A Cross Platform Framework for Social Media Information Diffusion Analysis

H.M.M.Caldera

198132D

Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

November 2023

A Cross Platform Framework for Social Media Information Diffusion Analysis

H.M.M.CALDERA

198132D

Thesis submitted in partial fulfillment of the requirements for the degree
Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

November 2023

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text. Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature of the candidate: Date: 21.11.2023

Mr.H. M. M. Caldera

The above candidate has carried out research for the PhD thesis under my supervision.

Name of the supervisor: Prof. G. I. U. S. Perera

Signature of the supervisor: .... Date: 21.11.2023

Acknowledgements

First and foremost, I want to thank my supervisor, Prof. Indika Perera, for his unbeatable support throughout the study. He has taught me, both consciously and unconsciously, how well experimental research is done. I appreciate all his time, ideas, and recommendation for the AHEAD grant to help fund my Ph.D. study.

Next, I would like to thank AHEAD Grants for financial support through research and give special thanks to Dr. Lochandaka Ranathunga as research coordinator for all financial approvals. Dr. Shalinda Adhikari for winning the research grant. I won't complete this degree if you have not won the grant.

As a fellow researcher, Mrs. Nadeera Meedin provided unbelievable support for success during the last few years. She has fully supported my success in many aspects. She shares personal research experiences and motivates me a lot for success.

Further, I thank all the research fellows who worked in the "social media analytics research group" for helping me. Thanks to reviews (including internal ones) for providing valuable feedback.

All authors develop valuable content and publish it on the web. I have gained much knowledge from your videos, documents, research articles, etc. Thanks to all the online content authors. Thanks to Alexander T., Research gate, and other online portals for making many scientific articles publicly available. Further, all scientists who kept their research data as a public asset and let other fellow researchers explore their research.

Finally, I would like to thank my father, mother, wife, son, and all other family members for motivating me and staying with me during this challenging period.

Abstract

In the current digital era, social media platforms have emerged as one of the most effective channels for the diffusion of information. People may readily access and exchange information, news, and opinions from anywhere worldwide because of increasing social media usage.

Information diffusion across multiplex social media platforms is one of the most prominent research problems ever. Social media content generators diffuse information on multiplex social media platforms by targeting many objectives such as popularity, online presence, hate targets, and customer engagement. Regardless of the ”content” posted on social media platforms, evaluating the dissemination velocity of each piece of content published on those platforms is essential. It will help to get an overall picture of ”how it flows” throughout the social media platforms. Most social media platforms have a platform-specific algorithm for calculating the degree of information diffusion on those platforms. The main objective of this research was to develop a method to calculate the velocity of information diffusion across multiplex social media platforms.

Existing literature on information diffusion strategies, effects, and measurements was used to develop the proposed algorithm. The information diffusion velocity of social media influencers varies according to the content. The platform-specific algorithms for diffusion strength detection vary based on the platform. Somehow, these platform-specific algorithms influence the community to engage with the trending content. i.e., platforms support increasing the strength of information diffusion.

Conventional information diffusion algorithms were designed to measure content diffusion speed on a simplex social media platform, which might be content-specific. The missing dimension is ubiquitous nature. Hence, regardless of the platform, it is mandatory to calculate a ubiquitous information diffusion velocity over multiplex social media platforms.

Both structured information diffusion in a graph for diffusion in a closed network and unstructured patterns in an open-ended coarse-grained information diffusion model check the importance of information diffusion on multiplex social media platforms. Time is another critical factor in defining velocity. i.e., a time series of information diffusion provides a rich picture of information diffusion.

Event-driven architecture is a well-known software architectural approach that facilitates the implementation of microservice-based solutions. The suggested algorithm utilizes an event-driven architecture to manage the information flow by processing social media events. Eventually, this research uses the event-triggering process to understand how information is propagated through an event-driven microservice architecture.

Data science and artificial intelligence are being employed in information diffusion

studies. Understanding how information spreads and the variables and features that influence it is another crucial study area of this research. There are several techniques for studying information dissemination using artificial intelligence. Applying artificial intelligence to information diffusion studies might improve our knowledge of ”How information travels” and ”how to disseminate information” in various circumstances efficiently. The research used natural language processing to evaluate the textual content of the social media post. That is to find a general textual meaning given by the end-user reactions.

