NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Graduate School of Business

Charalampos Tsiolis

Predictive Models for Content Virality in Online Forums
MASTER'S THESIS

Field of study: Business Informatics

Degree programme: Business Analytics and Big Data Systems

Reviewer
Doctor of Science
Natalia A. Zhukova

Academic Supervisor
Associate Professor, PhD
Ignatov I. Dmitry

Moscow 2024

Abstract

The main objective of this Master's thesis is to develop predictive models for virality content analysis in online forums, particularly Reddit. Through analyzing 108,036 posts from many different subreddits the votes, comments, and posting times, the research aims to determine those features that can predict content with viral potential using ML-based models and Time Series Analysis.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Context and Background

In the digital era, a content virality construct has become a piece of the main plan for marketing strategies, specifically on social media, where the ways that brands and content creators involve with the audience differs from each other (Fan, 2023; Le Meunier-FitzHugh, 2021; Rosário & Dias, 2023). The phenomenon of the virality means that the content experiences a rapid diffusion through the networks, and therefore achieve a great deal of attention and engagement. On the other hand, the evaluation of the virality is mostly done through the likes, shares or comments. This phenomenon is to reach a wide public audience and to affect them in a way that really results in improvement in behaviors and perceptions (Dašić et al., 2023; Fan, 2023; Le Meunier-FitzHugh, 2021). Moreover, it affects the culture completely.

Exploiting the underlying factors that drive content to go viral requires the analysis of the various elements of the social media dynamics. It is essential to point out that the factors which determine content's importance, for example, tone and theme, as well as those which determine creator's influence, for example, credibility and popularity, are equal when it comes to establishing the level of content's appeal and shareability (Han et al., 2020). Such interactions suggest that it is the content that matters as well as its owners and that their reputation and credibility are important too.

In addition to this, the architecture of a social network is a determinative factor for the occurrence of the virality of the content (Han et al., 2020; Joseph et al., 2021). The scale of the network, the communities and the likelihood of the existence of the highly interlinked circles can to a very great extent determine how information is dispensed through social media networks (Joseph et al.,

2021). Here it can be seen that the network topology of the digital environment is also the factor here, some of the network configurations can decrease the rate of information spread while some can accelerate it.

Emotional engagement is another critical driver of content virality. They are more likely to share the content that evoke strong positive feelings such as happiness, inspiration or excitement as these emotions create engagement and the probability that content will be passed on grows. It is this emotional connection that accentuates the fact that creating messages has to be in such a manner that they consume a personal level with viewers and also are shareable (Tellis et al., 2019).

To add to that, the presentation and environment in which the content is displayed also play an important role in how the consumer engages with it and in the eventual spread of the content. For instance, different platforms may prioritise the content of different type, and the same content may vary in how it performs if presented in a text format or an image or a video format (Shahbaznezhad et al., 2021). Such kind of variety emphasizes on the fact that the content strategies must always be customized to take into account both the medium and the message and thus optimize the engagement.

As the last point, the virality of the content is to be looked at by the way the content is cycle within the digital world. The content, alongside with it is discovery process through which users are consuming it by likes, shares and comments, bears upon the content's prospects for spreading and having an effect (Yang et al., 2023). Such interactive activities have the potential to shed light on what kind of content get the public to retain and share endlessly on various platforms. Therefore, the market research will be very helpful for refining future content creations.

1.2 Research Problem

In the present online world, social platforms such as Reddit is taking the lead as a central place of lightning fast dissemination of latest information as well as an important platform for engagement and conversation (Chi & Chen, 2023; Proferes et al., 2021). These platforms, besides the function of users discussing and sharing their points of view, have turned into strong amplifiers of content, making it go viral.

Despite its intriguing nature as it is the core factor shaping public discourse, influencing cultural tendencies, and, eventually, impacting big decisions the phenomenon of content virality is a relatively complex one (Elmas et al., 2023; Han et al., 2020; Joseph et al., 2021). Thus, the reasons behind the posts popularity on Reddit provide the necessary understanding of online information spread mechanisms and also the essence of major trends shaping the social media world.

Research problem

The main focus of the research paper will be to organize the different aspects that drive a post to become viral on Reddit (Chi & Chen, 2023; Proferes et al., 2021). The analysis carries out tasks to observe themes like the type of content, when is it posted, the engagement level of the communities, and the overall sentiment of the content. This study has high significance since it can offer insights into how digital content may interact with human behaviors and what people value.

Through highlighting the most important factors of virality, this research would give insights for seeming content creators and digital marketers of successfully involving their audiences. In addition, it could help developers of platforms in creation of algorithms that are truly reliable in what they curate and

recommend so as to improve the consumers' experience in the platform and prevent bad engagement (Arjona-Martín et al., 2020; Shahbaznezhad et al., 2021; Tellis et al., 2019). The core objective of this study is to cover the existing educational gap by offering an in-depth examination of dynamics of virality, which makes it a very useful tool for both university researchers and media practitioners.

1.3 Research Objectives

The purpose of this study is to create prediction models that can significantly improve the likelihood of content going viral analyzing Reddit data. Analysis of different factors such as votes number (upvotes), comments number, and posting time helps to detect the route and spread of posts. By a systematic study of these determinants at the subreddit level the goal is to clarify the patterns that work behind content virality and, in a consequence of this, to develop such models of viral potential of a post that can be considered robust.

Today this project will be achieved through the avanced of machine learning techniques that will be used to process and analyze 108,036 of Reddit posts from 2006 to the end of 2023. Among the steps will be data preprocessing, extraction of the features using TF-IDF model and the application of algorithms, like Random Forest, LightGBM and Gradient Boosting (Mohammed et al., 2016; Sarker, 2021; Yeturu, 2020). The objective is to also better these models in order to make them more accurate and precise on the predicted virality. As mentioned before, the research findings on this topic will help to develop the academic field of digital communication and will as well supply practical strategies, on how to increase engagement and reach among the content creators and social media strategists, to those using the online platform.

1.4 Research Questions

This thesis aims to address several key questions to better understand and predict content virality on platforms like Reddit (Proferes et al., 2021):

a. **Which indicators would be considered the most reliable to predict virality for a post?** *This question will undertake an investigation to find out the scores of the post in terms of likes, comments, and the time posts are made.*

b. **Are machine learning models capable of real-time, pre-virality forecasting by these indicators? If so, what the level of accuracy can be achieved?** *This question probes the expertise of the machine learning algorithms in locating and forecasting the most viral content. The purpose is to check the accuracy and reliability of these models in real-time.*

c. **To what extent, do the user's feelings about the content affect the level of popularity of the content in the Internet's forums?** *The purpose of this question is to find out whether the sentiment of the users are able to influence the impact and engagement of posts at all. Through this process it becomes clear that whether the sentiments are positive, negative or there are no sentiments at all (neutral) that the way content is perceived and shared can be very influential.*

d. **How is it possible to create implementable strategies from the predictive models so as to enable the content developers and social media strategists to make their content trend?** *By asking this question,*

there will be a possibility of transforming the analytical results of the predictive models into the steps which may be implemented by the creators to maximise the viralness of their content via social media platforms hence improving their engagement and influence.

e. **Does the number of comments play a decisive role in making a post viral?** Answering this question will refer to whether the increased number of comments on a post also implies that the post is viral.

1.5 Structure

The thesis is structured in the form of several chapters, each chapters cover on a different aspect of the research which in turn makes it a comprehensive study of the factors which are affecting the content virality of Reddit. Every part of the proposal is structured to be followed by the previous one, making a clear, logical flow from the basic principles to the real-life applications and the conclusion.

Chapter 2: Literature Review

This chapter gives a comprehensive overview of the studies that have been done on content virality, and the use of machine learning techniques and time series analysis in social media platforms. The section deals with the evolution of virality, the theoretical models, and the preceding studies. The literature review besides the evaluation of the machine learning models employed in the prediction of content virality and the decision-making process which influences the user engagement with the viral content is also discussed.

Chapter 3: Methodology

The methodology chapter describes the research design, the way the data was collected and the preprocessing methods. This is the paper that explains the data collection from Reddit using the Python Reddit API Wrapper (praw) and the ways of choosing subreddits and keywords that are indicative of viral content. Besides, this chapter gives the information on the steps that are done to make sure the integrity of data and the way of treating users' data with ethical considerations (Gold Nmesoma Okorie et al., 2024; Proferes et al., 2021). The data preprocessing section is devoted to the cleaning, normalization, and feature engineering processes, that the dataset is being processed for the analysis.

Chapter 4: Results

This chapter gives the results of the analysis of data exploration. It encompasses the visualizations and the interpretations of the elements that are the prime movers of the virality of the content. Furthermore, the performance analysis of different machine learning models, for example Gradient Boosting, K-Nearest Neighbors, Random Forest, XGBoost, and LightGBM, is given, and in addition, the evaluation metrics of these models are also provided. The section goes to explain the hyperparameter tuning and cross-validation results, which in turn makes it possible to pick the best model for predicting the virality of the content.

Chapter 5: Discussion

The discussion chapter understands the results delivered from the analysis, thus, giving the hints of the factors that the content turns into viral. It provides the results with the already established literature and elucidates the similarities and

the differences. The so-called section is dedicated to the evaluation of the research outcomes which are later applied in practice. For content creators and marketers, the section offers them with the tips to boost the engagement and the virality on the social media platforms.

Chapter 6: Conclusion

In this section, it is provided an argument that is strongly persuading in favor of the implementation of the proposed course. This last chapter presents the most important results of the research, which are then analyzed in the light of their impact on content creators and digital marketers. It is the solution to the problems the study faced and suggests the topics for the future research. The aforementioned recommendations are the ones that will be used to capitalize on the results derived from this research in order to enhance the content strategies and to get more engagement on Reddit and the similar platforms.

2. Literature Review

2.1 Theoretical Framework

With the spread of content on the internet, the mechanism of digital communication that facilitates the rapid and far-ranging diffusion across social networking platforms, is the virality of content (Joseph et al., 2021; Muller & Peres, 2019). This literature review is based on recent studies carried out by researchers. It is a concise analysis of the underlying mechanisms that make content viral and the decision making process that influences the users activities on the viral content.

Firstly, (Arjona-Martín et al., 2020), in their work, shed light on virality's historical development which starts from traditional media power and ends up in digital networks' current version. They suggest the virality of information operates as a hybrid communication medium taken from the entities that are integrated in terms of old media with the influence that is enhanced by digital networks. Such a perspective shows the difference from the centralized ways of broadcasting and local environment with the users who are the drivers and are themselves active participants of, the new digital media communications.

The research in the forefront by (Xu & Qian, 2023), is the introduction of the ViralGCN model, a new approach that makes use of a temporal, spatial cascade convolutional learning framework. That model is being different from the conventional approaches and the double agent is identifying. The dynamic information of user relationship and interactions over the time helps to occur the complicated structure sequence of content and user engagement. This shows the significance of considering both the extent the content is being disseminated and the user-to-user interaction patterns complications while ViralGCN is predicting the virality content on platforms like Weibo.

Furthermore, going far beyod the interaction in the content realm, (Reichstein & Brusch, 2019) article on viral marketing focuses on the decision making of the users as the ultimate mechanism of the marketing phenomenon. They differentiate between these two stages of the dynamic model by starting with the perception process, in which users are required to assess the content and then, based on this evaluation, they will decide whether to interact with it. This framework spotlights individual motivation factors and what drives the readers to eventual posting of viral content, pointing the cognitive path leading to the existence of viral content.

And taking a step further, (Garcia-Salirrosas, 2020) investigates viability of viral marketing as a strategy that takes advantage of the network effect that is embedded in customer interactions. The author supports that in the viral marketing process the emotional and social triggers are responsible for the growth in the virality. Especially, the user involvement is vital here being a vertex in the viral communication path.

Moreover, (Sanders, 2020) article encompasses the immediate impacts of virality on YouTube Influencers. (Sanders, 2020)'s work brings into light how the strategic play pushed by the content developers along with that of the algorithm of the platforms comes together in creating the viral potential of the digital content. Virality research demonstrates how big platforms sustain delicate equilibrium between pushing user-content and advertisement placement, which, due to advertising interventions, can define the visibility, reach, and views of viral content.

Overall, thus, researching in path of content virality, makes apparent the theoretical causes and practical settings of viral dynamics on the digital age. This review is built on the integration of those computing, psychological, and planning dimensions, which, in turn, not only enriches our theoretical understanding fo virality but also guiding the development of effective strategies of utilizing virial mechanisms in and content production and marketing.

2.2 Previous Research

The present development of machine learning has influenced in a great manner how social media platforms customize content to the individual user. According to research, (Taherdoost, 2023) examines how the neural networks employed to users' data in order to enhance the content and interaction with the users. However, the data collection about the privacy and ethical use of the data is an issue (Gold Nmesoma Okorie et al., 2024; Rajaraman & Ullman, 2012). The machine learning technology is indeed the figurative of this context. Thinking from a broader point of view it becomes obvious that this technology is going to play a major rolein the processing massive data, and in the process of detecting patterns assisting decision makers in content delivery.

Subsequently, (Shang et al., 2022) apply the techniques of machine learning in social media to predictive modeling concentrating on the changeability of social influence and homophily. Their work into graph neural networks alerts to the intricacy of social connections and the role of their pattern in content popularity. This is crucial for platforms like Reddit because the way users interact very much determines the reach of the contents (Proferes et al., 2021). This way, the accuracy of the prediction increases, and it is also does underline the significance of the intricate modeling to understand the spread of the information.

Furthermore, a deep learning model is investigated by (Gao et al., 2021) where they present a new model, namely CasWarn, which is destined to predict the spreadability of public opinion/sentiment. The competence of the virality model to conduct fast analysis and prediction independent of traditional network configurations gives an example of an advanced AI social media analytics that are flexible and adaptable . Those tools are the key of live monitoring and management of public relations, especially when there is crisis or misinformation periods concerns.

Regarding content marketing, by the time look for (Y. Chen, 2023), he demonstrates how the machine learning approach helps in the analysis of consumer behavior & optimizing engagement strategies. This research shows that data-driven insights can be the sources that produces a set of successful actions on social media that gives digital brands the edge in the market which is enhanced by their improvement in how well they lure their target audiences (Fan, 2023; Rosário & Dias, 2023; Yang et al., 2023). With this transition from common user engagement to specific marketing techniques, one can see the power of machine learning in the use of contextual strategy for any business success.

(Zhou et al., 2022), synthesized the models of information cascade indicating the different factors in determining the virality of content, which is a crucial subject for the scholars and practitioners. In doing so this work simultaneously accomplishes two vital goals by virtue of its systematic comprehensive categorization of literatures and methods. The two goals include promoting academic research and also bridging the gap between research and its practical applications in social media marketing and public information management (Chan et al., 2023; Dwivedi et al., 2021; Jamil et al., 2022; Shahbaznezhad et al., 2021). This way research insights get into actionable strategies used in media marketing and advertising.

Also, (L. Lopez et al., 2022) try to solve the demand of predicting virality of revealing press articles with machine learning techniques. They exemplified how NLP can assist media in present times by improving article impact prediction in their use of BERT. This proves that AI can be a critical means of gauging the distribution and impact of news in the digital world (Gupta & Kumar, 2023; Madakam et al., 2022).

Moreover, (Alassafi et al., 2023) stressed the importance of machine learning in huge datasets that social media platforms generate to predict trends. The role of this work is to emphasize the usability of different machine learning algorithms in the search of helpful information embedded in the complex user

data. Such technology is crucial for effective strategic decision making both for business and for content management (Earnshaw et al., 2020; Mikko Dufva, 2020).

In their study, (Elmas et al., 2023) enhance the measurement of virality on social media by setting the new measuring standards using empirical data. They conclude that note viral measurement are essential in determining how content about these issues is shared and is what ensures the dissemination of information is ethical and most engaging.

Following, (G. Chen et al., 2019) explored neural popularity prediction (NPP) model which applied deep learning to predict the virality potential of social media content. Their model that takes advantage of the time series and the user engagement data offers a fairly sophisticated analyzing tool for content creators and marketers to evaluate which posts can make it top/viral in this media by strengthening the predictive analytics in social media.

In the study (Ng et al., 2023), the authors examine the effectiveness of Shifted, Hawkes processes, and ARIMA models for social media activity forecasting. They discovered that the Shifted baseline does well for the short predictions, while ARIMA is more useful to the estimate of the total activity in the cases of the important real-world events. The study underlines the fact that choosing the right baseline depends on the length of the forecast and the metrics that are available. Hence, a mix of models will be more accurate in the prediction.

Finally, the author (Sadrnia, 2023), examines the impact of predictive modeling on marketing, she presents the ability to increase the effectiveness of campaigns by using historical data to predict the future outcomes. The study stresses the outstanding improvements in the performance indicators, e. g. response rates and ROI, caused by big data and AI progress. To sum up, the text talks about problems such as data quality and algorithmic bias, and then proceeds to advise a coordinated approach to be able to make the most of the predictive analytics for a competitive edge also using Time Series Analysis.

The table below present the overview of the studies:

"Enhancing Social Media Platforms with Machine Learning Algorithms and Neural Networks"	Hamed Taherdoost	2023	Machine Learning in Social Media	Neural Networks, Machine Learning	Through ML, determined users' preferences on social media
"Predicting the Popularity of Online Content by Modeling the Social Influence and Homophily Features"	Yingdan Shang, Bin Zhou, Xiang Zeng, Ye Wang, Han Yu, Zhong Zhang	2022	Predicting the Popularity in Online Forums	Graph Neural Networks, Attention Mechanisms	Prediction enhancement by simulating the social dynamics
"Public Opinion Early Warning Agent Model: A Deep Learning Cascade Virality Prediction Model Based on Multi-Feature Fusion"	Liqun Gao, Yujia Liu, Hongwu Zhuang, Haiyang Wang, Bin Zhou, Aiping Li	2021	Virality Prediction	Deep Learning, Multi-Feature Fusion	Efficient viral impact prediction by means of feature fusion method
"Comparing content marketing strategies of digital brands using machine learning"	Yulin Chen	2023	Analysis of Social Media	Ensemble Learning, Social Media Data Analytics	Identified important communication/behavior practices in digital marketing
"A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances"	Fan Zhou, Xovee Xu, Goce Trajcevski, Kunpeng Zhang	2021	Popularity Prediction, Information Diffusion Modeling	Stochastic Processes, Deep Learning	Extensive analysis and information taxonomy of propagation of information
"Artificial Intelligence Model to Predict the Virality of Press Articles"	Yesid L. Lopez, Didier Grimaldi, Sebastian Garcia, Jonatan Ordoez, Carlos Carrasco-Farre, Andres A. Aristizabal	2022	Virality of Press Articles	BERT, Natural Language Processing	A BERT model estimates the virality factor using article headlines as inputs
"Measuring and Detecting Virality on Social Media: The Case of Twitter's Viral Tweets Topic"	Tuğrulcan Elmas, Stephane Selim, Célia Houssiaux	2023	Virality Metrics	Transformers, Social Media Analysis	Development of a new approach able to detect viral tweets.

"Machine Learning for Predictive Analytics in Social Media Data"	Madini O. Alassafi, Wajdi Alghamdi, S. Sathiya Naveena, Ahmed Alkhayyat, Absalomov Tolib, Ibrokhimov Sarvar Muydinjon Ugli	2023	Predictive Analytics	Machine Learning Algorithms	Gain of knowledge concerning the user trends and conduct on social media.
"NPP: A neural popularity prediction model for social media content"	Chen Guandan, Kong Qingchao, Xu Nan, Mao Wenji	2019	Popularity Prediction	Machine Learning, Neural Networks, Time Series Analysis	Model that predicts the most popular social media content trends.
"Experimental evaluation of baselines for forecasting social media timeseries"	Kin Wai Ng, Frederick Mubang, Lawrence O. Hall, John Skvoretz & Adriana Iamnitchi	2023	Forecasting social media activity	Evaluation of Shifted, Hawkes processes, ARIMA	Short-term benefits for (Shifted), ARIMA is better for big events, the combination of models improves the accuracy.
"The Future of Marketing: How Predictive Modeling Optimizes Campaign Strategies"	Leila Sadrnia	2023	Optimizing marketing campaign strategies	Literature review, Mixed-method Research, Time Series Analysis	Predictive modeling improves KPIs

Table 1. Overview of Studies

Altogether these above studies have one thing in common they demonstrate an absolutely dynamic field in which machine learning powers social media “from communication channels” into “complex ecosystem capable of proper data analysis and predictive modeling” (Jordan & Mitchell, 2015; McCrea et al., 2023; Mohammed et al., 2016; Sarker, 2021). Machine learning if at advanced stage and is more likely to be linked to social media analytics giving firms access to more precise tools of playing a part in determining user engagement and standardization across various social media platforms.

3. Methodology

3.1 Data Collection

In this chapter , the method used in the collection of data from Reddit using the Python Reddit API Wrapper (praw) is being explored. This was kicked off by setting the praw using the necessary credentials like client_id, client_secret, and user_agent. This is those credentials which are vital for API authentication and keeping adherence to reddit API usage policies (Anugrah & Fakhruddin, 2020; Preibisch, 2018). The selection of subreddits and keywords was refined and poised precisely to target posts which had the potential to be viral. Communities in which engagement might be robust and keywords indicative of the trending content were the focus.

The figure below shows the whole group of subreddits and key words for the research on contents virality on Reddit which were used in the script for this study. The type of subredits explode, and there is a total different range of communities like “videos”, “AskReddit”, “gaming”, and many others that are selected, based their influence on users and the capability to make them viral. Keywords search, such as “viral”, “trending”, and “update”, that are likely commensurate with posts that have realized or have the potential of achieving viral status on the platform were specifically used in the key-word search. The implementation of this selection strategy ensures at the same time only data of the needed type and target with an indirect contribution to the research goals which are made up by understanding the content virality (Arjona-Martín et al., 2020).

```
subreddits = [
    'videos', 'AskReddit', 'gaming', 'pics', 'worldnews', 'todayilearned',
    'news', 'funny', 'aww', 'explainlikeimfive', 'dataisbeautiful',
    'PublicFreakout', 'Damnthatsinteresting', 'therewasanattempt',
    'Whatcouldgowrong', 'CrazyIdeas', 'facepalm', 'InternetIsBeautiful',
    'lifehacks', 'LifeProTips', 'bestof', 'oddlysatisfying', 'MadeMeSmile',
    'tifu', 'memeeconomy', 'crypto', 'economy', 'stocks', 'YouShouldKnow',
    'IAmA', 'science', 'AskScience', 'Documentaries', 'listentothis',
    'Futurology', 'philosophy', 'Art', 'space', 'gadgets', 'books', 'History',
    'music', 'upliftingnews', 'technology', 'sports', 'television', 'Earth',
    'movies', 'food', 'travel', 'photography', 'Fitness', 'educationalgifs',
    'nottheonion', 'architecture'
]

keywords = ['viral', 'trending', 'breaking', 'update', 'announcement',
            'launch', 'popular', 'top', 'discussion', 'news', 'hot']
```

Figure 1. The list of subreddits and the set of keywords used for data collection.

