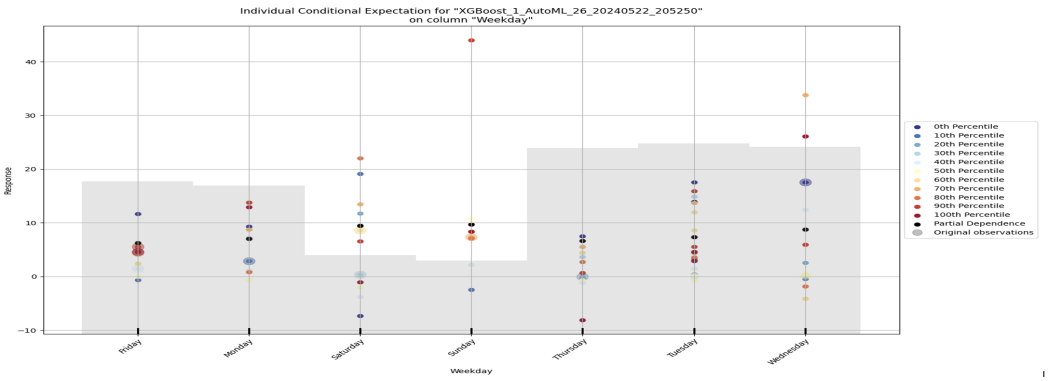

FIG. 50. Individual Conditional Expectation (ICE) Plot for "Weekday"

Individual Conditional Expectation (ICE) Plot for "Sentiment" The ICE plot for "Sentiment" indicates a non-linear relationship between sentiment values and the model's predictions. There are notable fluctuations and spikes at different sentiment levels, particularly around positive sentiment values. Different percentiles exhibit varying response patterns, with higher percentiles showing more pronounced changes. This suggests that sentiment has a complex impact on the model's predictions, with positive sentiments having a significant effect. The variability across percentiles highlights the importance of sentiment in the model's decision-making process.(Fig. 51)

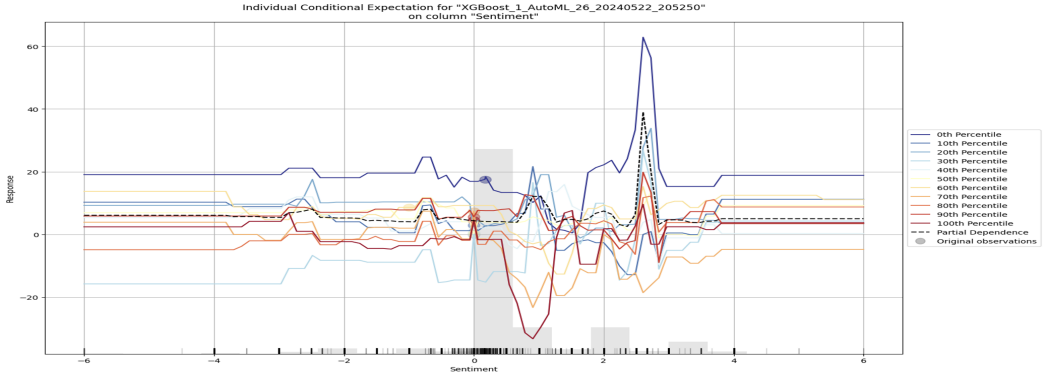

FIG. 51. Individual Conditional Expectation (ICE) Plot for "Sentiment"

Individual Conditional Expectation (ICE) Plot for "best emotion" The ICE plot for "best emotion" demonstrates how different emotional states affect the model's response across percentiles. Optimism shows the highest response levels, particularly at higher percentiles, indicating its strong influence on the model's predictions. Other emotions, like joy and anger, have varied impacts, with optimism consistently leading to higher predictions. The variability across percentiles suggests that the emotional context plays a significant role in shaping the model's outputs. This highlights the importance of incorporating emotional features for more accurate predictions.(Fig. 52)