Event-driven architecture is one of the best possible for information diffusion analytics. Using event-driven architecture, data may be delivered in real-time to various analytics services, allowing for the speedy and effective processing of enormous amounts of data. This is especially true in today's data-driven world, when businesses and organizations must make quick, well-informed decisions based on real-time data. Because of its event-driven nature, it is also simple to interface with other systems and services, making it a highly adaptable and versatile option for information distribution analytics. Since the diffusion of information starts with an event's occurrence, it follows numerous steps to flow among the community. An event-driven micro-services architecture that uses artificial intelligence methods (like natural language processing to evaluate textual information) has been experimented with to propose a simple solution for this complex problem.

As per the research work, I can summarize the key findings. I have proposed a tree-structured diffusion tree that can explain how information flows through multiplex social networks. Under this multiplex context, I have experimented with multiple trees and a more robust graph that focused on the diffusion of information. The diffusion strength was based on the SIR model, and the time series analysis focused on how quickly information spread throughout the network. The proposed solution was tested in several real-world cases. Technique-specific tests like seasonality and autocorrelation were conducted to evaluate how the time-series model works in a graph context. Further tests like cohesiveness and robustness were tested, and the proposed algorithm achieved good robustness (an average of 75%) and cohesiveness (an average of 70%) in each case. The best experimental results show an average of more than 80% accuracy in any given instance, and it constructs the tree in less than a second. Most of the predicted values generated an average accuracy of around 70%.

In summary, social media platforms have emerged as prominent channels for information propagation within the contemporary digital landscape. Quantifying the velocity at which information propagates across diverse social networks presents a notable challenge in research. While algorithms tailored to specific platforms influence community engagement, a ”universal metric for information dissemination strength” is necessary across multiple social media platforms. The envisioned algorithm considers time series data, integrating structured and unstructured patterns during construction.

Keywords: Information diffusion analysis, Social Media Data Analytics, Graph Learning, Time series analysis, Event-driven micro-services, Artificial Intelligence, Natural Language Processing.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial intelligence

API Application Programming Interfaces

ARIMA Autoregressive Integrated Moving Average

BC Betweenness centrality

DLR Dynamic Linear Regression

EC Eigenvector centrality

ETS) Exponential Smoothing

FMTS Fixed model time series

LDA Latent Dirichlet Allocation

OMTS Open model time series

SNA Social network analysis

SNS Social Network Sites

STL Seasonal Decomposition of Time Series

SVM Support vector machine

TDC Total degree centrality

TFP Total-From-Partial

Chapter 1
INTRODUCTION

1.1 Overview

De facto, social media platforms are the most prominent online platform for user engagement and information diffusion. Total global social media usage in January 2023 was 4.76 billion[1]. Total social media users in Sri Lanka in January 2023 were 7.20 million, i.e., 32.9% of the total Sri Lankan population[2]. Furthermore, Facebook has 6.55 million users, YouTube has 7.04 million users, and there are 37.31 million users on Twitter in Sri Lanka[2].

Social networks can proliferate with new social nodes and interactions among nodes. Moreover, those nodes create new social groups. Rapidly growing social media content contains various information, including social issues, discussions related to politics, comedies, teledramas, stories, and hate speech. Further, social media platforms are comfortable and one of the simplest methods to connect without common barriers in a physical environment, such as location and time. Information diffusion is the technique that describes" how information flows from one end to another." Under the context of social media, information diffusion has several facades, such as influencing diffusion, herd behavior, and information cascades. Users can access various channels through multiplex social networking sites like Facebook, Twitter, Instagram, and LinkedIn to interact and share their relationships. Identifying significant users, communities, and trends can be aided by analyzing the dissemination of information on such platforms.

The "Information Diffusion Analytic Framework" proposed here, is an invaluable resource for researching the dissemination of information across multiplex social media platforms. The three steps of this approach are data gathering, network building, and diffusion analysis.

Under the data gathering section, information about user interactions, such as

likes, comments, shares, and mentions, is collected. A social network that depicts the connections between people is built using this data. It is possible to visualize this network to find prominent users and communities. Graph theory is used in network construction to depict the social network, where nodes are users and edges are their relationships.

Diffusion analysis examines how information spreads over a network. The pace and direction of information flow may be revealed through this study, along with bottlenecks and key users who can influence how quickly or slowly information spreads. The research may also be utilized to pinpoint user-favorite patterns and subjects.