It is noteworthy that collecting the data was carried out by Python to do it by searching each given subreddit and keyword combination, taking down important data points like a post ID, post title, upvotes, number of comments, timestamps and URL. This involved checking for duplication during data entry in which any new submission would be added to the existing Excel file as long as it was available and hence data integrity and completeness were maintained. Furthermore, the script implemented a minimal delay ("time.sleep" as 0.1) between requests in order to avoid the exceedance of Reddit's API rate limits and to adhere a sustainable data collection practices.

The process of data collection ended in the storing of all the details collected into one Excel workbook that was structured, preceded by a completion message from the script's output.

```
Collecting data on r/architecture for 'update'
Collecting data on r/architecture for 'announcement'
Collecting data on r/architecture for 'launch'
Collecting data on r/architecture for 'popular'
Collecting data on r/architecture for 'top'
Collecting data on r/architecture for 'discussion'
Collecting data on r/architecture for 'news'
Collecting data on r/architecture for 'hot'
Data collection complete.
```

Figure 2. The terminal output that has been successful with collecting data which involves different keywords with a confirmation at the end.

This data is structured, and it is complete and that is sufficient for the purpose of analyzing viral content on Reddit and in addition, that is the foundation on all analytical exploration in the next chapters of the present study. A total of 113,772 data were collected initially with dates from 2006-12 to 2024-04. Later the number of data was reduced to 108,036 obtain more representative results.

	post_id	title	subreddit
0	eqia5m	Since we're talking about one of the first vir...	videos
1	99ng2v	Paul Flart loses his job due to his viral fame.	videos
2	18ztjyf	San Francisco break-in suspect seen in viral M...	videos
3	tjselj	Man whose car was damaged by the viral "Tesla ...	videos
4	sa8dos	Bob Saget hosting the first episode of AFV, se...	videos

	upvotes	num_comments	created_utc
0	45279	3181	2020-01-18 18:16:14
1	52173	6310	2018-08-23 16:37:52
2	3150	883	2024-01-06 08:28:22
3	42018	4435	2022-03-22 04:12:08
4	2297	146	2022-01-22 20:11:50

Figure 3. A snapshot of the data frame.

Finally, it is noteworthy that it is fundamental to be fully committed to ethical standards (including the preserved and protected rights of Reddit users), while presenting observations gathered from such online platforms (Floridi & Taddeo,

2016; Parthasarathy et al., 2024; Proferes et al., 2021). For example, in this project, the post IDs (post_id), is being considered and it does not in any way include personal user information. These IDs are utilized solely as unit post distinguisher blocks and to handle the assessment of content with viral aspects. Utilization and manipulation methods of data gathered according to Reddit's terms of service. Moreover, major purpose of using this data is research for academic purposes where it is important to understand the content virality patterns, its causes and not to disclose about the identity of Reddit users. Through such ethical approach, it is emphasized that this research process is supposed to be ethically correct and the user privacy at the forefront.

3.2 Data Preprocessing

First steps of the analysis and preprocessing of the Reddit dataset were dedicated to this section. Having no missing values and errors in the data, especially the outliers, is the key to getting data properly for further analysis. At the beginning, the dataset was carefully loaded from the Excel file and testing the validity of the total data points. Recording this time was paramount in knowing dataset temporal span and volume. To clean a dataset up and free the stream of data handling from the redundant information, the unnecessary columns were removed such as the URL column. Such action facilitated a more focused data handling process.

The thorough document checking process helped to maintain the data on the ground. Particularly, the "created_utc" column was altered into the datetime format, thus there could be an easier time analysis. With the convertion data couldbe more detailed breakdown like when the earliest and latest post has been carried out. It was also paid due care to the issue of unexpected encoding issues in both "title" and "subreddit" columns. This could be of help during the data entry process or when collecting special cases of data points. Besides, fetching

“created_utc” column and then creating the posts' year grouping to uncover ongoing data trends helped in revealing important parameters such as total upvotes per year which is significant for understanding engagement dynamics adjusted to time line flow.

It is noteworthy that a crucial decision of limiting it to only the data up to December 31 2023 was made purposefully. As for the choice to consider data only from posts published before a certain time period, it was partially motivated by the belief that the engagement figures based on very recent posts could be not representative enough, because it takes long for such content to be noticed and get the shares and comments. It could also be a part of the analysis to exclude an excessive amount of data which may cause a wrong perception of trends and ranking. It was employed this threshold-setting in order to acquire a dataset with enhanced stability and trust-worthy content. The total entries after this threshold were 108,036 of dataset.

```
Data columns (total 7 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   post_id       108036 non-null  object
 1   title         108036 non-null  object
 2   subreddit     108036 non-null  object
 3   upvotes       108036 non-null  int64
 4   num_comments  108036 non-null  int64
 5   created_utc   108036 non-null  datetime64[ns]
 6   year          108036 non-null  int32
dtypes: datetime64[ns](1), int32(1), int64(2), object(3)
memory usage: 6.2+ MB
```

Figure 4. The types of dataset after the threshold.

The preprocessing spilled over text data in titles, where text technique (normalization) was done to ensure better analysis. First, all text was converted to lowercase for the purpose of maintaining consistency. Second, a process of cleaning the data by removing special characters was done. Then, in preparation for natural language processing, words were tokenized and finally, only the stop

words were removed to focus on more significant words. Lemmatization was applied successively to disintegrate the words into the basic forms or dictionary form, that plays the crucial role in the dimension reduction with the help of natural language processing tools, e. g. sentiment analysis (Beysolow, 2018; Teahan, 2018).

Original title	**Tokenization/Remove Stopwords**	**Lemmatization**
Since we're talking about one of the first viral videos. This went viral before youtube even existed.	['since', 'talking', 'one', 'first', 'viral', 'videos', 'went', 'viral', 'youtube', 'even', 'existed']	since talk one first viral video go viral youtube even exist
Paul Flart loses his job due to his viral fame.	['paul', 'flart', 'loses', 'job', 'due', 'viral', 'fame']	paul flart lose job due viral fame

Table 2. Example of Text Preprocessing

3.3 Exploratory Data Analysis (EDA)

Subreddit popularity

The table below is presented exactly how many posts there have been in every subreddit. The popularity metric is the number of the submissions or the volume of the engagement in each subreddit, which gives the information about where users are the most active.

	subreddit	popularity
0	aww	2476
1	IAmA	2411
2	science	2397
3	funny	2394
4	worldnews	2341
5	technology	2340
6	pics	2337
7	videos	2321
8	news	2317
9	movies	2314
10	gaming	2300
11	AskScience	2300
12	todayilearned	2271
13	explainlikeimfive	2264
14	CrazyIdeas	2239
15	economy	2210
16	tifu	2206
17	Futurology	2201
18	television	2194
19	photography	2175
20	dataisbeautiful	2171
21	Art	2170
22	books	2149
23	facepalm	2129
24	stocks	2120
25	MadeMeSmile	2117
26	PublicFreakout	2113
27	music	2101
28	food	2097
29	Fitness	2073
30	space	2065
31	gadgets	2052
32	LifeProTips	2029
33	History	2028
34	sports	2004
35	travel	1998
36	memeeconomy	1968
37	architecture	1962
38	AskReddit	1903
39	upliftingnews	1866
40	Damnthatsinteresting	1855
41	education	1776
42	nottheonion	1770
43	bestof	1762
44	Documentaries	1719
45	YouShouldKnow	1663
46	therewasanattempt	1506
47	oddlysatisfying	1496
48	listentothis	1467
49	philosophy	1454
50	Whatcouldgowrong	1109
51	lifehacks	1006
52	crypto	1005
53	InternetIsBeautiful	787
54	educationalgifs	384
55	Earth	154

Figure 5. Subreddit popularity.

This distribution is presented the diversity of topics that Reddit users are interested in, and also a pointer to the communities that are the most engaging ones. It can be distinguished that 'Aww', 'IAmA', and 'science' subreddits are the most popular reddits, imply those communities enjoy a diverse spectrum of user interactions and content submissions. For stakeholders or advertisers, these insights are going to be pretty valuable for them in making the targeted strategies, while on the other hand, the new users will be benefitted in getting the community where they might find engaging content or like-minded discussion.

The number of monthly post volume from year 2006 to 2023

The figure below illustrates the monthly post volume from year 2006 to 2023.

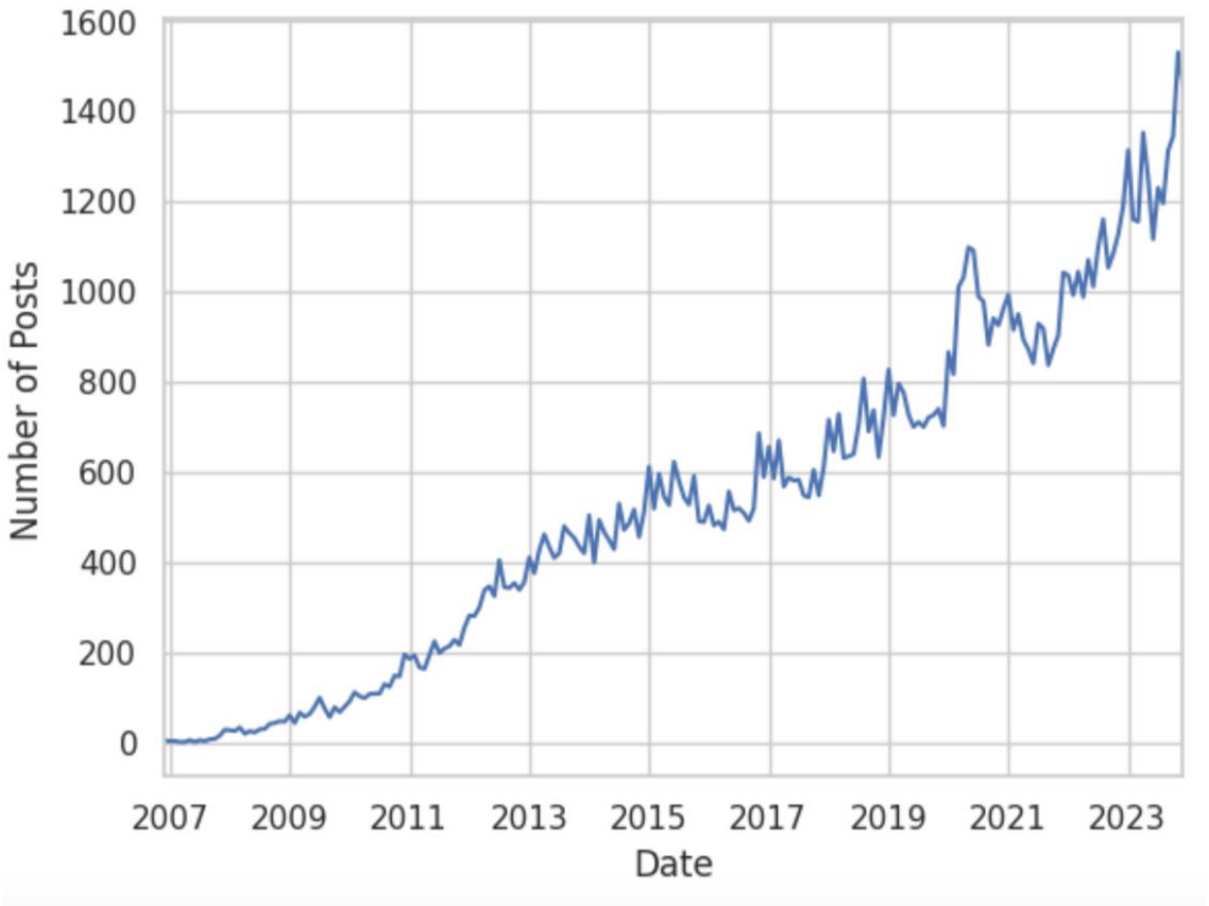


Figure 6. The number of monthly post volume from year 2006 to 2023.

This figure (line chart) above is presented the number of posts increasing with time on social media platform having a longitudinal perspective on user involvement from 2006 through 2023. Especially, there are noticeable spikes and plateaus which indicate varied interaction of different users at different points.

The initial growth was slow at first, then it became faster after that, probably because the platform became more popular or its user base widened.

Post 2015 the growth rate becomes more unstable but with an upward trend. However, there were some significant fluctuations. Furthermore, it is noteworthy that during 2020 a sharp upward slope is observed. After that, there is some decline until 2022, and then there is a strong upward trend until the end of the year 2023. This high peak could corresponds to specific events or platform innovation that could either help user to behave differently or use the application more frequently.

Boxplot of the spread in upvote count and comments

This pair of boxplots below takes attention to the distribution of upvotes and comments. Each boxplot is designed to present the range and concentration of these interactions, which are of great importance in understanding the user engagement on the platform. The first boxplot of upvotes is a graph that reveals the number of upvotes a post receives, while the motivation for the creation of the boxplot of comments is the involvement depth expressed over comments count.

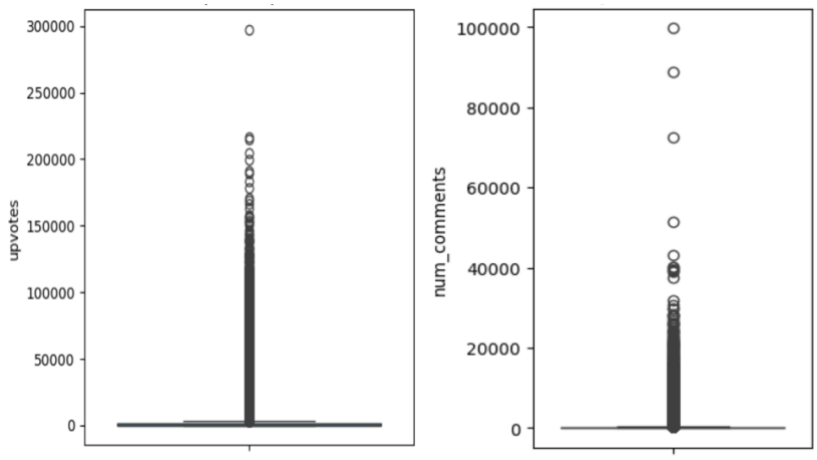


Figure 7. Boxplots of the spread in upvote count and comments.

Especially, boxplots analysis above shows the wide spread in the distribution of both upvotes and comments and plenty of outliers which mean that there are posts that have far exceeded typical interaction levels. This, in turn, provides the information of widespread popularity by well-liked posts and also reveals the norm of social communication of the rest of the content.

These insights are very important in determining the catalysts that drive users' connection to the content ,or causes them to forget. Such results could give content creators and moderators the means to formulate strategies for engaging and creating a lively dialogue in their communities. This is, therefore, the practical use of data-driven content management on social media platforms.

3.4 Feature Engineering

The extraction and engineering of features are a key factor when developing a machine learning model in the area of social media postings engagement study for improving predictive capability of statistical models. A first stage in this procedure implies that temporal features are extracted. This entail implicating various attributes of the data that can be derived from the timestamp of each post, such as day of the week and hour of the day. These factors, therefore, play a key role in the quantification of the increased or decreased user activity and the number of interactions' posts obtain, most of the time, during different time periods, e.g. days or weeks, day, night. Recognizing these temporal regularities help to forecast occasions of high level activity when posts are much likely to achieve more success and different emotions.

	post_id	title	subreddit	\
0	eqia5m	[since, talking, one, first, viral, videos, we...	videos	
1	99ng2v	[paul, flart, loses, job, due, viral, fame]	videos	
3	tjselj	[man, whose, car, damaged, viral, tesla, stunt...	videos	
4	sa8dos	[bob, saget, hosting, first, episode, afv, ser...	videos	
5	4gjg4q	[little, girls, tantrum, went, viral, russia]	videos	

	upvotes	num_comments	created_utc	\
0	45279	3181	2020-01-18 18:16:14	
1	52173	6310	2018-08-23 16:37:52	
3	42018	4435	2022-03-22 04:12:08	
4	2297	146	2022-01-22 20:11:50	
5	6047	734	2016-04-26 18:53:53	

	title_lemmatized	day_of_week	hour_of_day
0	since talk one first viral video go viral yout...	Saturday	18
1	paul flart lose job due viral fame	Thursday	16
3	man whose car damage viral tesla stunt go wron...	Tuesday	4
4	bob saget host first episode afv serve viral v...	Saturday	20
5	little girl tantrum go viral russia	Tuesday	18

Figure 8. Columns for day of week and hour of day of posting from the engineered features.

Then the identifying of the temporal features lines up with the text transformation utilising the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization (Beysolow, 2018; Havrlant & Kreinovich, 2017; Teahan, 2018).

With this algorithm which is actually the conversion of the post titles lemmatized into the digit codes that highlights against a documents word frequency, hence signifies the words that have the possibility to influence the popularity of any post. A number of 1000 unique head terms provided by the analytical model helps the model to harness the elements that influence the most in purchase prediction. Especially, vectorizations ensure that no NaN values exists in text data and there are no issues in using data analytical methods on the processed data.

For TF-IDF scores to be merged into the initial dataset it should be gone through process of the conversion of TF-IDF matrix into a DataFrame. The output DataFrame is then unobtrusively joined with original data that have been

addressed for the unexpected overlapping columns with the corresponding columns being renamed and another step of data checking and cleaning being done at this point for data consistency and reliability (Beysolow, 2018; Havrlant & Kreinovich, 2017; Teahan, 2018). This integration of the dataset gives it an added goodness whereby instead of being entirely unstructured and qualitative something is now quantitative and easy to evaluate the meaning of textual content in social media posts.

	10	100	11	12	13	15	16	18	20	2012	...	xpost	year
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

	yesterday	yet	york	young	youre	youtube	youve	ysk
0	0.0	0.0	0.0	0.0	0.0	0.321103	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

[5 rows x 1000 columns]

Figure 9. The TF-IDF matrix, that illustrates the1000 features, based on their relevance among several documents.

Then, the dataset goes through another step of enrichment by adding features which show post engagement level. As for instance, a binary indicator is developed to post that shows a mostly affecting 75th percentile of comments and put them alone with those highly engaged posts. This unique mode comes in very handy in identifying posts with an outstandingly high number of views and others with minimal ones. In addition to sorting posts by time created, the summit will be divided into four separate time slots: morning (between 5 a.m. - 12p.m.),

afternoon (between 12 p.m. – 17 p.m.), evening (between 17 p.m. – 21 p.m.), and night (between 21 p.m. – 5 a.m.).

Grouped indicators to estimates like the average number of comments of high-engagement posts for each time category are calculated to give adds value to the process. This type of classification evens allows marketers and content creators to identify in which times posts are more expected to engage users and initiated by them, which most importantly is useful if content publication is maximized.

	is_morning	is_afternoon	is_evening	is_night	high_engagement
0	False	False	False	True	0.227031
1	False	False	True	False	0.262391
2	False	True	False	False	0.311324
3	True	False	False	False	0.207008

Figure 10. The calculation of the proportion of high engagement posts in each part of the day (morning, afternoon, evening, night).

Lastly, regarding the virality of posts, a threshold of high-scoring posts assigned might help to develop an understanding of posts that fall within the top 20% of upvotes level being highly viral. Such classification play an important role in revealing and understand those features that contribute to the virality of a post. Some descriptive statistical data on what these highly shared posts have in common are provided, facilitating the process of getting to in-depth understanding of traits (that are dominant among viral posts).

```
count     21603.000000
mean      15270.074480
std       19629.602923
min        2125.000000
25%        3639.000000
50%        7398.000000
75%       18453.500000
max      297524.000000
Name: upvotes, dtype: float64
```

Figure 11. The statistics of high score.

As well the analysis includes determining the number of words in the post's title to investigate whether there is a correlation between the level of details of the title and how users interact and engage with the post. This extensive type of feature engineering helps in building an extensive base for social media engagement analysis as well as prediction. Moreover, using this method enables the more accurate and insightful determination of what drives user interactions on the mentioned social media platforms.

```
      post_id                                              title  \
55905  bxu6sf                [dominos, pizza, slice, life, 2017]
37245  punl30  [lpt, despite, may, said, either, commercials,...
23129  fsftqh  [oc, minecraft, vs, minesweeper, google, searc...
12698  2l0gz1  [til, downtown, seattle, actually, sits, top, ...
68199  1j9ph9  [reza, aslans, book, 1, amazon, distribution, ...

            subreddit  upvotes  num_comments         created_utc  \
55905    Documentaries       11            10 2019-06-07 15:45:51
37245      LifeProTips        6             6 2021-09-24 19:39:51
23129  dataisbeautiful       89            10 2020-03-31 19:17:59
12698    todayilearned     6728           390 2014-11-02 00:03:59
68199            books     1721           402 2013-07-29 13:46:15

                                      title_lemmatized day_of_week  \
55905                      domino pizza slice life 2017      Friday
37245  lpt despite may say either commercial doctor p...      Friday
23129  oc minecraft v minesweeper google search trend...     Tuesday
12698  til downtown seattle actually sit top original...      Sunday
68199  reza aslans book 1 amazon distribution custome...      Monday

       hour_of_day   10  ...  youtube  youve  ysk  high_engagement  \
55905           15  0.0  ...      0.0    0.0  0.0            False
37245           19  0.0  ...      0.0    0.0  0.0            False
23129           19  0.0  ...      0.0    0.0  0.0            False
12698            0  0.0  ...      0.0    0.0  0.0             True
68199           13  0.0  ...      0.0    0.0  0.0             True

       is_morning  is_afternoon  is_evening  is_night  high_score  \
55905       False          True       False     False       False
37245       False         False        True     False       False
23129       False         False        True     False       False
12698       False         False       False      True        True
68199       False          True       False     False       False

       title_length
55905            28
37245           107
23129            56
12698           189
68199           114

[5 rows x 1016 columns]
```

Figure 12. The new dataset with the addition of the column title length.

In the last feature-engineering step, a new column “subreddit_popularity” is added to the DataFrame. This column counts for each subreddit the number of posts and this has the value of how each subreddit is active or popular. The method “groupby” is used in combination with transform(‘count’) to partition the counts of posts into the corresponding rows of the DataFrame for each subreddit.

One hot encoding

Then, it is done one-hot encoding for three columns.