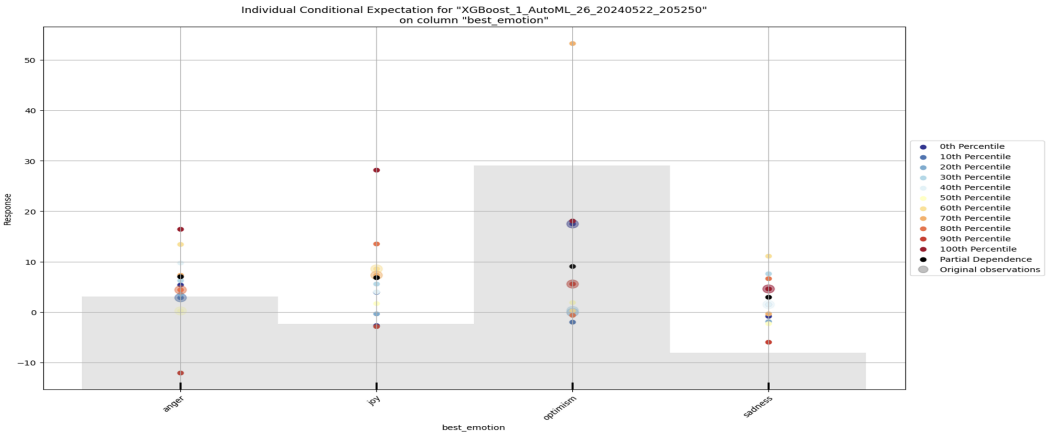

FIG. 52. Individual Conditional Expectation (ICE) Plot for "best emotion"

CLICK HERE TO ACCESS THIS ENTIRE CODE ON GITHUB

5. Conclusion and Recommendations

5.1. *Conclusion*

In this research Twitter dataset was loded, cleaned, preprocessed and then analyzed using 'Cardiffnlp/twitter-roberta-base-emotion' transformer model and best probable emotions were captured. This data is then appeneded to the original dataset and was used to train H20 AutoML, which is an open source, predictive analytics platform for building machine learning models. Once the model was trained, the best model was selected and then used the H20's model explainability feature to plot various useful plots such as Residual Analysis, Learning Curve Plot, Variable Importance, SHAP Summary, Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots to test on validation data to learn the impact of independent variables on 'Likes' and 'Retweets' which indicate the degree of social media engagement (In this case, Twitter) These plots provided valuable insights on the importance of various factors on 'Likes'and 'Retweets'. key take outs from this entire analysis are as follows. Reach, Klout and Weekday in which a tweet is posted impacts a posts probability getting higher likes. Although it is known that reach and klout impacts the probability of getting likes, the other features such as weekday, emotion and sentiment plays a crucial part in increasing the chances of impacting likes. Reach, Klout and Sentiment are top three factors impacting the likelihood of a text getting retweeted. Apart from Reach and Klout, factors such as sentiment,

emotions such as optimism and anger plays a significant role in impacting retweets.Optimistic and Angry tweets are top 2 emotions to get high user engagement on a tweet. Gender and Language does not have major impact on user engagement.PDP and ICE plots give deeper understanding of the impact of each variable on user engagement. This analysis is restricted to train 10 models only and is time restricted to strike a balance between accuracy and computational costs, keeping in mind the timeframe and scope of this project.

In future, there is a scope for performing sentiment analysis on the dataset using more advanced and efficient multilingual machine learning models for better emotion prediction. This is because in reality, a tweet might have more than one emotion, which when captured effectively can train machine learning model to predict an emotion more accurately. This data is mainly technical which might limit a wide range of emotions displayed in a normal tweet. In future, the same analysis can be made on a more diverse dataset to better capture emotions. With the advancements in Large Language Models, there is a possibility of greater feature extraction and more efficient models, which uses a fraction of present computation to predict better.