‘day_of_week’ to prefix ‘dow’
‘title_length_bin' to prefix ‘title_length’
‘sentiment_category’ to prefix ‘sent’,

so that the development of predictive models is done afterwards (mainly based on these).

3.5 EDA of Engineered Features

Average daily upvotes

The figure “the average daily upvotes” below depicts the density of occurrence of the posts that are getting the average upvotes on Reddit throughout a week. The goal of this analysis is to find out the possible patterns of user engagement, which can be the basis of the most active days for interaction within the community. This way data provides with the essential understanding of the instances for the posting for the highest possible visibility and user involvement.

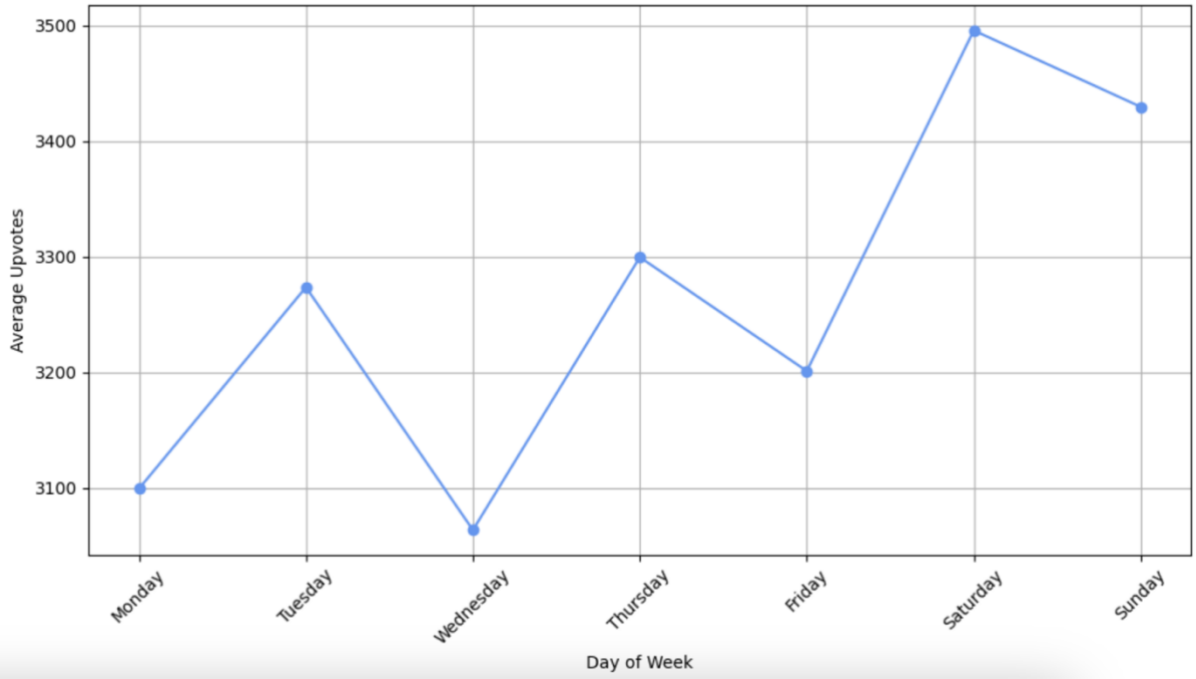


Figure 13. The average daily upvotes.

This figure above which presents the average daily upvotes fluctuates remarkably over a week. Especially, this indicates a sharp peak on Saturday, which means that posts made on this day usually get more upvotes than those made on other days of the week. On the other side, Wednesday seems to be the least involving day (because of its smaller average number among upvotes).

This pattern could be a way for it to be established that weekends, in particular Saturdays, are when the Reddit users are most active or ready for the content, so that it could be used strategically for the scheduling of the posts, such that interaction and visibility on the platform are increased.

Average upvotes per hour of day

The figure below of "Average upvotes per hour of day" is presented the fluctuations in average upvotes obtained by posts on this social media platform over separate hours in the day. The specific goal of this analysis is to pinpoint the maximum engagement period when the users are most active and average posts receive.

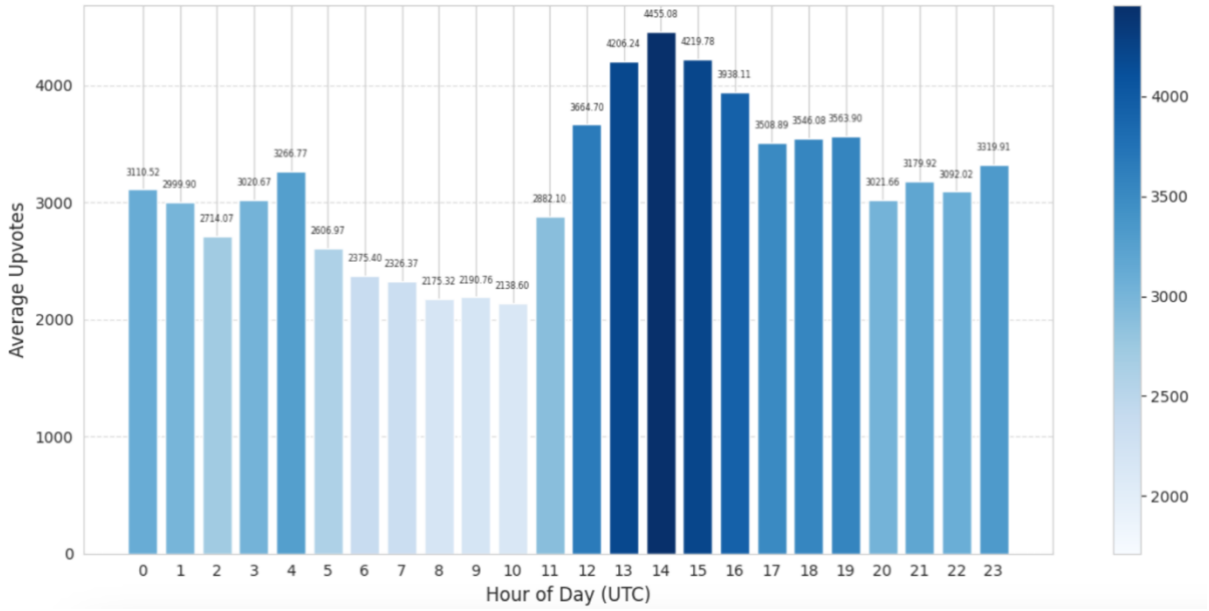


Figure 14. Average upvotes per hour of day (UTC).

This figure above, along the X-axis which represents the hour of the day, starting from 0 (midnight) and ending at 23 (11 PM), shows the average number of upvotes that were received during different hours of the day. This graph obviously shows that the posts get most upvotes in the afternoon and evening hours (between 12 p.m. – 19 p.m.), with a rise in the late night.

These data bring to the surface several crucial points. Firstly, the morning hours experience the least amount of engagement and it shows up in that there always be the lowest upvotes. The engagement starts to rise in the late morning, it reaches a peak at 14.00 p.m. Hence, this observation could means that the

audience is usually active and available in the afternoon hours, which may result from spare time of after work or school (between 13 p.m. – 15 p.m.). Such findings are important in the sense that they can help content producers as well as the marketers who aim at the maximum engagement.

Average upvotes by day of the week and hour of the day

The figure (heatmap) below offers a comprehensive visualization of average number of upvotes, indicating the day of week and hour of day as discriminating criteria. This kind of visual is very important in knowing when users are most active and when content gets the most engagement. Also, this makes it possible to present a transparent matrix of the hours by the days and then to detect the patterns that could be unseen through the simpler data aggregation methods. Through this in-depth analysis, it can be gained knowledge about user engagement patterns over various time periods and different days.

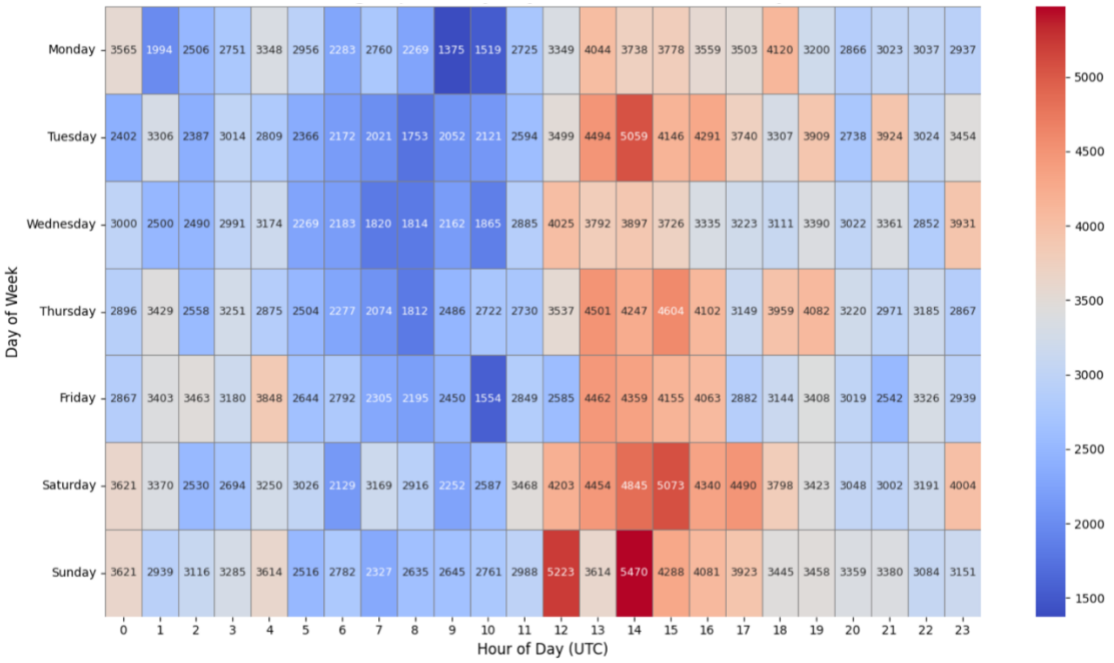


Figure 15. Heatmap of average upvotes by day of the week and hour of the day (UTC).

After looking at this heatmap above, it is seen that there are specific peaks at particular hours of certain days. Such as through the weekend, where it is presented that average upvotes are higher compared to other times, hinting at increased user activity and content interaction. The color scale from blue to red showcases the rating levels ever so slightly,and represents higher average upvotes with warmer colors. This way of visualization helps in the identification of the most active zones during the week.

Especially, the above figure helps determine the times of the day or the week when the visibility and user engagement of the posts are at the peak. Such numbers are of paramount importance both to business and content makers, as they help in planning posts and marketing strategies. Academically, it is a starting point for further research on the factors behind user interaction on social networks. The study of them makes it easier to conceive theories about the behavior of online users, giving directions about the improvements in the platform design process and performance, to make them work according to the peak usage times.

Word Cloud for Viral Post Titles

The figure below depicts the word cloud i. e. the frequency of the words created by the names of popular viral posts. The color depicts the frequency of words in these viral posts. Different font size in the cloud reflects word occurrence with bigger font component being the common word in the data set. This visualisation intends to show the most popular themes and topics which have attracted the community, thereby revealing what is of interest to the group or creates a lot of engagement.

Figure 16. Word Cloud for Viral Post Titles.

Especially, this figure above represents a simple word cloud where “new”, “top” “til”, and “break” words are at the top line connecting them with the fact that at discussions they always discuss topics that are actual, relate to the audience and have a great impact on the participants. The words like “think”, and “want” indicates the level of engagement between the user and the content.

In addition, this use of opinion or personal-based content leads to a strong engagement that works. These insights are of the most importance for the people who are involved in the creation of content and marketing who need to know the trends and to create the messages that will be an impact to their target audience.This graphic visualization will give the sense of what are the main topics and the tendency of different discussed patterns during the period. Consistantness of terms such as the word “launch” and “announce” could provide with information regarding the community and its interest in development.

The Connection Between Sentiment Polarity and Upvotes

The figure below called “The Connection Between Sentiment Polarity and Upvotes” shows the correlation between the sentiment polarity of the posts and the number of upvotes they get. Sentiment polarity is a metric directly derived from the content of the post, scale value -1 to +1, from which the language is quantified into highly negative (negative) to highly positive (positive). The analysis which is to follow will try to grasp a hold on the extent to which emotional tone of a post and its engagement in number of upvotes are correlated with each other, which will be a great insight into how the users interact with the content and the emotional preferences they have.

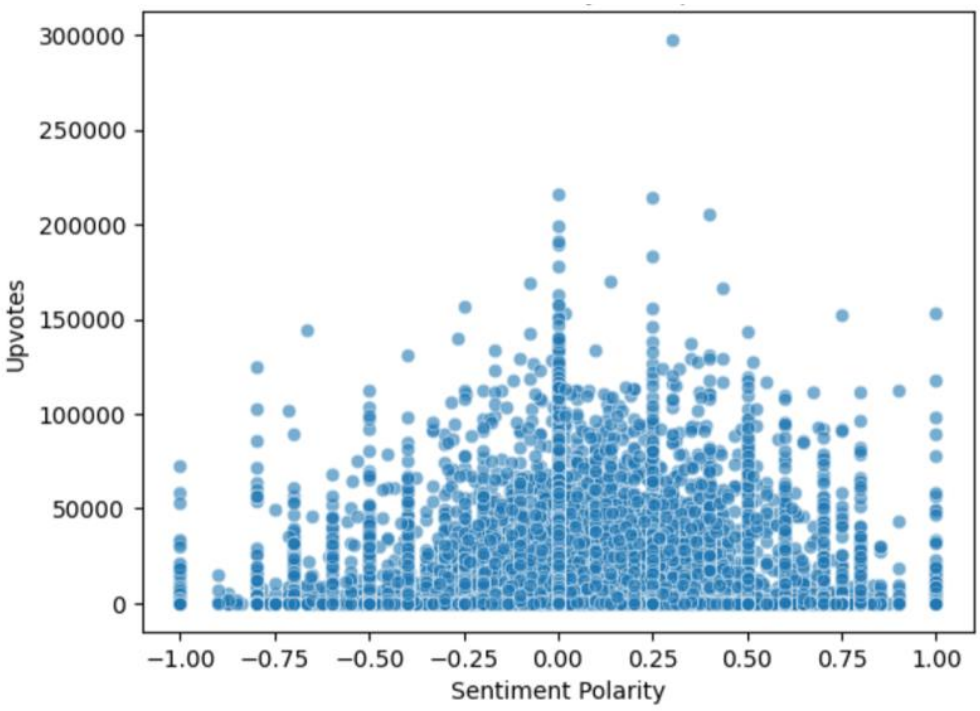


Figure 17. The Connection Between Sentiment Polarity and Upvotes.

When looking at the scatter plot above, it is obvious that most of the data points are located around the center of the sentiment polarity spectrum, which means that the posts contain a lot of neutral towards positive sentiments. Also, such wide variety of posts can manage to get high levels of engagement.

This distribution opens up a lot of different ways to look at it. Next, it reveals that emotional content that impacts negatively has less potential in being viral. This analysis shows the complicated interactions of the users and the engagement of them on the social platforms, it is highlighting that the sentiment is an important factor of the content.

Especially, in the figure below (Pie Chart), it can be distinguished the percentage of viral posts by sentiment category (Negative, Neutral, Positive) of viral posts. That is, posts that belong to the top 20% of upvotes on the platform).

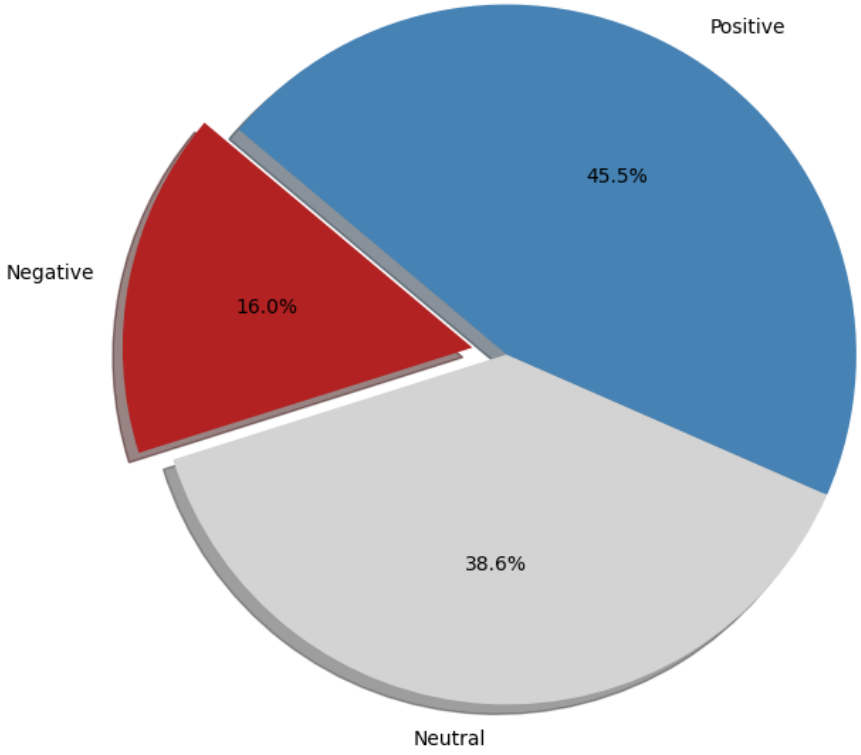


Figure 18. The percentage of viral posts by sentiment category.

Matrix of the scatter plots dealing with post metrics

The figure below presents a matrix of scatter plots where each of them reveals different sets of relationships in the data set that encompass diverse metrics of social media posts. These are upvotes, number of comments, sentiment in titles and title length. The key reason for creating these plots is to understand how these variables are related and how they influence each other, thus providing a multidimensional analysis of user engagement and content reception.

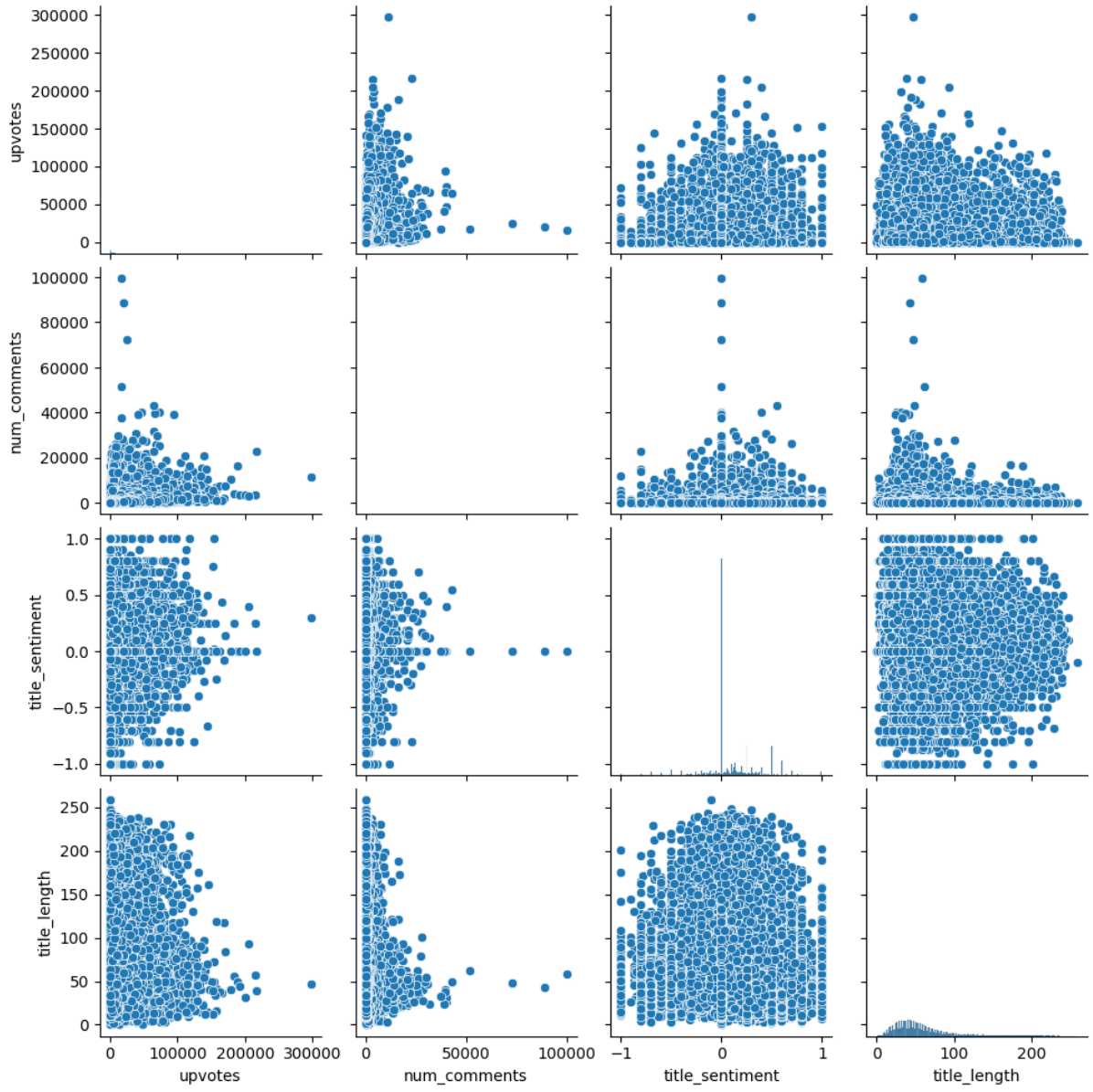


Figure 19. Matrix of the scatter plots depicting interactions of post metrics.

It is noticeable that each scatter plot is properly labeled using caption in order to show the link between variables in the plot. Initially, the cases within the top row dissect the association of upvotes to the number of comments, title sentiment and title length, revealing the trends and outliers. These relationships help to identify the factors that may determine the popularity or the reception of a post on social media(Fan, 2023; Rosário & Dias, 2023; Solomon & Tuten, 2016).

Matrix rows such as the second, the third and fourth expand on how things such as comments, title sentiment and title length all relate. For instance, the plot of title sentiment versus number of comments, can show if the posts with a higher degree of emotional portrayals, evidenced by the higher number of comments get more user's interactions. A closer look at the plots shows that there are some trends and deviations that are worth mentioning:

1) It can be seen that comments are not quite positively correlated with the number of upvotes. That is, a viral post may get less comments than a non-viral post.

2) Most viral posts tend to have a positive/neutral title sentiment

3) Most comments have a more neutral title sentiment

4) Most comments are in title length between 0 and 100

The consequences of these findings are relevant for content producers and social media planners, as they emphasize the need to know the interaction patterns of the audience. Through this, those who are interested can create and shape relevant strategies in order to boost reach and interaction. Thus, they could determine decisions on the composition of the content whether it is about the title, the post length, or the emotional impact of the content.

3.6 Development of Predictive Models

3.6.1 Data Splitting

In this section, the preparation stage of the dataset was carefully segmented into training and testing subsets to perform as a neutral indicator of the performance evaluation of the machine learning models. Such stratification is done via the 'train_test_split' performance from scikit-learn, in this case the split took about 80% for training and the remaining 20% for testing.

The split was done in a way that viral and non-viral posts were equally distributed in both the training and testing sets. One the most important this method is the factor needed to keep the level balance in all the data sets. It is evident that there are two groups of data, that is large (non-viral group) and small (viral group), that was 86,433 and 21,603 respectively.

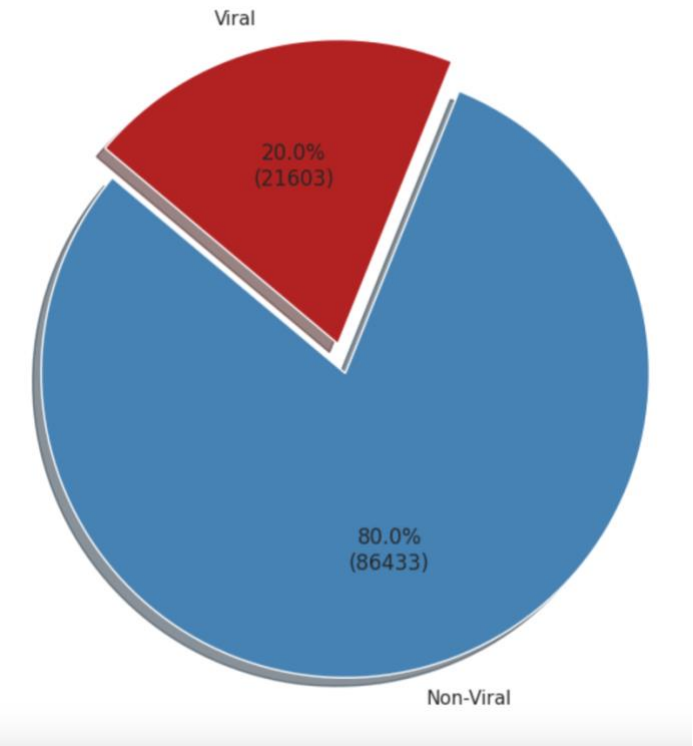


Figure 20. The distribution of Viral/Non-Viral posts.

3.6.2 Feature Selection

Feature selection, which was executed for high-accurate and relevant model, was optimized for only numeric and boolean data types. It is noteworthy that, unnecessary columns like 'post_id', 'title,' 'url,' and other columns that could introduce bias or were irrelevant to the prediction task were taken out of the feature set.

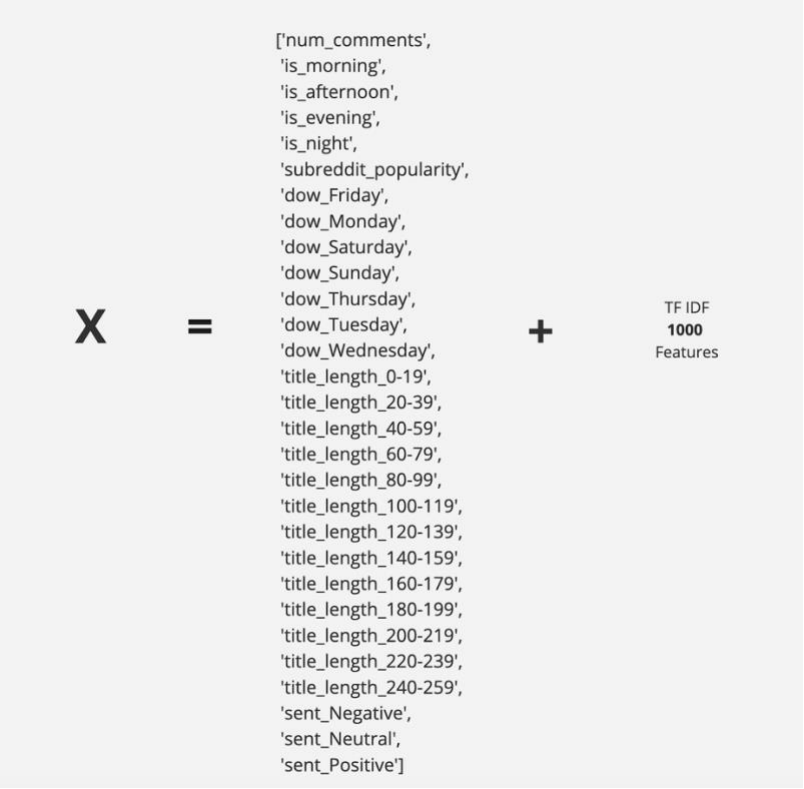


Figure 21. The definition of the feature set X

While the feature set Y is the column: ‘viral’

3.6.3 Machine Learning models

The chosen models below are five different machine learning algorithms as they are appropriate for the case dataset and they will be handy in attaining the research objectives:

1) Gradient Boosting: Responsible for successful prediction, Gradient Boosting was chosen due to its capacity to work with data sets that has a class distribution skewed (imbalanced dataset) and to model the most complex of patterns. By making a tree, one after another where every next tree corrects errors performed by the previously built trees, Gradient Boosting models are able to do well even on the data sets where a target variable distribution it skewed (Géron, 2019).

2) K-Nearest Neighbors (KNN): KNN was picked for its simplicity and efficiency in the classification tasks. It operates on a basis of several clustering like comparing the efficiency of a new post's data with the older one as well as making decisions based on the samples in a feature space. This method is especially important for situations, where decision boundary goes through non-linearity. KNN's dependence on feature similarity lets it make very intuitive predictions, which is exactly what it is needed when there is a dataset where the patterns among the data points need to be closely examined (Géron, 2019).

3) Random Forest: This model was built for the reason of its resistance to overfitting and, therefore, it is very reliable for the use with big datasets which contain both numerical and categorical data. The Random Forest algorithm is a combination of different decision trees into one in the effort of getting a more reliable and sturdy result. A model which is capable of performing both prediction

as well as regression has a broad spectrum and is applicable in solving a wide range of complex decision problems (Géron, 2019).

4) XGBoost: The reason why this algorithm was chosen is because it is very fast and it has high performance, which is an advantage in large datasets and in competitive scenarios such as machine learning competitions. XGBoost of gradient boosting algorithm sees an efficient implementation, and also it provides advanced features for the model design can be tuned and regularized to avoid overfitting (Géron, 2019).

5) LightGBM: This is distinguished to be the most expeditious one in terms of the training speed and apparition and hence, can be employed to address highly-dimensional data. This model is an improved version of the gradient boosting that employs tree-based learning algorithms to provide faster and more efficient processing. LightGBM's ability to manipulate big datasets evenly without having to shift the accuracy of its model is an asset for any data scientist, where computational time is limited or during when data volumes are crucial (Zhang et al., 2019).

Finally, the RandomizedSearchCV will be used for the hyperparameter tuning of the five machine learning models, and finally will become the prediction from selected model using Explainable AI with SHAP for a viral post (Aas et al., 2021; Géron, 2019)

3.7 Time Series Analysis

Firstly, the analysis is started by pandas library importing data from the data file (for the Time Series Analysis). Also, the decisive one date\time variable, 'created_utc', is cast on the datetime type, and thus, it can be performed accurate in the analysis. It is important to get this conversion right, allowing to separate the data based on specific time frames. It is a strong basis for all the temporal analysis operations.

Secondly, the first step of processing is after the setup of which a thorough check is done to ensure that there will be no missing values within the various columns, which are for further analysis completion and accuracy. As such, any inconsistencies or inadequacies in data that might lead to unbalanced results in the later stages are taken into account to ensure that the outcomes are objective with the use of basic techniques. The data is set to filter to contain statements that are limited to the period before 31, December, 2023, this can improve the validity and relevance of the analysis by considering all the recorded data within the defined period (as previously mentioned).

Thirdly, in this stage, a comprehensive exploratory study which includes trends, cyclical patterns and seasonality as well as that of the 'upvotes' are explored with the statistical decomposition technique. A seasonal approach is used in decompose the series in its different component (trend, seasonality and residuals). This operation is critical to see underlying trends that could lead to such appropriate corrections or smoothing techniques to be applied in a predictive forecast.

Fourthly, the stationarity of the time series is determined using the Augmented Dickey-Fuller test (ADF) to tell whether the series can be suitable for model ARIMA modeling (if it is differenced). Consequently, according to ADF test results may be the series differencing in order to try and achieve stationarity, thereby making the subsequent ARIMA modeling assumption to be satisfied. This

however is a crucial step since the usage of non-stationary data may lead to the production of biased results and inaccurate prediction.

Fifthly, the autoregressive integrated moving average (ARIMA) model, which has its seasonal version as SARIMA, is fitted to model the analyzed time series data (Arunraj & Ahrens, 2015; Ng et al., 2023). Furthermore, the autocorrelation and partial autocorrelation analyses are used to select the model parameters so that the fitted model is able to capture the dynamics of the dataset inherently which includes any seasonal patterns identified in the previous study stage. The last 12 months of data (that is, from January 1, 2023 - December 31, 2023) are kept aside as a test set to assess the model's prediction accuracy. It will be a sound measure of the model's performance on the recent and unforeseen data.

Sixthly, it is noteworthy that further forecasting will be made with Prophet model, for the next 12 months from the end of the data. That is, from January 1, 2024 to December 31, 2024. This very well covers all seasonal variations, that can have enormous impact on social media indexes like upvotes. Prophet, in turn, outperform very well under conditions with high seasonality and fluctuations, thanks to the flexibility of its models that take into account the seasonal components and the effects of holidays directly (Toharudin et al., 2023).

Finally, the model is thoroughly trained on the dataset that is comprised of all types of anomalies for the precise description of every temporal dynamics. Furthermore, post-training, it employs a decomposable time series model with main components: trend, seasonality, and holidays. This underlies the most critical part of the process that directs towards distinguishing specific times of the year and events which significantly affect the activity of upvoting, hence for all the actionable insights regarding the content engagement patterns. Therefore, this level of precision is quite helpful in planning content strategies during the periods when target audience is highly active.

4. Results

4.1 Performance Analysis of Machine Learning algorithms

- Gradient Boosting

```
Gradient Boosting - Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     17287
           1       0.78      0.75      0.77      4321

    accuracy                           0.91     21608
   macro avg       0.86      0.85      0.85     21608
weighted avg       0.91      0.91      0.91     21608

Gradient Boosting - Confusion Matrix:
[[16393   894]
 [ 1086  3235]]
Gradient Boosting - Accuracy Score:
0.9083672713809701
```

Figure 22. Metrics of Gradient Boosting algorithm.

The Gradient Boosting model obtained an accuracy of 90.84% in the classifying posts. Also, the precision and the recall scores of 0. 94 and 0. 95, respectively. On the other hand, regarding viral content, the precision was 0.78 and 0.75 was the recall of the controlled and instructed review of the statement that was given to non-viral posts. Also, the F1-score is 0.77, which means it has moderate effects in detecting viral posts. The confusion matrix denotes the vast number of false negatives (1086), hence, the difficulty to detect all the possible viral incidents. This is the evidence that the model is competent in classifying non-viral posts and at the same time the weaknesses in viral post detection are exposed to which the improvements can be achieved to avoid the missed opportunities

- K-Nearest Neighbors (KNN)

```
KNN - Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94     17287
           1       0.75      0.75      0.75      4321

    accuracy                           0.90     21608
   macro avg       0.85      0.85      0.85     21608
weighted avg       0.90      0.90      0.90     21608

KNN - Confusion Matrix:
[[16228  1059]
 [ 1065  3256]]
KNN - Accuracy Score:
0.9017030729359496
```

Figure 23. Random of KNN algorithm.

The K-Nearest Neighbors (KNN) model turned out a very high score, an accuracy of 90.1% which a post can differentiate between viral and non-viral content. The model reached the top grade in the precision and recall fields and achieved the scores of 0.94 which is extremely impressive. Also, F1-score reached 0.94, which shows its high efficiency in the recognition and classification of non-viral contents.

However, the precision and recall of viral posts are lower while both are 0.75. The result of the F1-score turned out to be 0.75, which means that the performance of the subjects in the experiment was moderate in correctly identifying viral content. The confusion matrix shows that there are more false negatives (1065) than other mistakes, which means the model is more likely to misclassify viral posts as non-viral and thus it can be missed the trend of that content. This means that there is a necessity to enhance the model to boost its effectiveness on the detection of viral posts.

- Random Forest

```
RandomForest - Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.96      0.94     17287
           1       0.80      0.73      0.76      4321

    accuracy                           0.91     21608
   macro avg       0.87      0.84      0.85     21608
weighted avg       0.91      0.91      0.91     21608

RandomForest - Confusion Matrix:
[[16513   774]
 [ 1182  3139]]
RandomForest - Accuracy Score:
0.9094779711218067
```

Figure 24. Metrics of Random Forest algorithm.

The Random Forest model has a high overall accuracy rate of 91% in differentiating the posts as viral or non-viral. It is a perfect at detecting non-viral posts with a high precision of 0.93. Also, the recall of 0. 96 means that it can correctly classify the content that is going to be not viral. Nevertheless, its effectiveness in the detection of viral posts is not very high, because a precision of 0.80 is reached and a recall of 0.73, which shows that some difficulties are encountered in all the viral content that is going to be captured. This difference has shown that there is a necessity to adjust the model to improve its predictive performance for viral posts.

- XGBoost

```
Gradient Boosting - Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     17287
           1       0.78      0.75      0.77      4321

    accuracy                           0.91     21608
   macro avg       0.86      0.85      0.85     21608
weighted avg       0.91      0.91      0.91     21608

Gradient Boosting - Confusion Matrix:
[[16393   894]
 [ 1086  3235]]
Gradient Boosting - Accuracy Score:
0.9083672713809701
```

Figure 25. Metrics of XGBoost algorithm.

The XGBoost model shows a strong performance, especially in the classification of posts that are not viral, with both precision and f1-score rates at 0.94.

Nevertheless, the precision of the viral class remains at 0.78 for precision. Also, recall is at 0.75 which implies a moderate level of the effectiveness in the prediction of the true viral posts. It is noteworthy that the model realizes a general accuracy of 0.91. The confusion matrix shows a small imbalance with more false negatives (1086) than false positives (894), thereby, indicating that the model is somewhat more cautious in predicting the viral status.

- LightGBM

```
LightGBM - Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     17287
           1       0.81      0.78      0.79      4321

    accuracy                           0.92     21608
   macro avg       0.88      0.87      0.87     21608
weighted avg       0.92      0.92      0.92     21608

LightGBM - Confusion Matrix:
[[16474   813]
 [  935  3386]]
LightGBM - Accuracy Score:
0.9191040355423917
```

Figure 26. Metrics of LightGBM algorithm.

The LightGBM model is able to recognize the non-viral posts very well with precision, recall and f1 score rates of 0.95, meaning high accuracy in this category. Nevertheless, its effectiveness lightly decreases for viral posts, as the precision is 0.81. Also, recall and f1-score is 0.78 and 0.79 respectively, thus this indicates a moderate success in the identification of viral content. The total accuracy is at 0.92 and that is a robust performance. The confusion matrix shows that the number of false negatives (935) is higher than the number of false positives (813), which indicates that LightGBM is more efficient compared to the other 4 machine learning algorithms in detecting the viral posts.

4.1.1 Hyperparameters

The RandomizedSearchCV was used for the hyperparameter tuning of the five machine learning models based on specific parameters in each case. The results are presented below:

- Best parameters for Gradient Boosting:

  N estimators = 50
  max depth = 1
  learning rate = 0.5

- Best parameters for KNN:
  p = 2
  n neighbors = 3
  leaf size = 20

- Best parameters for Random Forest:
  n estimators = 10
  min samples split = 4
  min samples leaf = 2
  max depth = 5

- Best parameters for XGBoost:
  subsample = 0.9
  n estimators = 300
  max depth = 7
  learning rate = 0.1

colsample bytree = 0.9

- Best parameters for LightGBM:
  subsample = 0.8
  num leaves = 31
  n estimators = 100
  learning rate = 0.1
  colsample bytree = 0.8
  boosting type = gbdt

4.1.2 Cross-validation

The best parameters for five machine learning algorithms obtained from the hyperparameter tuning process were used that go through 5-fold cross-validation to improvish their performance and ensure their robustness (Géron, 2019).

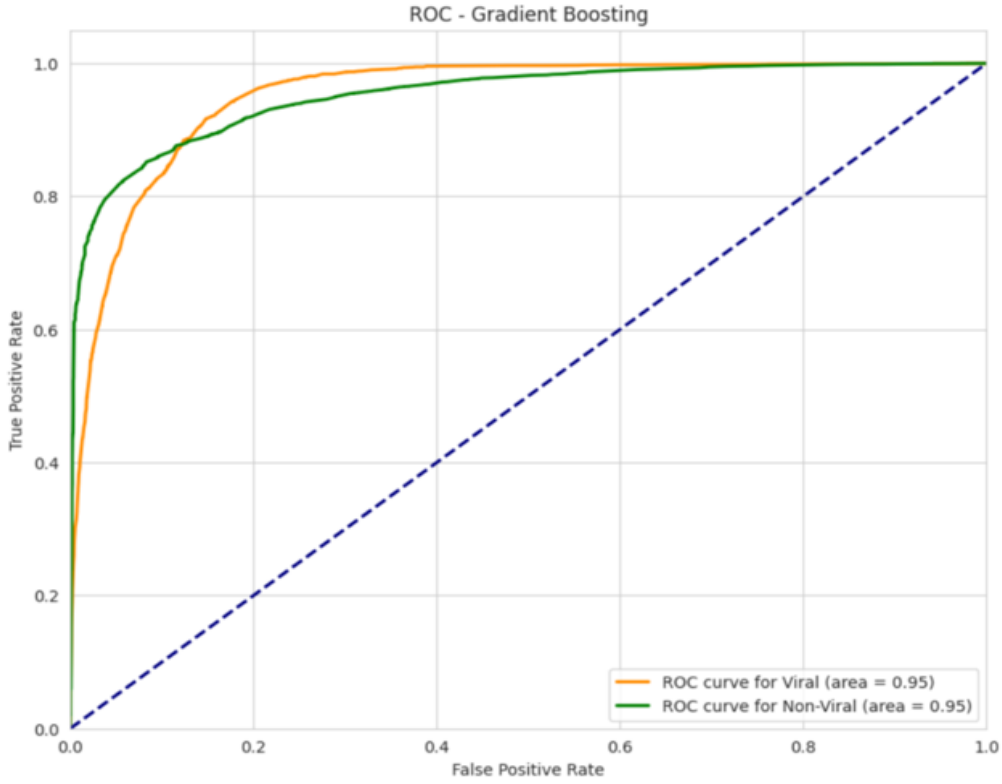
This is the way that the performance of each model is evaluated by looking at its predictions on different parts of the dataset, so that can have the whole picture of the generalizability and reliability of the model. The strict validation process is necessary in the choosing of the best model for the task, hence the validation procedure is the one that ensures the final choice is capable of giving the same and reliable predictions to the new data that is unseen.

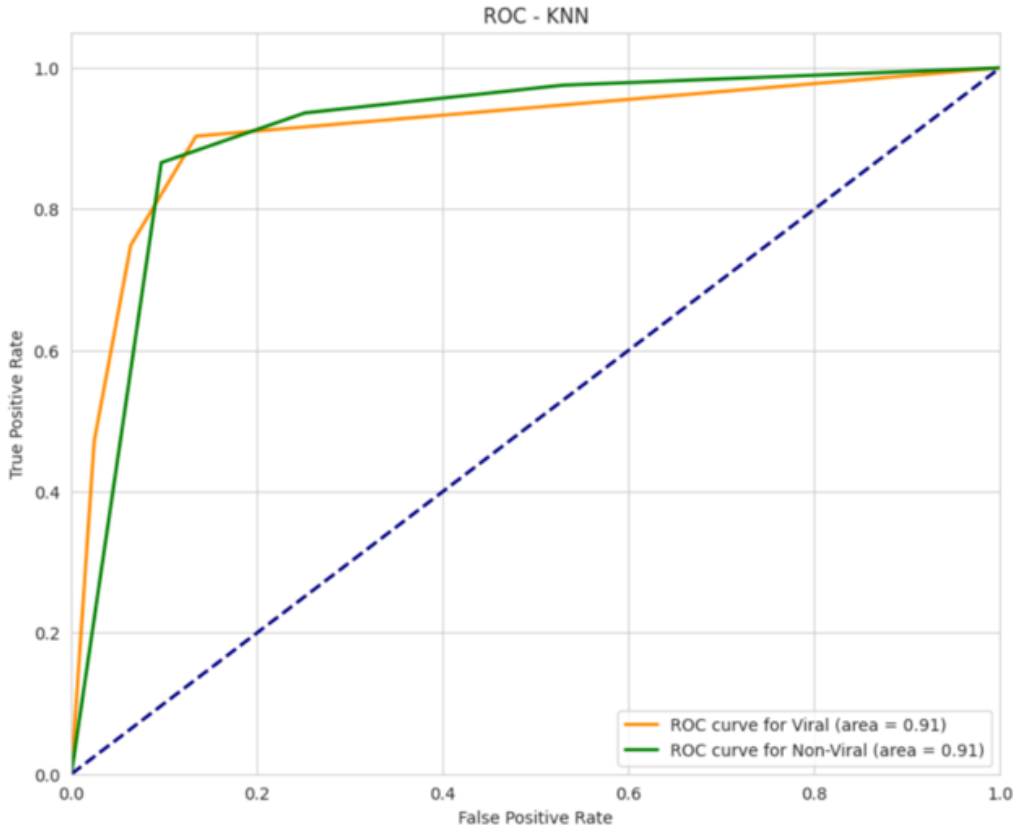
```
Gradient Boosting CV average score: 0.8419516857521474
KNN CV average score: 0.6945831360752229
Random Forest CV average score: 0.7999185450873483
XGBoost CV average score: 0.7547488256913188
LightGBM CV average score: 0.7611726244084892
```

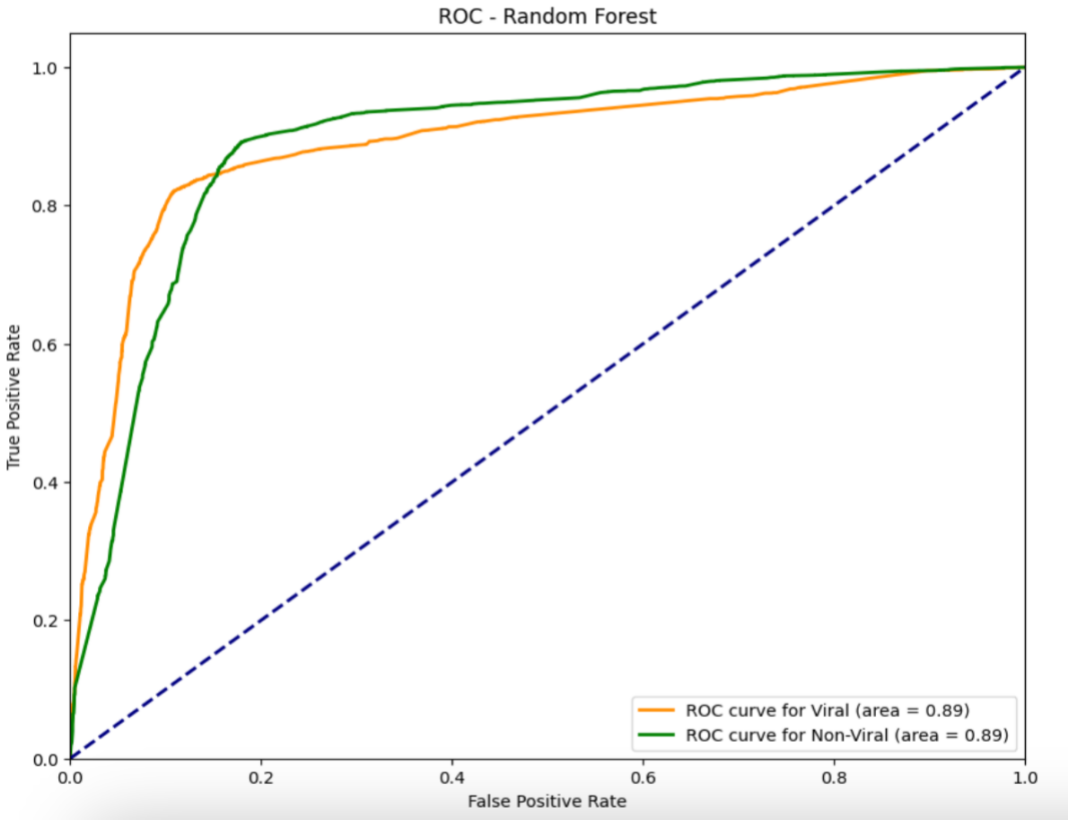
Table 3. Cross-validation scores of ML algorithms.