5.2. *Recommendations*

Based on the insights gained from this study, numerous guidelines are made for improving and, in addition, applying predictive models for social media analysis:

- **Integration into Chatbot Applications:** Insights from sentiment evaluation and interaction prediction can notably improve the performance of chatbot packages. Chatbots can tailor consumer interactions based on emotional sensitivity and engagement choices by combining these insights.
- **Continuous Model Refinement:** Continuous model refinement calls for inputting new statistics and enhancing predictive models. This ensures long-term relevance and accuracy. This is further allowing the organisations to better adapt to traits and personal conduct.
- **Collaboration with Social Media Platforms:** Partnering with social media structures allows groups to connect with databases and APIs in real time. This also enriches predictive models with platform-specific insights. These collaborations could make them thoroughly understand the behaviour of the users and developments.
- **Ethical Considerations:** It is essential to remember ethical implications when analysing user data and the use of predictive models for social media analytics. Organisations should prioritise the privacy of their users and data protection. They should follow applicable laws and suggestions and make certain that their data-handling practices are transparent and accountable.

6. Acknowledgment

I would love to express my sincere gratitude to all who contributed to the completion of this research project. I would like to thank my mentor for their treasured guidance and courses at some point inside the research system. I am also grateful to the contributors who helped obtain the information used in the study. In addition, I respect the insightful comments and encouragement from colleagues and friends. Their support and encouragement have been very important to success of the entirety of this project.

REFERENCES

1. J. An and W. M. N. W. Zainon. Integrating color cues to improve multimodal sentiment analysis in social media. *Engineering Applications of Artificial Intelligence*, 126:106874, 2023.
2. L. Bryan-Smith, J. Godsall, F. George, K. Egode, N. Dethlefs, and D. Parsons. Real-time social media sentiment analysis for rapid impact assessment of floods. *Computers & Geosciences*, 178:105405, 2023.
3. L. Chen, C. Li, and T. Tang. The impact of working from home on urban commuting in china: A comprehensive analysis using social media and recruitment website data. *Cities*, 148:104868, 2024.
4. S. P. Eslami, M. Ghasemaghaei, and K. Hassanein. Understanding consumer engagement in social media: The role of product lifecycle. *Decision Support Systems*, 162:113707, 2022.
5. H. T. Halawani, A. M. Mashraqi, S. K. Badr, and S. Alkhalaf. Automated sentiment analysis in social media using harris hawks optimisation and deep learning techniques. *Alexandria Engineering Journal*, 80:433–443, 2023.
6. H. Huang, R. Long, H. Chen, K. Sun, Q. Sun, and Q. Li. Examining public attitudes and perceptions of waste sorting in china through an urban heterogeneity lens: A social media analysis. *Resources, Conservation and Recycling*, 199:107233, 2023.
7. M. A. Khan and M. AlGhamdi. A customized deep learning-based framework for classification and analysis of social media posts to enhance the hajj and umrah services. *Expert Systems with Applications*, 238:122204, 2024.
8. Y. Li, J. Chan, G. Peko, and D. Sundaram. Mixed emotion extraction analysis and visualisation of social media text. *Data & Knowledge Engineering*, 148:102220, 2023.
9. E. P. Meshram, R. Bhambulkar, P. Pokale, K. Kharbikar, and A. Awachat. Automatic detection of fake profile using machine learning on instagram. *International Journal of Scientific Research in Science and Technology*, 8(1):117–127, 2021.
10. P. Pandey and M. M. Pandey. *Research methodology tools and techniques*. Bridge Center, 2021.
11. J.-H. Park and H.-Y. Kwon. Cyberattack detection model using community detection and text analysis on social media. *ICT Express*, 8(4):499–506, 2022.
12. V. P. Rodrigues and M. A. L. Caetano. The impacts of political activity on fires and deforestation in the brazilian amazon rainforest: An analysis of social media and satellite data. *Heliyon*, 9(12), 2023.
13. W. Wu, L. Huang, and F. Yang. Social anxiety and problematic social media use: A systematic review and meta-analysis. *Addictive Behaviors*, page 107995, 2024.
14. W. Zha, Q. Ye, J. Li, and K. Ozbay. A social media data-driven analysis for transport policy response to the covid-19 pandemic outbreak in wuhan, china. *Transportation Research Part A: Policy and Practice*, 172: 103669, 2023.