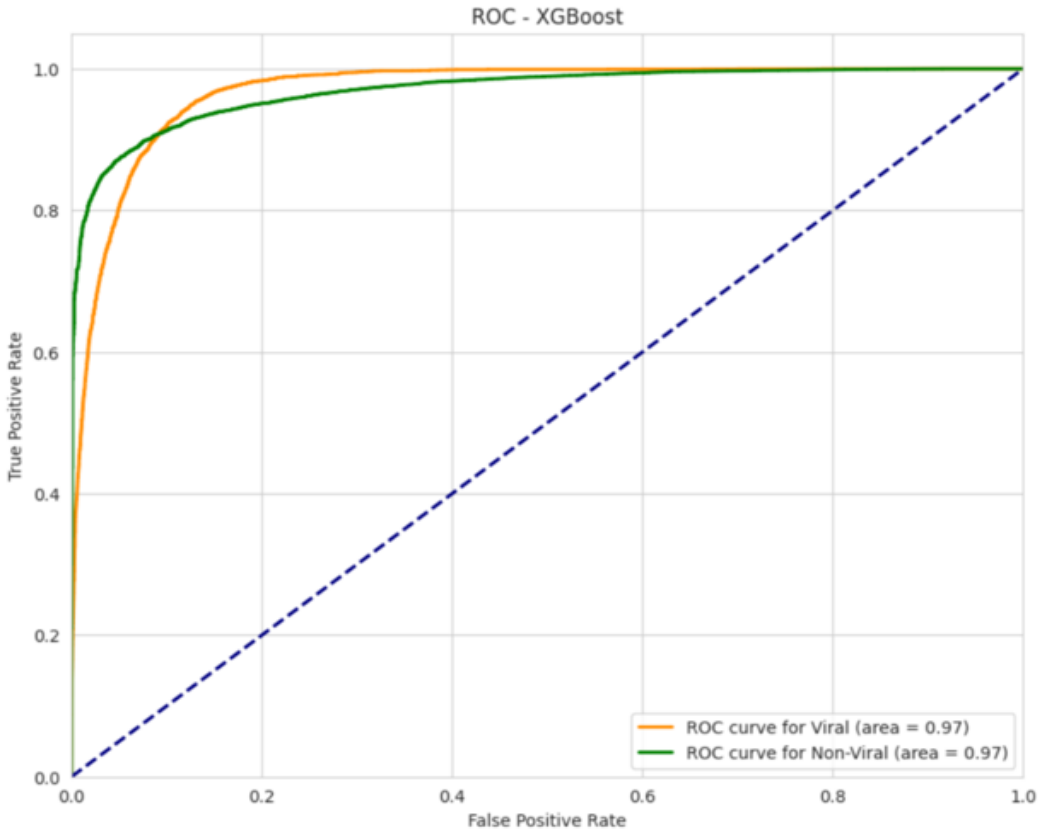
It is noticed according to the table above that the highest score has the Gradient Boosting CV with 0.84, followed by Random Forest with 0.80, LightGBM CV with 0.76, XGBoost CV with 0.75, and last KNN CV with 0.69.

4.1.3 ROC Curves of the Models






ROC - Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for Viral (area = 0.89)
ROC curve for Non-Viral (area = 0.89)


ROC - XGBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for Viral (area = 0.97)
ROC curve for Non-Viral (area = 0.97)

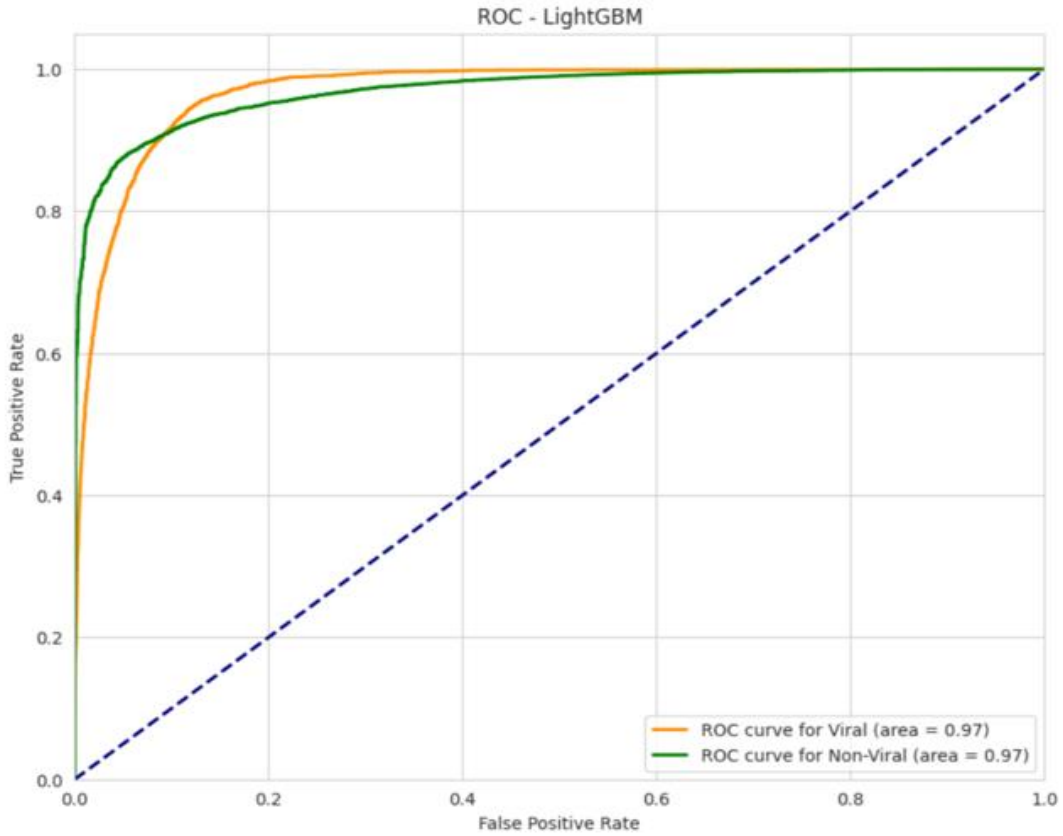


Figure 27. ROC curves of the models.

It can be noticed that according to figure above, the best performing models (related to best parameters) are LightGBM, XGBoost with ROC curve for Viral/Non-Viral 0.97 respectively, followed by Gradient Boosting with ROC curve for Viral/Non-Viral 0.95, KNN with ROC curve for Viral/Non-Viral 0.91 and last Random Forest with ROC curve for Viral/Non-Viral 0.89.

4.1.4 Model Selection

For the analysis, the LightGBM model was chosen as the best model. The choice to select this model as the final model, even though the highest average cross-validation score was not achieved, is made on the basis of its better performance in other important evaluation metrics. Especially, it is specified that the LightGBM model performed very well in the ROC analysis, thus it was proven

to have the better balance between sensitivity and specificity and in the predictive power. Besides, the LightGBM model got the highest accuracy of all models having an accuracy score of 0.92 across the 5 models.

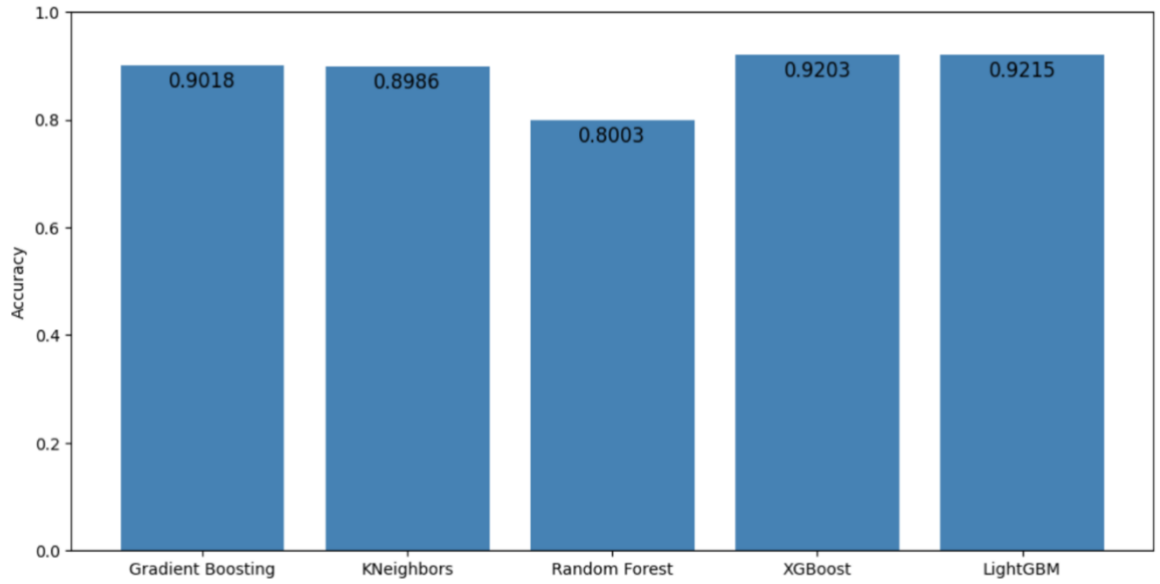


Figure 28. Accuracy scores between the 5 models.

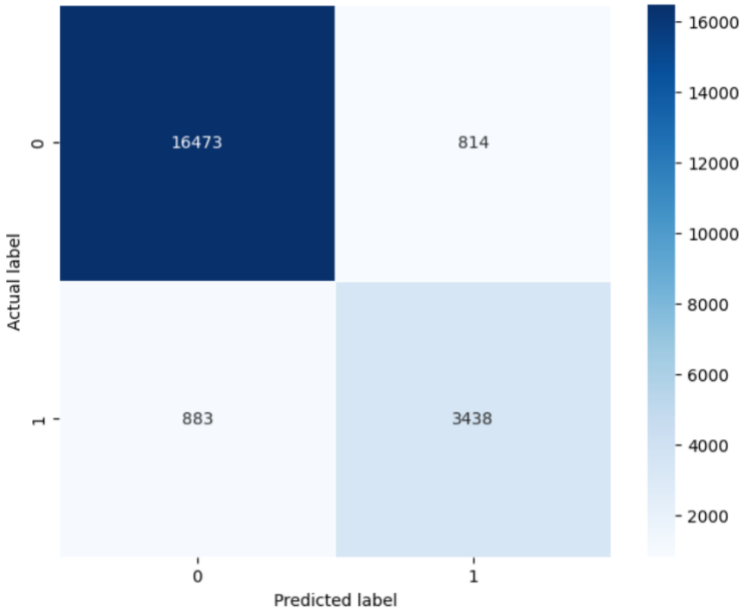


Figure 29. Confusion matrix evaluation for LightGBM.

4.1.5 Feature Importances

This figure below is the visualization of the feature importances of the LightGBM model. Each rectangle inside the treemap is a feature, and the size of each rectangle represents the feature's importance in the model's predictions.

Figure 30. Feature importances of the LightGBM model.

The feature named “subreddit_popularity” takes over the visualization as shown in the figure above, proves that the popularity of the subreddit is the most important variable in the decision of a post's virality in Reddit.

The second most notable characteristic, “num_comments”, also takes a significant place, thus, once again, emphasizing the role of user engagement through comments. Besides, thorough analysis shows that there are a number of other factors that also contribute to the success of a viral post though they are of lesser importance than the top two.

The factors such as “discussion”, “homemade”, “ama”, and the “title_length” bins are also part of the model and this shows that both the type of content and the specific characteristics of post titles are the two things that contribute to the prediction outcome.

The appearance of these components means that virality is affected by several things, such as the content type and structure of the post, for example, the title length. The layout of the treemap provides a visual representation of the feature importance, thus, allowing stakeholders to know which features the LightGBM model is most dependent on. The representation of the post virality

phenomenon in this way shows the complexity and the multiple aspects that contribute to it, such as the user engagement metrics and the content-specific attributes. Through the hierarchical representation of feature importances, the treemap makes it possible to grasp the degree of the influence of different factors in the model's decision making process.

4.1.6 Explainable AI

- SHAP

SHAP (SHapley Additive exPlanations) is a procedure that helps to explain the decisions of the machine learning models. Game theory serves as the basis of this technique and it offers a method to assign the output of a model to its input features (Aas et al., 2021). SHAP values are characterized by a uniform and theoretically sound way of giving explanations of the model predictions by spreading the contribution of each feature to the final prediction. This lets to know the features that are the cause of the model's decisions and the degree of the features which is very necessary in this study.

One of the most important merits of SHAP is that it offers both global and local interpretability. Furthermore, the SHAP values worldwide can illustrate the total influence of each feature on all the predictions. The local one can illustrate the different predictions and also make it clear what specific features that were used to make the specific output. The two distinct features of the SHAP makes it a powerful tool for model interpretation, eventually it brings the transparency and trust in machine learning applications.

SHAP Force Plot

The SHAP force plot below is a type of visualization which shows the effect of features on the prediction of a one instance by the LightGBM model for a viral post.

Figure 31. Visualization of the SHAP values for the viral post.

The plot above presents the way different features determine the shift of prediction either higher or lower compared to the base value (the predicted value without any specific feature input). In the context of this plot:

➔ Base Value: The base value (which is in the center of the plot) indicates the mean prediction of the model for the whole data. In this case it is -3.88.

➔ Positive Contributions: Elements that cause the prediction to rise (towards the right) are shown in red. It is the case that for this prediction, “subreddit_popularity” and “num_comments” are the major positive factors. These values are higher, showing that they are the major factors to the enhancement of the prediction.

➔ Negative Contributions: The non/viral-inducing features are represented in blue as they move the prediction more low (towards the left). For instance, the feature “may” is a factor which is not good for this particular situation, thus lowering the prediction.

Interpretation of the instruction of the SHAP Force Plot for the Viral Post is to extract the underlying causal facts of the post. In this particular prediction for a possibly viral post, it can be observed that “subreddit_popularity”, and “num_comments” greatly determine the estimate made by the model towards a higher value.

It is noteworthy that for Reddit “subreddit_popularity” being a high value shows that posts from more popular subreddits are more likely to go viral. The same is with the case when a large volume of comments is present on the post, the higher is the engagement which is also the best predictor of the virality. However, that the number of comments may often not be representative. And this is because, a post that is not viral (that is, it does not have enough likes), can gather too many comments in relation to a post that is viral.

Overall, the figure below represents the results in a SHAP waterfall plot for this specific viral post.

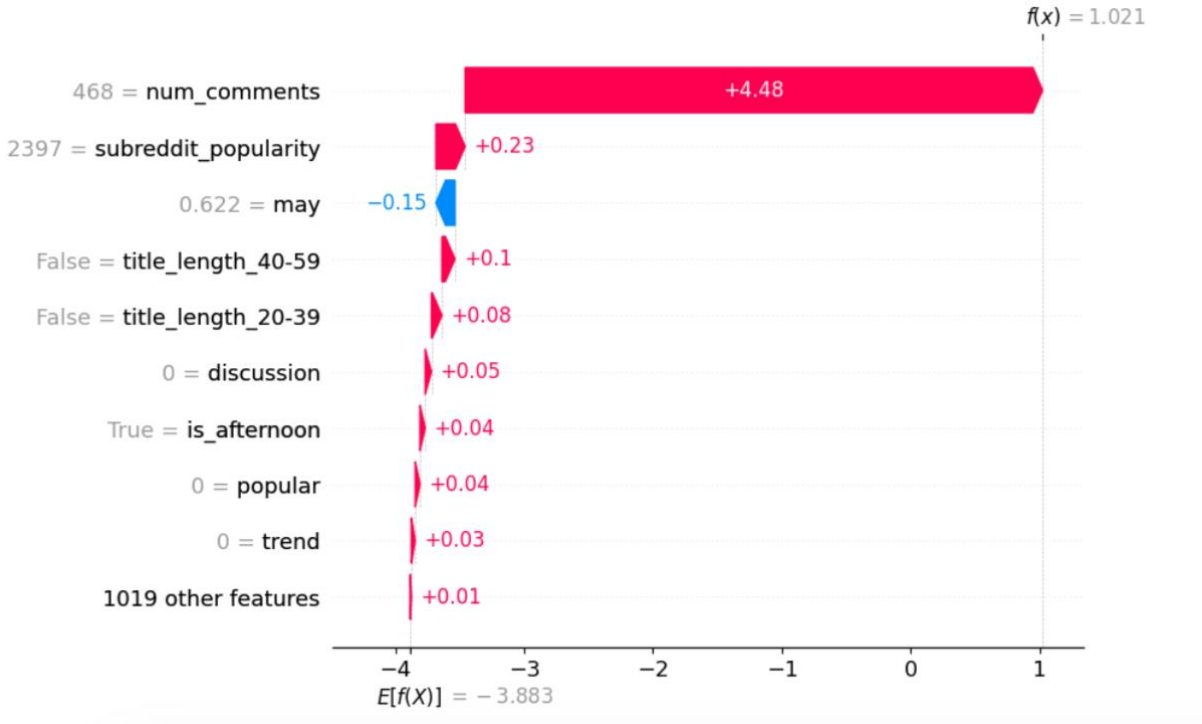


Figure 32. SHAP waterfall plot for the viral post.

4.2 Time Series Forecasting

In this section, the results obtained from the Time Series analysis will be analyzed.

4.2.1 Preliminary Analysis

Firstly, the column of ‘created_utc’ set as a datetime index. Furthermore, the data frame converted to a monthly format (resampled it by month, calculating the mean) and is mainly about ‘upvotes’ and ‘num_comments’ as presented in the table below. This conversion process amounted to accumulating daily records into the month average to capture the longer period trends and fluctuations.

created_utc	upvotes	num_comments
2006-12-31	168.750000	19.000000
2007-01-31	63.666667	26.000000
2007-02-28	5.666667	0.333333
2007-03-31	24.000000	18.000000
2007-04-30	9.000000	10.000000
...	...	...
2023-08-31	2313.863484	218.684255
2023-09-30	1939.206712	252.717773
2023-10-31	1767.113179	490.596426
2023-11-30	1401.070634	280.063440
2023-12-31	1555.841019	216.922918

205 rows × 2 columns

Table 4. The monthly data for Time Series Analysis.

The table above given presents the data from December 2006 to December 2023 with the average ‘upvotes’ and ‘num_comments’ shown for each month . It can be noticed that in the first months the mean of the monthly data is quite low, until the mean starts increase until the end of the data. As an example, the mean number

of upvotes in early 2007 from the data is relatively lower, but it has a clear increase by 2023 that might indicate growth in users' interaction or change in the content popularity. The first display on the table is made up of the time of user engagement that is measured through upvoting and commenting. These serve as the quantitative background for further analysis on the possible causes for these changes.

4.2.2 EDA

- Mean of upvotes by months, from Dec 2006 to Dec 2023

The figure below titled 'Mean of upvotes by months, from Dec 2006 to Dec 2023' exhibits the average quantities of upvotes generated on the posts in the period from December 2006 – December 2023.

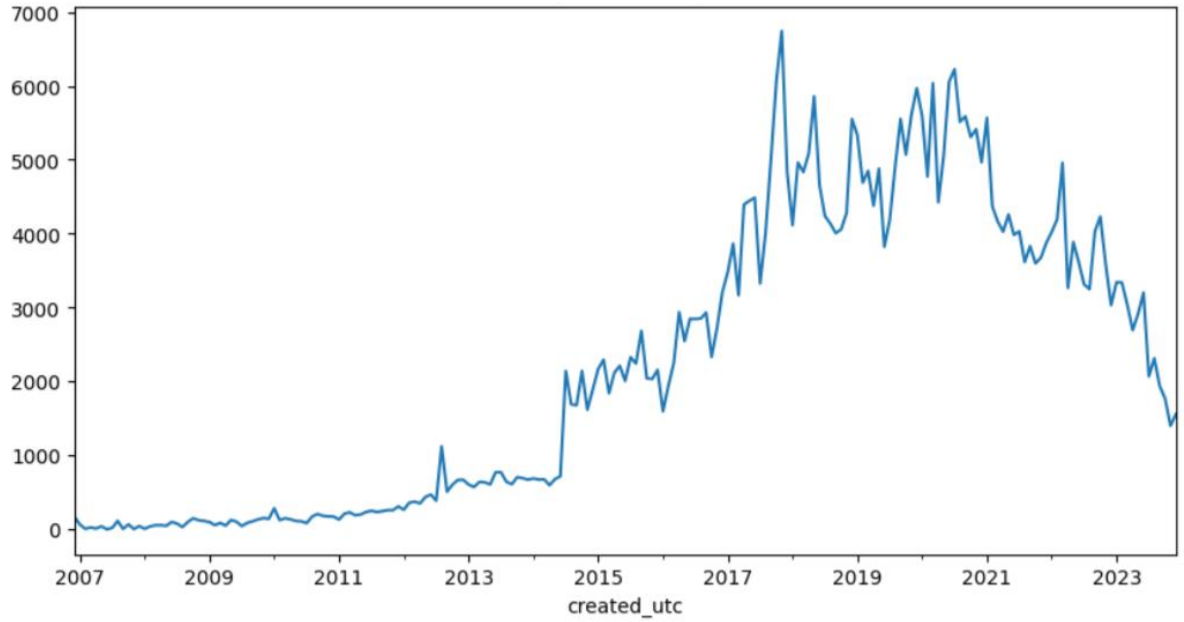


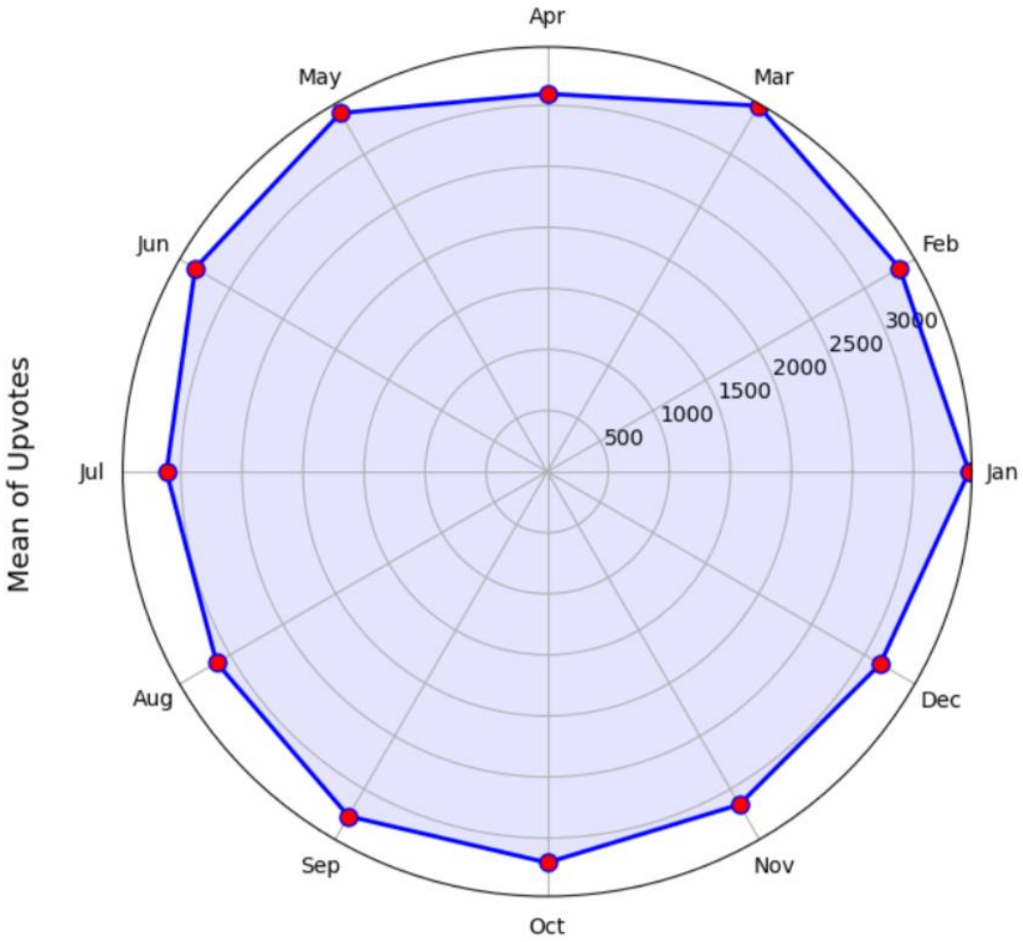
Figure 33. Mean of upvotes by months, from Dec 2006 to Dec 2023.

In the beginning, it appears that the trend is slowly increasing, that is, the mean amount of upvoted of posts. This stable condition takes place until approximately 2012, and later when the activity increases, this can be seen in terms of noticeable rise mainly in mid-2014. From that year onwards, however, there is a highly oscillatory model that can be seen on the graph as very high spikes, which obviously means that in these times, engagement was extremely high and possibly the platform dynamics or user activity changed.

The graph, from 2017 to the end of 2018, has significant fluctuations, reaching a peak mean upvotes around 2018, shows an increase in posts (e.g. popularity, etc.). This might be related to specific events or modifications of algorithms of a platform aiming to change the way posts are shown to users and influencing interactions with these posts. From 2021 until the end of 2023, the mean of the monthly figures of upvotes shows that there is a declination which can be used to speculate on changes in the engagement or of social media policies on consumption of content.

- Monthly mean of upvotes in a polar plot

The figure below shows the mean of upvotes monthly, and it helps to identify a detailed pattern of the user engagement over a year.



Month	Mean of Upvotes
Jan	3455.25
Feb	3332.99
Mar	3458.93
Apr	3094.16
May	3389.94
Jun	3329.86
Jul	3118.75
Aug	3127.45
Sep	3256.88
Oct	3200.64
Nov	3147.68
Dec	3149.31

Figure 34. Monthly mean of upvotes in a polar plot.

This polar plot above underline that the greatest number of votes is during January and March. However, an increase is also observed in the months of May and June. Therefore, users may be more active or content may be more engaging all through these months because of seasonal topics of the year, which are compelling to the communities.

The drop in mean upvotes in summer as shown in this plot is becoming more striking with July, August showing a gradual decrease. Consequently, during summer, there may be a slight fluctuation in the engagement level as the trend will plateau. The month of April is the time of the year when user engagement is the least and that could be due to the presence of the holiday season (when the users may not be that active online) (Fan, 2023; Solomon & Tuten, 2016). This data could prove very useful for content creators and marketers who seek to improve their approaches of maximizing their visibility among others by focusing on greater engagement rates during peak seasons.

- Rolling Mean & Standard Deviation

This figure below represents the rolling mean and standard deviation of monthly upvotes for the whole period. Especially, it shows the trends and the volatility in user engagement.

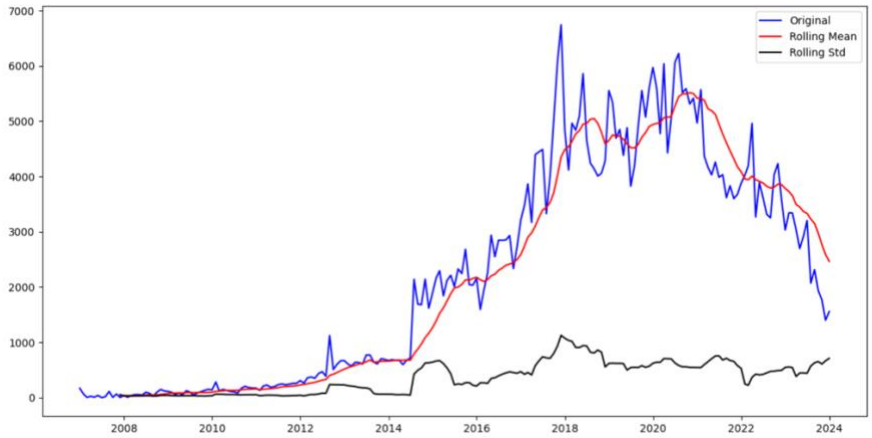


Figure 35. Rolling Mean & Standard Deviation (Time Series Analysis).

As can be seen from the diagram above, the red line represents the rolling mean, which shows the significance of the average upvotes starting from 2008 and reaching its peak in 2021, after which it gradually reduces. The black line means the rolling standard deviation, which was not changing much at first, but started to be more variable as the mean grew, which shows that the upvotes were more dispersed (fluctuated) as the user activity intensified. The general trend shows the fact that user engagement on the platform is continuously changing.

4.2.3 Stationarity Testing and Data Differencing

- Dickey-Fuller test

For the monthly series of upvotes, Dickey-Fuller test statistics showed ADF value of -1.35 with implications of non-stationarity. This is determined to be more than the critical values at 1%, 5%, and 10% significance levels as shown in the figure below. Additionally, the p-value is equal to 0.60 which is too high (more than the p of 0.05), it is usually used as a threshold. Therefore, it cannot be rejected the null hypothesis. The related findings indicates that the series depends on time and this fact should be addressed through differencing.

```
ADF Statistic: -1.3508932605916792
p-value: 0.6055792111138419
Critical Values: 1% -3.4645146202692527
Critical Values: 5% -2.8765564361715534
Critical Values: 10% -2.5747745328940375
```

Figure 36. Initial Dickey-Fuller test on the 'upvotes'.

- Decomposition

The figure of decomposition of the monthly mean ‘upvotes’ is presented below. This is into trend, seasonal, and residual elements which provide a holistic view of what the patterns be.

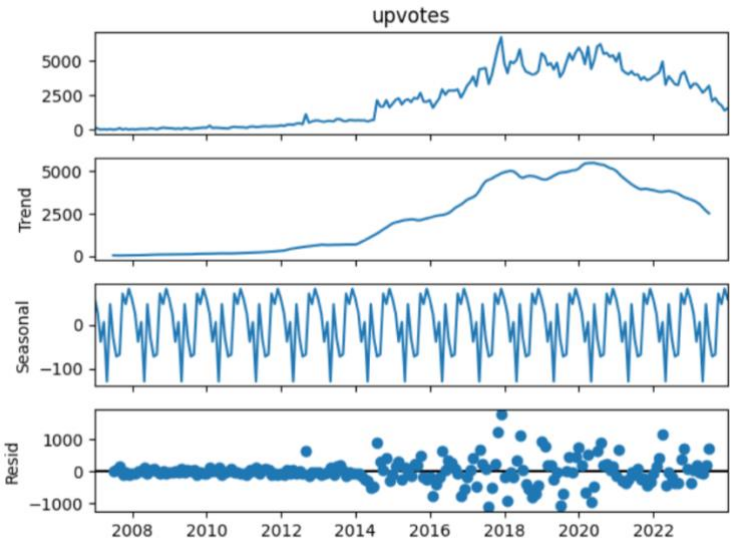


Figure 37. Decomposition of the monthly mean ‘upvotes’.

The trend component specifically indicates a trend line that progresses with time, being steady during the start of 2010s and then gradually going up towards the end of 202. This might be a sign that they have more interest to explore new functions or have a change on the algorithm on how the platform will view and popularize the released content. After that there is a reduced trend until the end of 2023.

Seasonal subcomponent represents periodical movements within every year and is seen in the form of peaks and troughs in which possibly users activity happens during some times when user activity is more and during sometimes it is less. Such events and behaviours may be identified as methods of measuring web traffic in terms of cultural, holiday or seasonal patterns.

Related to residuals, the remaining upon decomposition though is indicative of any concentrate of dropouts that were not noticed from the trendline or seasonal components. The units of residuals represent the random deviations around zero. This implies that the regression model does a good performance in capturing the repeated patterns, but there may still be random unexpected errors which might stem from outside factors or scattered data.

This analysis above provides the basis in understanding what the different factors of this trending are and how they affect it in its whole on the platform.

Detailed decomposition makes the whole process of diagnosis for the data so easy and gives the planner targeted strategies for constructing interactions between users based on their cognitive and emotional states. This step by step analysis is helpful for predictive modeling because it results in adjusting models to be more precise and provide better forecasting regarding the future activities based on the found patterns.

- Differenced series graph plot with Dickey-Fuller test results

The figure below shows the differenced monthly mean of upvotes. It is used this technique because in time series analysis takes the trend and cycles components out of the data to make it stationary. The reliability of this tool is compromised as the visible fluctuations represent the changes from one month to the next and not the absolute values. This approach brings those spikes and declines to light which,

specifically, could suggest seasonality tendencies or other activities contributing to the user engagement.

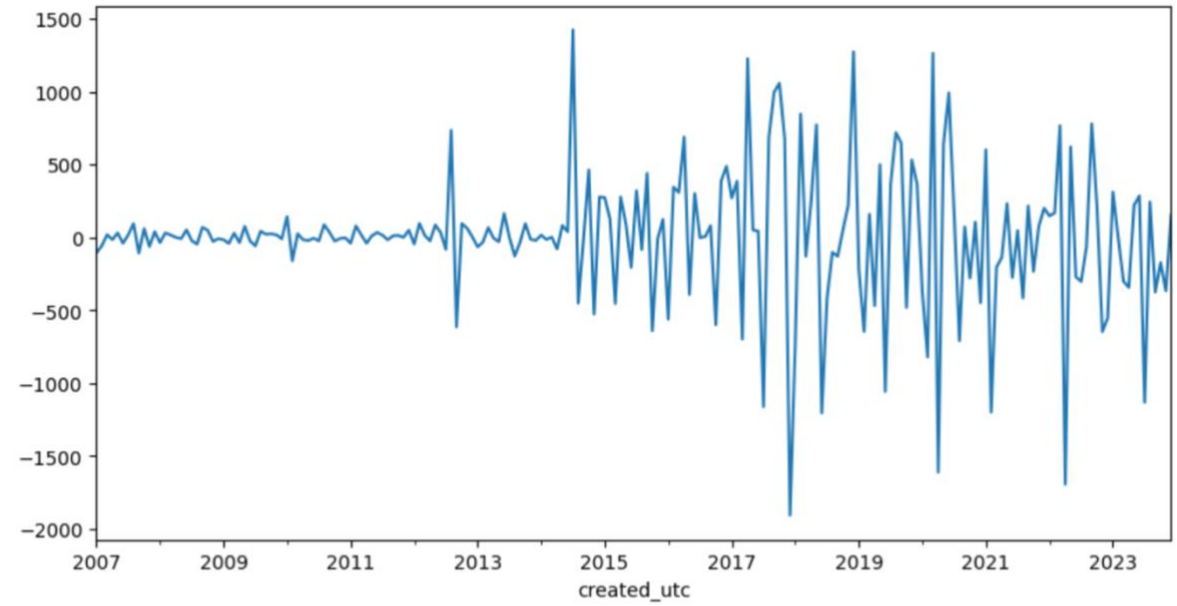


Figure 38. Differenced series with Dickey-Fuller results.

The results from the Augmented Dickey-Fuller test in the above figure show that the data is stationary after differencing. The value of -3.14 of the ADF Statistic and the p-value being 0.023 means significance level of rejecting null hypothesis, a unit root, is at the 5% level. This proves that the data, after differencing, simplifies the analysis further and forecasting without the need to deal with any trends within the data (Huang & Petukhina, 2022). Therefore, this knowledge is valuable in that help to accurately model time series in order to predict the future trends in the upvotes based on the historical data.

- Autocorrelation and partial autocorrelation (ACF – PACF)

Below are presented the plots of the ACF and PACF produced by the differenced monthly upvotes data.

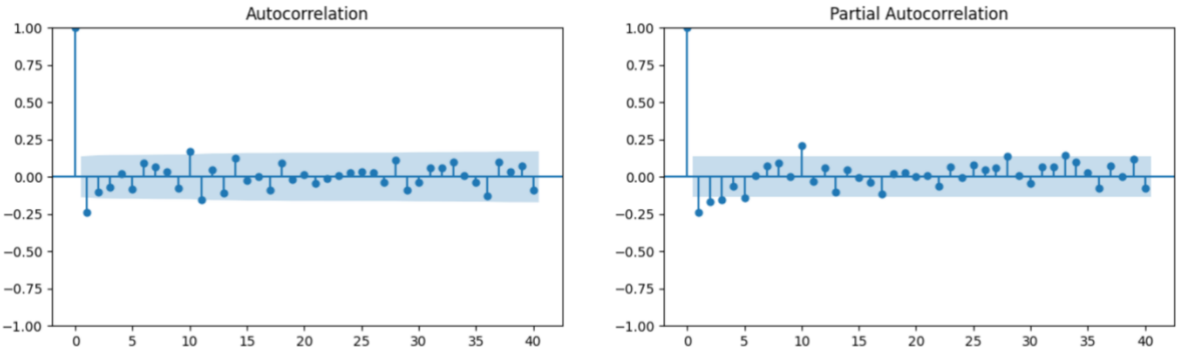


Figure 39. ACF and PACF plots of differenced monthly upvotes.

In the ACF plot, where there is decrease after the first lag and the most subsequent values staying within the confidence interval. This presents that the differencing has been effective in removing the autocorrelations from the time series (Huang & Petukhina, 2022). Thus, the differentiation process seems to have achieved its goal to stabilize the mean of the series and the exactness of the forecasting methods depends on the assumption which is stability.

The PACF plot reveals this important feature of an AR series, which is the significant peak at the first lag. The AR(1) model might be an appropriate fit for the differenced data as the spike at lag one in the PACF plot indicates that the process consists essentially of a randomly selected term and does not involve any terms at higher order. These hints are of use in modeling the data either through ARIMA or SARIMA models, which impose the correct order of autoregressive terms needed for seizing the patterns in the data accurately.

4.2.4 ARIMA/SARIMA Modeling and Validation

- Data splitting

First, the 205 monthly data (from 2006 to 2023) will be split, as follows, the last 12 for the test (of 2023) and all remaining 193 in total for training.

- Auto-model

```
Best model:  ARIMA(2,1,2)(2,1,0)[12]
Total fit time: 148.907 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  193
Model:             SARIMAX(2, 1, 2)x(2, 1, [], 12)   Log Likelihood               -1372.269
Date:                             Mon, 29 Apr 2024   AIC                           2758.538
Time:                                     00:41:00   BIC                           2780.888
Sample:                                 12-31-2006   HQIC                          2767.600
                                      - 12-31-2022
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3312      0.053     25.237      0.000       1.228       1.435
ar.L2         -0.6780      0.056    -12.068      0.000      -0.788      -0.568
ma.L1         -1.6838      0.038    -44.564      0.000      -1.758      -1.610
ma.L2          0.9000      0.036     24.682      0.000       0.829       0.972
ar.S.L12      -0.5974      0.054    -11.002      0.000      -0.704      -0.491
ar.S.L24      -0.2586      0.057     -4.568      0.000      -0.370      -0.148
sigma2      2.383e+05   1.83e+04     12.993      0.000    2.02e+05    2.74e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                42.31
Prob(Q):                              0.76   Prob(JB):                         0.00
Heteroskedasticity (H):              30.40   Skew:                             0.03
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.37
===================================================================================
```

Table 5. SAMIRAX results using auto-arima.

The results of ARIMA(2,1,2)(2,1,0)[12] for SARIMAX model has the capacity to predict with efficiency upvotes. The historic data starting from the month of December in 2006 through December in 2022 (it is the train data) from which the model correctly identifies regular and seasonal patterns in the data. The presence of significant coefficients and confirmation by model diagnostics suggest that this is a good model with a high degree of predictive ability. As for the final model, it can be seen that this structure is a reliable framework that is used in the disposal

of accurate forecasts which are of help to the future trends and forecasts of upvotes.

- Model diagnostics

The diagnostics results for the SARIMA model offer a broad coverage to analyze the model operation and effectiveness.

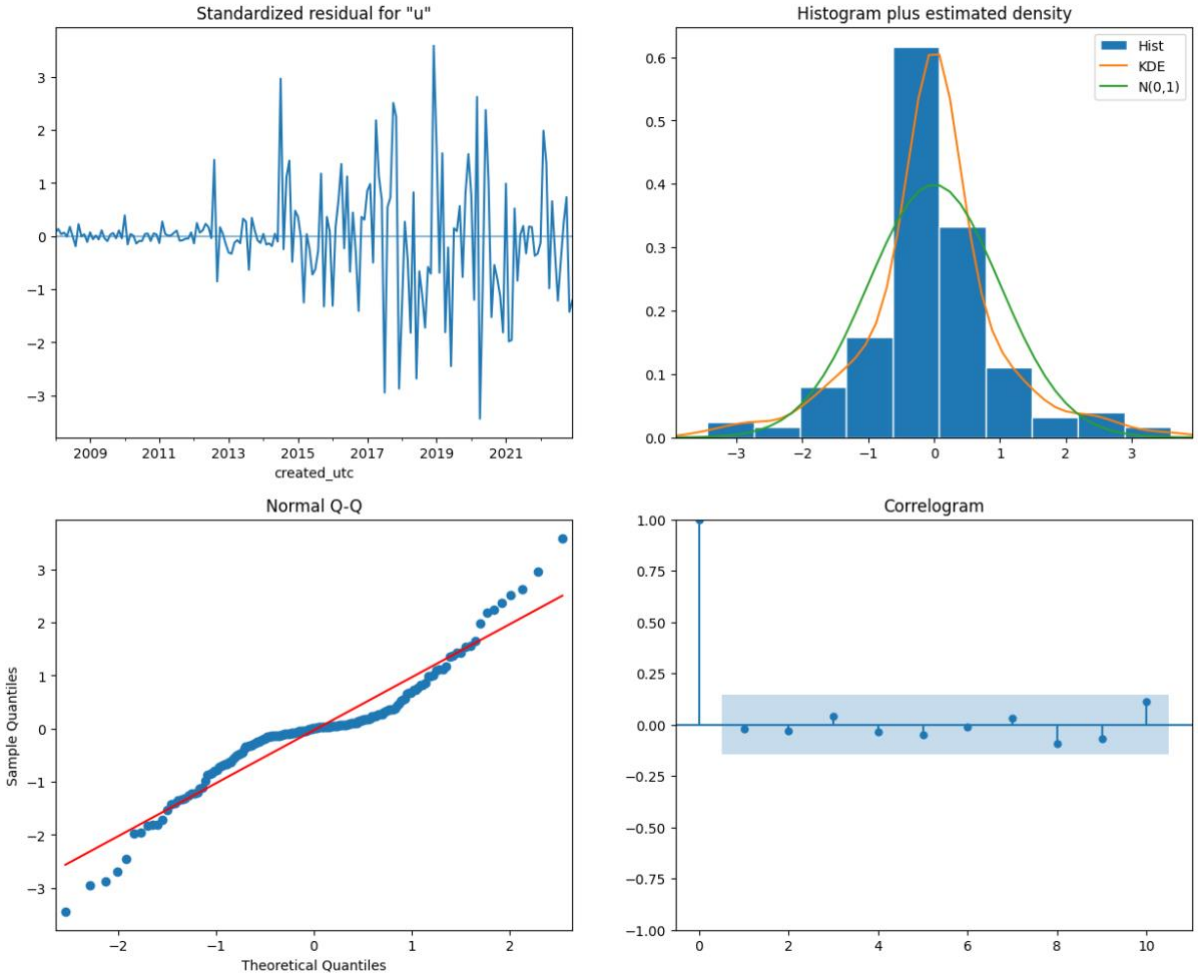


Figure 40. Model diagnostics of Sarima.

➔ The graph in the left-top corner shows the standardized residuals over time (they should resemble white noise and if the model fits then well it will be the case). The residuals basically wiggle around zero until 2012, with no structure pattern and this tend to indicate that that the model fitted the data well by its main features. But in many cases, these are only related to the periods during which the volatility spikes, for example from 2015 onward, which can indicate the possible periods of model misfit or that the model does not take into account the changes that take place in data generation process. Over this period from 2015 to 2023, significant fluctuations are observed.

➔ Furthermore, the graph in the right-top, the histogram of the residuals together with a Kernel Density Estimate (KDE) and the normal distribution curve give an insight into the distribution of the residuals. The histogram indicates a quite symmetrical distribution around zero, whereas the KDE matches with the shape of normal distribution curve. It is the visual realization of this assessment that is supported by the green normal distribution curve that generally goes together with the histogram and KDE, because it indicates the residuals have approximately normal distribution. This type of distribution is preferred to many other types of distribution as many of the inferential statistics are based on normality assumption.

➔ The normal Q-Q plot is at the lower-left corner, and it would be comparing the quantiles of the residuals to the expected quantiles of a standard normal distribution. When the data lies along the red line, the model is considered as a well fitted. Here Q-Q plot shows reasonable alignment with the line (but not absolutely). However, there are deviations towards the tails, especially in the extreme values. So, this implies that there might be certain

difficulties concerning normality at the tails which are stating that while most of the residuals follow the normal distribution, some outliers deviate from this rule.

➔ The bottom-right plot depicts the correlogram (where there is the estimation of the residual autocorrelation). It is noteworthy that, the residuals show no significant autocorrelation, which describes that it is uncorrelated between residuals of each time point. Especially, the correlogram over here illustrates that the majority of autocorrelations are in the blue confidence intervals lines suggesting there is no significant autocorrelation present. Absence of noticeable autocorrelation thus does reflect that the model has really accounted for the time-dependent structure of the data hence leave residuals that are independently distributed.

Overall, these diagnostics suggest that SARIMA has done a good performance of fitting the data by capturing main temporal patterns (without retaining significant autocorrelated residuals).

- Sarima Forecast with the test data

The graph below named “Sarima Forecast with the test data” clearly demonstrates the success of the SARIMA model in the prediction of the monthly upvotes from a historical data that was collected for a long period of time.

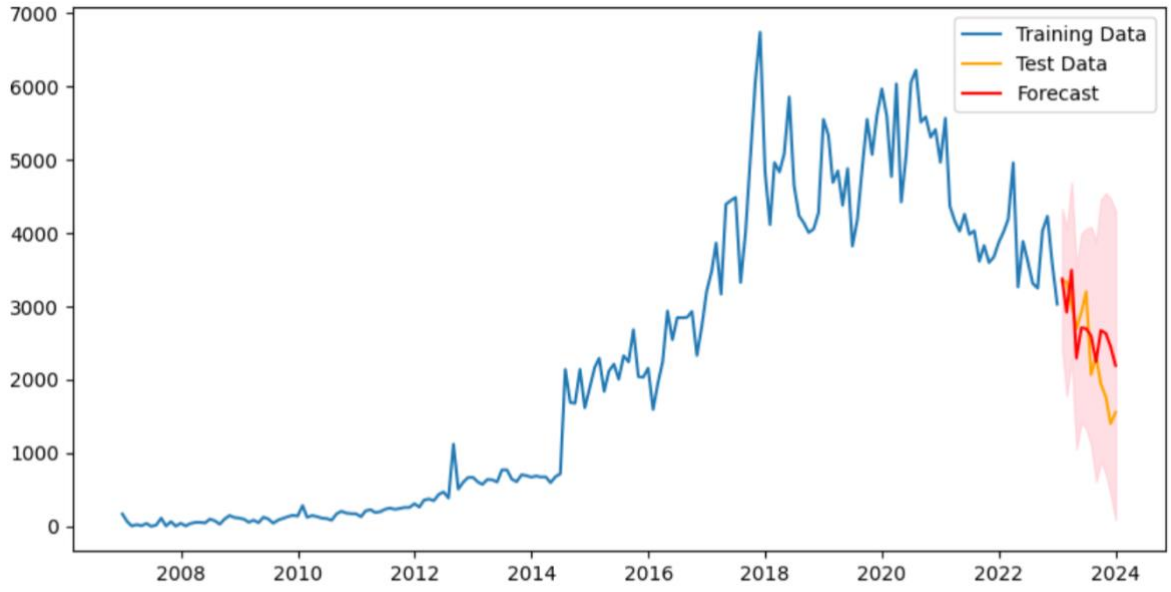


Figure 41. Sarima Forecast with the test data.

The blue line depicts the training data which is the data where the model discovers the hidden trend and the seasonal patterns, and the orange line is the test data which is the real one analog of the model evaluation. The red line and shaded area presents the forecast of the future values and their confidence intervals. The visual comparison results in the model's high capability of grasping the past data trends and also of using these trends to predict the future user engagement, therefore, the model's practical application. The match between the forecast and the actual data around the transition point from training to testing especially the model's accuracy and the future predictions' reliability is very effective.

Furthermore, the figure below, in a simple way, displays the comparison between the actual and the predicted upvotes using the SARIMA model in the test period, hence, it shows the model's ability to describe the trend of the user interactions. The blue line is the actual upvotes, whereas the red line is the forecast (as presented), and the red shaded zone is the confidence interval, which thus shows the forecast's reliability and the potential variations in predictions. This visualization confirms the model's accuracy in the user behavior trends and at the same time, it points out the lay hatch transitions at the peak and the trough transitions that means that the model needs to be fine tuned in these areas to enhance its precision and use in the strategic content management.

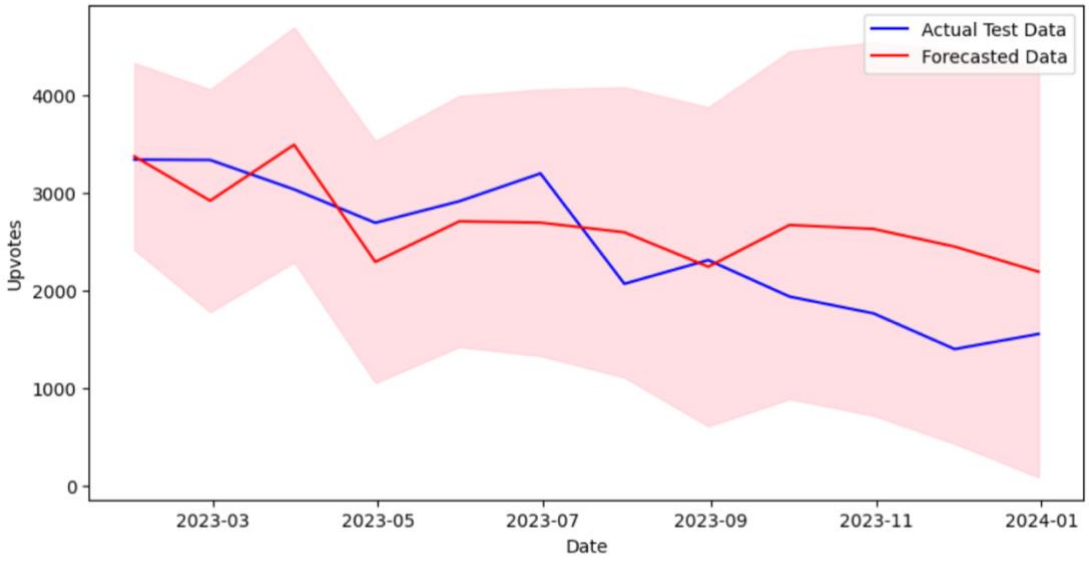


Figure 42. Comparison between the actual and the predicted upvotes.

The table below contains the metrics (RMSE = 571.72, MAE = 491.80, MAPE = 24.49%) that show the evaluation of the model's performance in predicting the upvotes. The values might seem a little high, though this is because of the high volatility and unpredictability in the social media engagement data, especially in the case of predicting upvotes which are affected by various shifting factors. A larger error in predicting might also show the complicated dynamics of online

interactions rather than the lack of the model, which means that the model is indeed catching and reacting to the deep complexities in the data thus, it is creating the possible insights in spite of the difficult prediction task.

```
Root Mean Squared Error: 571.7250287924275
Mean Absolute Error: 491.796176820688
Mean Absolute Percentage Error: 24.786487221903723%
```

Table 6. Accuracy metrics of forecast.

4.2.5 Advanced Forecasting with Prophet

- Forecast graph using Prophet

This figure below is the plot of the Prophet forecasting model and it illustrates the historical data as well as the predictions for the upvotes on social media posts, from December 2006 to the end of 2024 (including the twelve-month forecast).

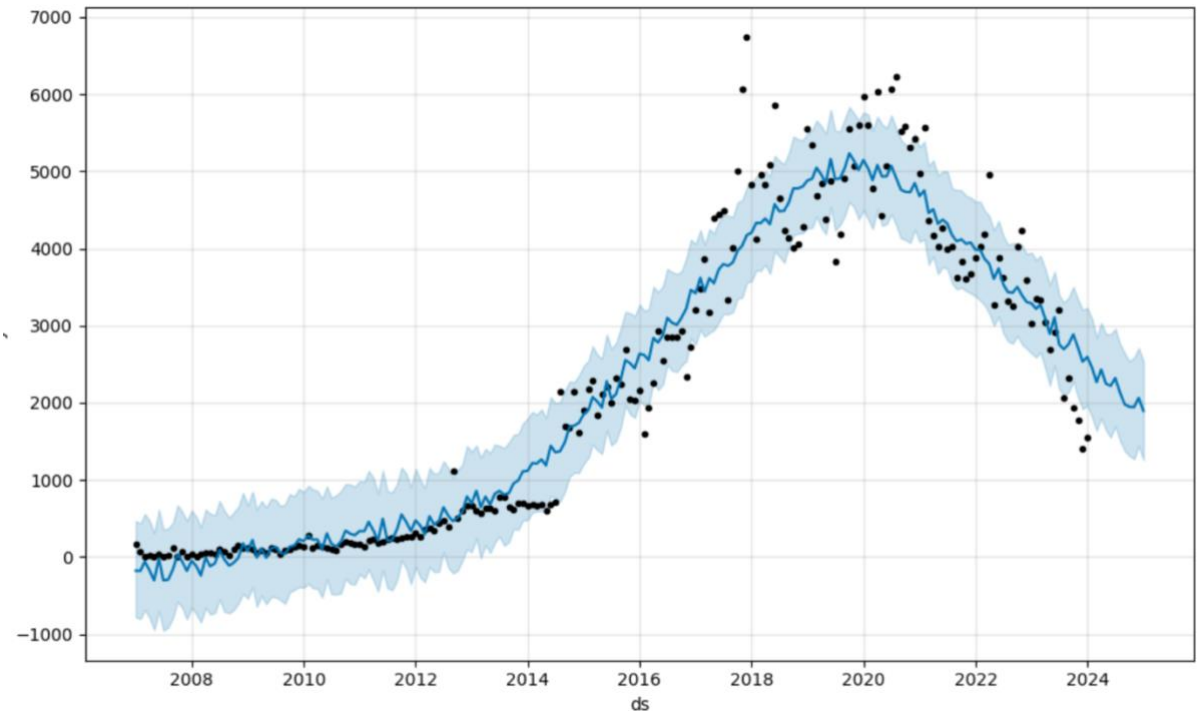


Figure 43. Forecast graph using Prophet.

As shown in the graph above, the model shows a definite upward trend until about 2020, then a decline is observed, and after that there is a decline expected (as forecast). The shaded area is the uncertainty interval and it is the graphic representation of the prediction's confidence and the range of the values that are expected.

This prediction is a great case for the understanding of the seasonal patterns and the long-term trends in user engagement which, afterwards, helps the strategists and marketers in the preparation for the future variations in the interaction rates. The features of the Prophet model which are the seasonality, holidays, and additional regressors are the reason why it can make accurate forecasts in data that has more than one seasonal pattern, thus, the plotted data is very much detailed and shows the trend and fluctuations.

The black dots which represent the existing data points and the curve for the forecast show the model's efficiency in its historical data fitting and future predictions. The predictive insight is the crucial part for the preparation of resources and the strategic planning of social media content management that allows stakeholders to have a perspective for the periods of high user engagement (that are predicted for the coming months).

- Forecast graph using points in Prophet

The figure below presented provides the illustration of the major trends change in the upvote patterns over the time, which are marked by the vertical dashed lines that are the sign of the major changes which are the results of the Prophet forecasting model.

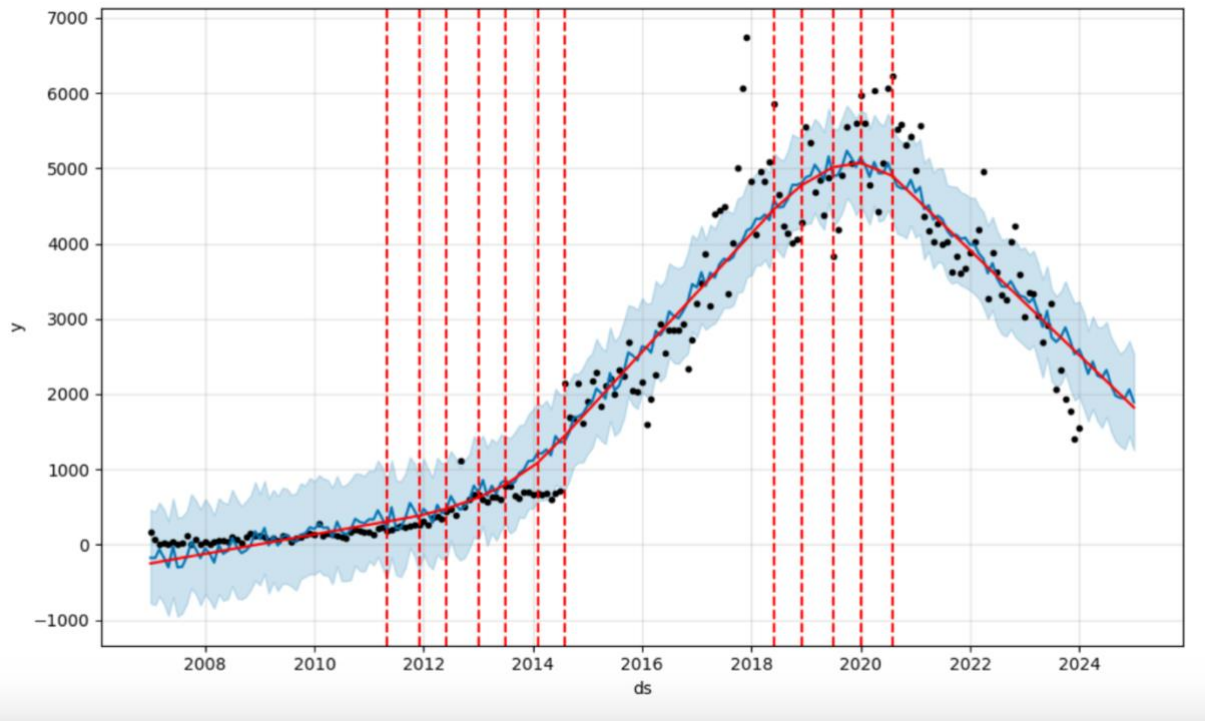


Figure 44. Forecast graph using points where the trend changes noticeably in Prophet.

As show in the above figure, the red markers in this graph illustrate the crucial turning points where the trend of upvotes either increases or decreases, thus, giving a clear idea of the user engagement shift either high or low. Those are the important clues that it is important to understand how the events outside the platform or the change in the platform dynamics can change the user interactions over time.

The trend line shows the seasonal changes in the upvote activity through the annual cycle (that is at the same time of the year). This scheme shows the repetition of certain behaviors that are seasonal in nature, probably due to the cultural or social events. Through this analysis of this seasonality, it can be predicted and be prepared for such periods when user activity is anticipated to be high or low, hence, the content and interaction strategies can be made more effective. Thus, this whole-process analysis tool enables the stakeholders to gain a deeper understanding of when and why the upvote patterns change. It also makes the strategic planning of the content and the engagement more effective.

- The forecasted 12 months for the year 2024 using Prophet

The figure below depicts the forecast of the next 12 months, and it shows that the upvotes will probably decrease slightly for 2024. This forecast (which is the result of the Prophet tool), shows the user engagement range that is usually consistent with a shaded area that represents the uncertainty interval that covers the possible fluctuations.

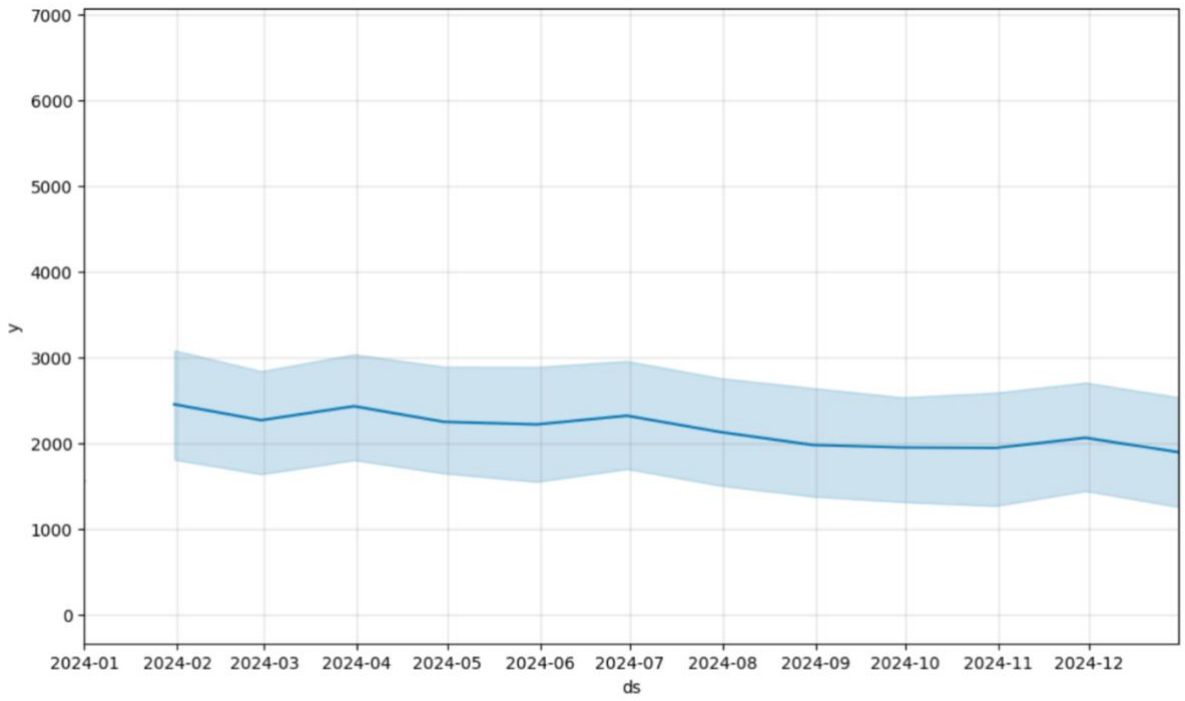


Figure 45. The forecasted 12 months for the year 2024.

- Components of the forecast

In this section, two diagrams of forecasting components using Prophet will be presented.

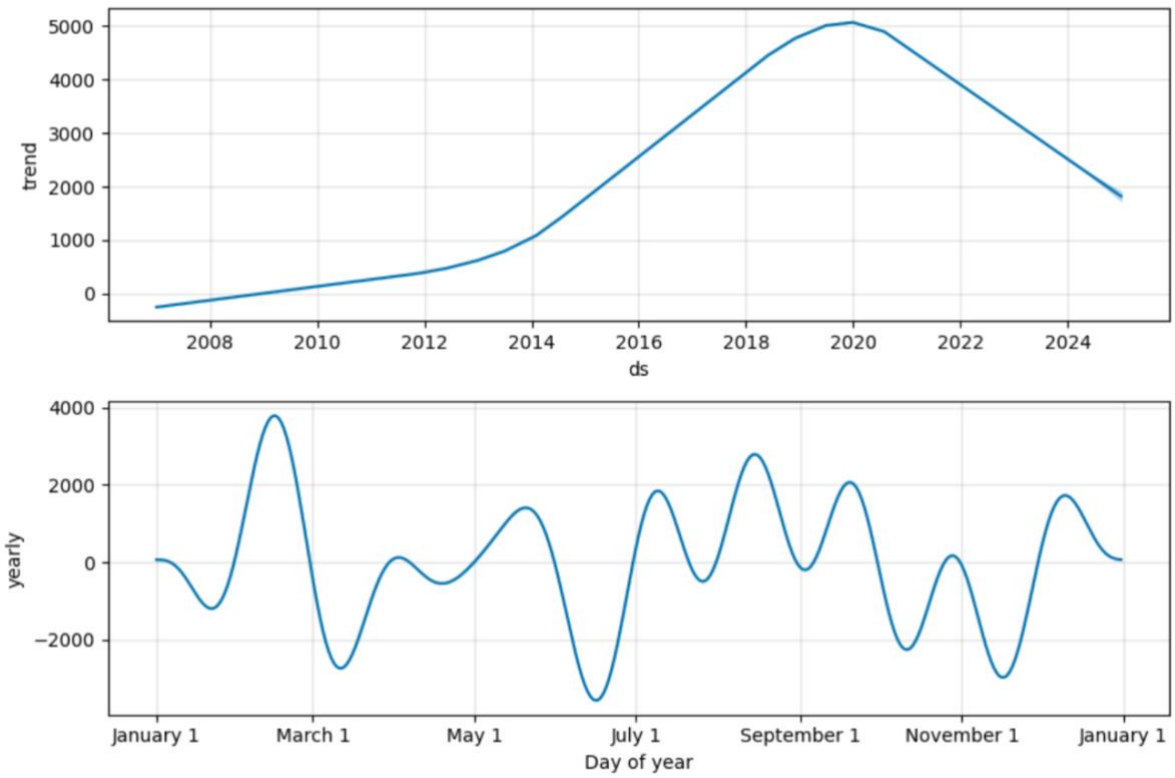


Figure 46. Components of the forecast using Prophet.

The first graph in the upper part of the figure shows the general trend component obtained by the Prophet prediction tool, that is, the pattern in the upvote data over years (historical data as known is through December 31, 2023, and the forecast is for 12 months in 2024). This trend graph shows a great increase of upvotes from 2008, reached the maximum level around 2020, and then it is going to decrease in 2023 and later on. This may be because community activity is getting bigger and bigger as the years go by and it might be the case that the community activity is evolving to its normal level or getting saturated because the community is very mature.

The second graph is the example of the seasonal fluctuations in the data that depicts the annual pattern that is used throughout a year. Regarding the prophet model, a great increase in upvotes is observed each year in February, while downward trends are seen around March, June and November. This pattern

could be the evidence of seasonal user behavior, which may be the result of the user activity during specific periods of the year. For instance, cultural events or social events, or simply the user has more free time and being interested in certain activities during the seasons. This broad explanation facilitates the comprehension of when to anticipate the higher or lower engagement, which gives useful information for the purpose of strategic planning and content planning.

- Performance metrics of Prophet

The figure below shows the performance of the Prophet model with the MAPE for the 365-days horizon. The key point of this metric is the forecast accuracy evaluation, which is very important for the understanding of the model's performance on the data as the forecasting period is extended.

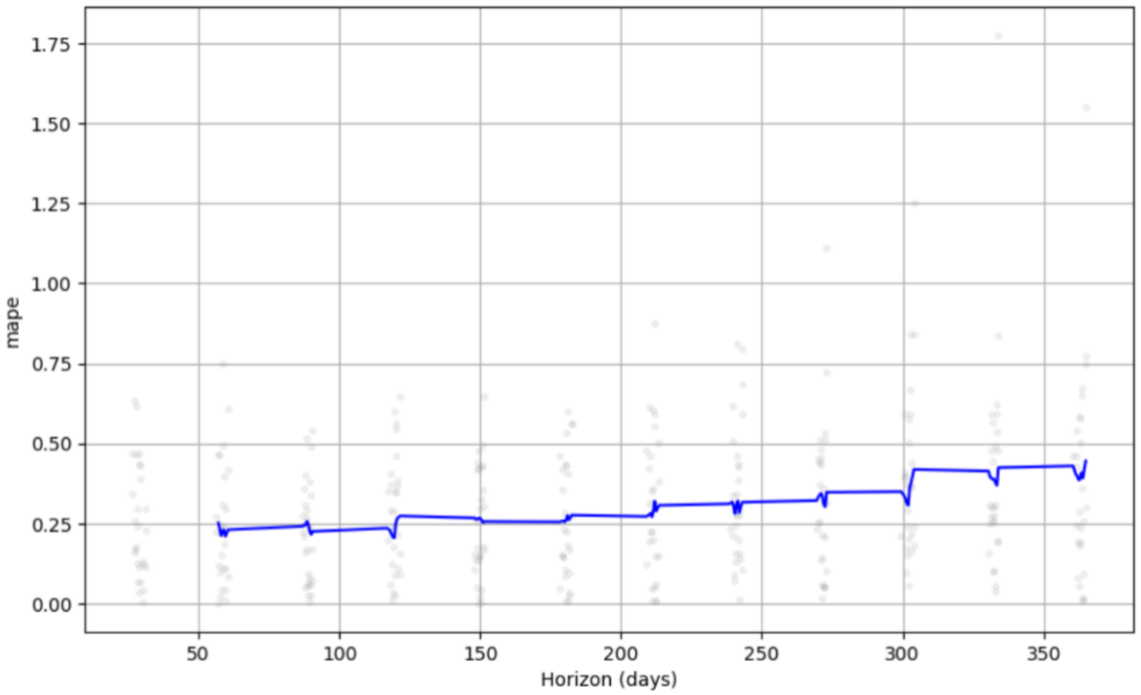


Figure 47. Performance metrics of Prophet

This line plot above shows that the model keeps a low error rate all year, which means that the model is able to catch and predict the trends in the data over different seasonal cycles and the error does not increase significantly. The level of the flat forecast line throughout the period implies that the model can be trusted to make predictions with the same accuracy, no matter much the length of the forecast. It is noteworthy that the stability in forecasting of these events is very important for strategic planning and the decision-making process.

5. Discussion

5.1 Interpretation of Results

The outputs from the predictive models and time series analysis give several clues about the dynamics of the viral contents on Reddit. The machine learning models, particularly LightGBM, demonstrated to be very good in the detection of the features that have a great impact on the virality of a post.

Prominently, subreddit popularity, and the number of comments turned out to be the most important factors. This shows the significance of the community involvement and the power of the certain subreddits in spreading the content virality. High subreddit popularity means a bigger, more active audience, thus enhancing the chances of a post to be seen, upvoted and shared. It is worth noting that the length of the titles also plays a decisive factor in making something viral, according to the results obtained. As for example, (title length = between 20 and 39) feature, proved to be the most important factor compared to other title lengths.

The EDA analysis discovered the obvious user engagement patterns. Weekend posts (and especially on Saturdays), those made in the afternoon and evening hours (12 PM to 7 PM UTC) received the most upvotes on average. Thus, it is obvious that the timing is the most important factor in the visibility and the engagement of the content.

Furthermore, the word cloud analysis of the viral post titles has revealed the significance of the certain themes and topics which the user actively engages more. The main topics on these posts were the current events, personal opinions and emotional triggers. This goes hand in hand to psychological studies on content sharing, which provides that the emotionally engaging content is the one that is more likely to be shared. Posts that are designed to appeal to the audience emotionally, are related to current and upcoming topics or offer personal views

are more likely to get the audience's attention and thus they will share, hence the posts will be more viral.

For time series analysis, ARIMA and Prophet models in terms of efficiency in detecting seasonal patterns show that the temporal features are of great importance in the content virality analysis. Especially, those models are the best to show the times of high engagement which can be used by marketers to plan their posts wisely. Through the foresight of the future, the producers can foresee the moments of the year in which the users are the most active and that is the early spring and the late fall. This, in turn, they will be able to plan the content releases in a better way and thus increasing the reach and the engagement of the users.

5.2 Comparison with Existing Literature

The study results are in accord with the previous research that shows the importance of emotional involvement and community dynamics in the content virality. For instance, (Tellis et al., 2019) underlined that the content that is related to strong positive emotions is more likely to be shared. This thesis analysis proves that the assertion is right, since it is shown that positive or neutral posts get more upvotes than the ones with negative sentiments. This connection implies that emotional resonance is a vital element of the process of users to let the content be shared and circulated.

The great importance of subreddit popularity in the models confirms the findings of (Joseph et al., 2021), who stressed the significance of network topology in the information propagation. The subreddit popularity being in the center of attention means that the communities with the bigger and more active ones are the ones where the content can get the visibility and traction. It is the

same as the theory that highly interconnected networks can speed up the spread of information hence the possibility of the content going viral will be on the increase.

The high significance of the number of comments is in line with the research of (Reichstein & Brusch, 2019), who investigated the user decision-making processes and the elements that contribute to the phenomenon of viral marketing. It is noticed from the study that posts which are widely discussed and generate a lot of discussion are more likely to go viral, thus, it is clear that the more interactive and engaging the content is, the more it will be the reason for user participation which in turn will increase the visibility and the shares.

Nevertheless, the results contradict some of the previous studies. In the given example, (Xu & Qian, 2023) studied the structural complexity of user interactions, whereas the results of this thesis show that the simpler metrics, like the subreddit popularity, can be the strong predictors of virality. This difference of opinions demonstrates the complexity of content virality and shows that even though the interaction patterns are the main factor, the simple engagement metrics can be as important as the complex interaction patterns as well. It is noteworthy that the focus on the timing and thematic elements in this study provides another perspective on the content virality issue.

Besides the research that has been usually concentrated on network dynamics and emotional triggers, this study underlines the significance of the posting time and the topics to be posted, like current events and personal opinions. The wider view that is emerging accentuates the necessity of a comprehensive approach to the study and use of content virality.

In general, this thesis supports many of the well-known theories while at the same time, it is introducing new aspects to the understanding of the factors that make a content go viral. Through the inclusion of the emotional connection, community interaction, the timing, and the thematic relevance, it is offered a

complete framework that can be used by content creators and marketers to optimize their strategies for the maximum impact.

5.3 Practical Applications

The information obtained from this thesis has a lot of possible uses for content creators and marketers who want to increase the level of engagement on social media sites. To start with, one can learn how the popularity of the subreddit is related to the engagement metrics like comments from which the creators can formulate their strategy to target their content more effectively. Through the use of highly engaging subreddits, content creators are able to pass on the message thus achieving virality.

More than that, the right moment is also important. Especially, the analysis demonstrates that posts made on the weekends and the ones made in the afternoon and evening hours are the ones that usually get more upvotes (as previously mentioned). Creators and marketers will be able to use this data to plan their posts at the right time so that they will be seen and liked by more people.

The word cloud analysis shows that there are some themes and topics that are more likely to be forwarded. Sentences that appeal to the emotions of the readers, talk about current things or provide personal opinions are the ones that are most likely to be discussed. Marketers can employ this information to create the messages that will make the people to feel a certain emotion or will relate to the trending topics, thus the content they post will be shared more often.

The seasonal trends figured out through time series analysis give a lot of useful information for the long-term content planning. To give an example, the marketers can anticipate the highest engagement periods in specific months such as from January to March and organize their campaigns accordingly. Such an

arrangement of the schedule can be made sure that the content is published when the user activity is at its peak which will in turn will result in the maximization of the reach and the impact of the content.

Besides, the use of the latest machine learning models such as LightGBM and time series forecasting tools such as Prophet can be of help to the content creators and marketers to predict the virality of their posts more correct. Through the repeated analysis of the engagement metrics and the changes in the strategies based on the predictive data, they can improve the content's performance in the long term.

Finally, the SHAP technique for the generation of explainable AI creates an extra layer of interpretation. Through the process of distinguishing which features are the most important for the model's predictions, creators can further develop their content strategy to concentrate on the most potent elements. To illustrate, if for instance subreddit popularity and the number of comments are the critical predictors, content creators will give priority to the engagement with their audience through comments and focus on popular subreddits. Also, at the same time they will target a specific title length (as previously mentioned) to enhance even more their content.

6. Conclusion

This thesis provided an insight into what exactly is it that makes content viral and how many factors such as the popularity of subreddits and engagement metrics like comment volume influence the virality of content on Reddit. Content posted during weekends and evenings was observed to have more total engagement, signifying that timing, as a factor, has a very significant impact on content visibility. In addition, personalized and emotional connection with current events or personal opinions have high-level effect on the virality of the content. These factors were strong indicators of how a post would perform on getting a lot of attentions and engagements which were the determinants of its popularity on the platform.

This research gives actionable insights to be used by content creators and online marketing specialists. Through use of the trendy subreddits and timing the posts at the right moment of the peak users' activity, the marketers have a chance of increased visibility for their content. Moreover, production of content that relates to the feelings or the general user trends can allow content to be shared at a very high rate. The method is optimizing the content to be engaging and also adapting it to the character of social media content consumption

Even though the research has brought a new insight, a major limitation is that it is Reddit-centered and a case of other social media platforms may having different user activities and content delivery mechanisms.

Overall, this thesis opens the way to analyse how fast-developing technologies such as the artificial intelligence and machine learning could help with the prediction of the virality of a content more accurate, providing digital marketers a toolset for preemptive content strategy. This shall help to diversify the research, and also provide useful experience in practical assignments.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

Alassafi, O. M., Alghamdi, W., Naveena, S. S., Alkhayyat, A., Tolib, A., & Muydinjon Ugli, I. S. (2023). Machine Learning for Predictive Analytics in Social Media Data. *E3S Web of Conferences 399(2009)*.

Anugrah, I. G., & Fakhruddin, M. A. R. I. (2020). Development Authentication and Authorization Systems of Multi Information Systems Based REst API and Auth Token. *INNOVATION RESEARCH JOURNAL*, *1*(2), 127. https://doi.org/10.30587/innovation.v1i2.1927

Arjona-Martín, J.-B., Méndiz-Noguero, A., & Victoria-Mas, J.-S. (2020). Virality as a paradigm of digital communication. Review of the concept and update of the theoretical framework. *El Profesional de La Información*. https://doi.org/10.3145/epi.2020.nov.07

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321–335. https://doi.org/10.1016/j.ijpe.2015.09.039

Beysolow, T. (2018). *Applied Natural Language Processing with Python : Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing*. http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=edsebk&AN=1892182

Chan, I. C. C., Chen, Z., & Leung, D. (2023). The more the better? Strategizing visual elements in social media marketing. *Journal of Hospitality and Tourism Management*, *54*, 268–289. https://doi.org/10.1016/j.jhtm.2022.11.007

Chen, G., Kong, Q., Xu, N., & Mao, W. (2019). NPP: A neural popularity prediction model for social media content. *Neurocomputing*, *333*, 221–230. https://doi.org/10.1016/j.neucom.2018.12.039

Chen, Y. (2023). Comparing content marketing strategies of digital brands using machine learning. *Humanities and Social Sciences Communications*, *10*(1), 57. https://doi.org/10.1057/s41599-023-01544-x

Chi, Y., & Chen, H. (2023). Investigating Substance Use via Reddit: Systematic Scoping Review. *Journal of Medical Internet Research*, *25*, e48905. https://doi.org/10.2196/48905

Dašić, D., Vučić, V., Turčinović, Ž., & Tošić, M. (2023). DIGITAL MARKETING - MARKETING OPPORTUNITIES AND THE POWER OF DIGITAL CONSUMERS. *Ekonomika Poljoprivrede*, *70*(4), 1187–1199. https://doi.org/10.59267/ekoPolj23041187D

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, *59*, 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168

Earnshaw, R., Liggett, S., Excell, P., & Thalmann, D. (Eds.). (2020). *Technology, Design and the Arts - Opportunities and Challenges*. Springer International Publishing. https://doi.org/10.1007/978-3-030-42097-0

Elmas, T., Stephane, S., & Houssiaux, C. (2023). Measuring and Detecting Virality on Social Media: The Case of Twitter's Viral Tweets Topic. *Companion Proceedings of the ACM Web Conference 2023*, 314–317. https://doi.org/10.1145/3543873.3587373

Fan, X. (2023). Social Media Marketing Strategies. *Advances in Economics, Management and Political Sciences*, *23*(1), 59–64. https://doi.org/10.54254/2754-1169/23/20230353

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Gao, L., Liu, Y., Zhuang, H., Wang, H., Zhou, B., & Li, A. (2021). Public Opinion Early Warning Agent Model: A Deep Learning Cascade Virality Prediction Model Based on Multi-Feature Fusion. *Frontiers in Neurorobotics*, *15*. https://doi.org/10.3389/fnbot.2021.674322

Garcia-Salirrosas, E. E. (2020). Conceptual framework of viral marketing: a review of the literature of the last years. *Global Business Administration Journal*, *4*(1), 18–27. https://doi.org/10.31381/gbaj.v4i2.2890

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow*. O'Reilly.

Gold Nmesoma Okorie, Chioma Ann Udeh, Ejuma Martha Adaga, Obinna Donald DaraOjimba, & Osato Itohan Oriekhoe. (2024). ETHICAL CONSIDERATIONS IN DATA COLLECTION AND ANALYSIS: A REVIEW: INVESTIGATING ETHICAL PRACTICES AND CHALLENGES IN MODERN DATA COLLECTION AND ANALYSIS. *International Journal of Applied Research in Social Sciences*, *6*(1), 1–22. https://doi.org/10.51594/ijarss.v6i1.688

Gupta, Dr. R., & Kumar, V. (2023). Role of Artificial Intelligence in Different Sectors of Society. *JOURNAL GLOBAL VALUES*, *XIV*(S.Issue), 34–40. https://doi.org/10.31995/jgv.2023.v14iS3.005

Han, Y., Lappas, T., & Sabnis, G. (2020). The Importance of Interactions Between Content Characteristics and Creator Characteristics for Studying Virality in Social Media. *Information Systems Research*, *31*(2), 576–588. https://doi.org/10.1287/isre.2019.0903

Havrlant, L., & Kreinovich, V. (2017). A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, *46*(1), 27–36. https://doi.org/10.1080/03081079.2017.1291635

Huang, C., & Petukhina, A. (2022). *Applied Time Series Analysis and Forecasting with Python*.

Jamil, K., Dunnan, L., Gul, R. F., Shehzad, M. U., Gillani, S. H. M., & Awan, F. H. (2022). Role of Social Media Marketing Activities in Influencing Customer Intentions: A Perspective of a New Emerging Era. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.808525

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joseph, N., Kar, A. K., & Ilavarasan, P. V. (2021). How do network attributes impact information virality in social networks? *Information Discovery and Delivery*, *49*(2), 162–173. https://doi.org/10.1108/IDD-08-2020-0094

L. Lopez, Y., Grimaldi, D., Garcia, S., Ordoez, J., Carrasco-Farre, C., & A. Aristizabal, A. (2022). Artificial Intelligence Model to Predict the Virality of Press Articles. *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, 221–228. https://doi.org/10.1145/3529836.3529953

Le Meunier-FitzHugh, K. (2021). The nature of marketing. In *Marketing: A Very Short Introduction* (pp. 1-C1.P28). Oxford University PressOxford. https://doi.org/10.1093/actrade/9780198827337.003.0001

Madakam, S., Uchiya, T., Mark, S., & Lurie, Y. (2022). Artificial Intelligence, Machine Learning and Deep Learning (Literature: Review and Metrics). *Asia-Pacific Journal of Management Research and Innovation*, *18*(1–2), 7–23. https://doi.org/10.1177/2319510X221136682

McCrea, R., King, R., Graham, L., & Börger, L. (2023). Realising the promise of large data and complex models. *Methods in Ecology and Evolution*, *14*(1), 4–11. https://doi.org/10.1111/2041-210X.14050

Mikko Dufva. (2020). *MEGATREND 4: Technology is becoming embedded in everything*. https://www.sitra.fi/en/articles/megatrend-4-technology-is-becoming-embedded-in-everything/

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine Learning*. CRC Press. https://doi.org/10.1201/9781315371658

Muller, E., & Peres, R. (2019). The effect of social networks structure on innovation performance: A review and directions for research. *International Journal of Research in Marketing*, *36*(1), 3–19. https://doi.org/10.1016/j.ijresmar.2018.05.003

Ng, K. W., Mubang, F., Hall, L. O., Skvoretz, J., & Iamnitchi, A. (2023). Experimental evaluation of baselines for forecasting social media timeseries. *EPJ Data Science*, *12*(1), 8. https://doi.org/10.1140/epjds/s13688-023-00383-9

Parthasarathy, S., Panigrahi, P. K., & Subramanian, G. H. (2024). A framework for managing ethics in data science projects. *Engineering Reports*, *6*(3). https://doi.org/10.1002/eng2.12722

Preibisch, S. (2018). API Authentication and Authorization. In *API Development* (pp. 61–105). Apress. https://doi.org/10.1007/978-1-4842-4140-0_5

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, *7*(2), 205630512110190. https://doi.org/10.1177/20563051211019004

Rajaraman, A., & Ullman, J. D. (2012). Mining of Massive Datasets. *Cambridge University Press.*

Reichstein, T., & Brusch, I. (2019). The decision-making process in viral marketing—A review and suggestions for further research. *Psychology & Marketing*, *36*(11), 1062–1081. https://doi.org/10.1002/mar.21256

Rosário, A. T., & Dias, J. C. (2023). Marketing Strategies on Social Media Platforms. *International Journal of E-Business Research*, *19*(1), 1–25. https://doi.org/10.4018/IJEBR.316969

Sadrnia, L. (2023). The Future of Marketing: How Predictive Modeling Optimizes Campaign Strategies. *IBusiness*, *15*(04), 249–262. https://doi.org/10.4236/ib.2023.154018

Sanders, A. M. (2020). UsTube - A Grounded Theory Analysis of the Relationship Between YouTube and Influencers. *Brigham Young University*.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shahbaznezhad, H., Dolan, R., & Rashidirad, M. (2021). The Role of Social Media Content Format and Platform in Users' Engagement Behavior. *Journal of Interactive Marketing*, *53*, 47–65. https://doi.org/10.1016/j.intmar.2020.05.001

Shang, Y., Zhou, B., Zeng, X., Wang, Y., Yu, H., & Zhang, Z. (2022). Predicting the Popularity of Online Content by Modeling the Social Influence and Homophily Features. *Frontiers in Physics*, *10*. https://doi.org/10.3389/fphy.2022.915756

Solomon, M. R., & Tuten, T. L. (2016). *Social Media Marketing*. Αθήνα: Εκδόσεις Δίαυλος.
Taherdoost, H. (2023). Enhancing Social Media Platforms with Machine Learning Algorithms and Neural Networks. *Algorithms*, *16*(6), 271. https://doi.org/10.3390/a16060271
Teahan, W. (2018). A Compression-Based Toolkit for Modelling and Processing Natural Language Text. *Information*, *9*(12), 294. https://doi.org/10.3390/info9120294
Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2019). What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence. *Journal of Marketing*, *83*(4), 1–20. https://doi.org/10.1177/0022242919841034
Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Lee, Y., & Chen, R. C. (2023). Employing long short-term memory and Facebook prophet model in air temperature forecasting. *Communications in Statistics - Simulation and Computation*, *52*(2), 279–290. https://doi.org/10.1080/03610918.2020.1854302
Xu, Z., & Qian, M. (2023). Predicting Popularity of Viral Content in Social Media through a Temporal-Spatial Cascade Convolutional Learning Framework. *Mathematics*, *11*(14), 3059. https://doi.org/10.3390/math11143059
Yang, B., Zhang, R., Cheng, X., & Zhao, C. (2023). Exploring information dissemination effect on social media: an empirical investigation. *Personal and Ubiquitous Computing*, *27*(4), 1469–1482. https://doi.org/10.1007/s00779-023-01710-7
Yeturu, K. (2020). *Machine learning algorithms, applications, and practices in data science* (pp. 81–206). https://doi.org/10.1016/bs.host.2020.01.002
Zhang, J., Mucs, D., Norinder, U., & Svensson, F. (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets. *Journal of Chemical Information and Modeling*, *59*(10), 4150–4158. https://doi.org/10.1021/acs.jcim.9b00633
Zhou, F., Xu, X., Trajcevski, G., & Zhang, K. (2022). A Survey of Information Cascade Analysis. *ACM Computing Surveys*, *54*(2), 1–36. https://doi.org/10.1145/3433000