Several social media sites may be used with the ”Information Diffusion Analysis Framework” offers valuable insights into user behavior and preferences. For instance, it may assist companies in identifying prominent people who can promote their goods or services, and it can assist marketers in comprehending the types of content that work best for attracting their target market.

Finally, I would like to introduce the ”Information Diffusion Analysis Framework” as a powerful tool for studying information diffusion on multiplex social media platforms. By collecting data, constructing a social network, and analyzing information diffusion, this framework can help identify influential users, communities, and trends, providing valuable insights for businesses and marketers.

1.2 Problem Statement

The problem statement of this research can be stated as follows - The existing literature lacks a comprehensive algorithm to evaluate the strength of information diffusion on a multiplex social media platform.

1.3 Motivation

Information diffusion among social media platforms can be involved in multiple aspects. The nature of information diffusion can impact various outcomes, such as social interactions. Unique business algorithms evaluate each piece of content published on social media platforms. The foremost hindrance behind these algorithms is that they are specified only for a given business platform. In other words, platform-specific information diffusion analysis algorithms are used by each platform to identify the influencing/trending content on that given social media platform. Hence, a proper algorithm that can evaluate the diffusion of information across multiple social media

platforms is required.

1.4 Research Questions

1. How to investigate a mechanism to identify information diffusion in multiplex social media platforms?
2. What are the methods to implement a proper mechanism for information diffusion in multiplex social media platforms?
3. What are the mechanisms for identifying the evolution of information diffusion in social media platforms?
4. How to implement a mechanism to identify the distribution of identifying message contents across multiplex social media networks.

1.5 Objectives

- To develop an algorithm to measure platform-independent information diffusion speed/information diffusivity. i.e., Investigate and implement a mechanism to analysis information diffusion in multiplex social media platforms regardless of specific content
- To design an information diffusion framework that calculates streaming information diffusivity. i.e., derive a mechanism to identify the information diffusion and its nature on multiplex social media platforms. The completed mechanism identified the trend regardless of the content and its nature.
- To implement a mechanism for identifying the distribution of information diffusion across multiplex social media platforms.
- To experiment and evaluate the analytical capabilities of the proposed solution.

1.5.1 Usefulness of the research

Proposing a method for information diffusion analysis has significant technological significance. Integrating dynamic network architecture, representing human behaviors, and responding to changing network circumstances in real-time is critical. Combining numerous data sources, allowing parameter adjustment and sensitivity analysis, and

combining community identification and impact analysis all improve the algorithm's capabilities. These technological advancements enable the exact analysis of information diffusion, facilitating informed decision-making in various fields. The foremost usefulness of the research is to narrow it down to "Understanding the Flow of Information in Social Networks.". Information diffusion analysis methods may aid in understanding how information flows across social networks. This is useful for marketers, academics, and policymakers in developing successful methods for reaching out to target populations and conveying critical information. Overall, offering an information diffusion analysis technique helps improve knowledge of how information travels, allowing for more informed decision-making and targeted interventions in various domains and applications.

Real world examples

- Early detection of negatively impacting information diffusion. e.g:-Fake news detection[3].
- Information diffusion analysis when an emergency situation. e.g:- Natural disaster [4],[5].

1.6 Digital social network

A digital social network [6] is a digital platform or website that enables people to communicate with one another, exchange knowledge, etc. Generally, a digital medium is used for interconnecting with each other. Social networks make it easier for people to communicate socially and exchange information[7]. Examples include social networking sites like Facebook, Twitter, Instagram, LinkedIn, and Snapchat. Users may often create profiles on social networks, interact with other users, share posts, images, and videos, and join communities or groups [7, 8]. These networks have grown in popularity and changed how individuals engage online and offline. This research's base is identifying data distributions among social media networks.

"Digital society analysis" means identifying, tracking, analyzing, and defining proper methods in all digital encampments of humans engaging in digitalized social activities based on various technological implementations. It has many aspects of social engagement, like social groups based on different interests and personal updates—digitalized social engagements based on social psychology. Hate speech is an egregious sociological problem in the community. Besides physical society, hate

speech is inculcated in digital society via social media platforms, and hate speech content is distributed among different digital sociological clusters, flourishing for distinct reasons.

Social innovations are embedded into social media platforms[9]. Online social media platforms are transforming sociological interactions and interpersonal communication. The multiplex social echo system drastically changes human life, specifically how humans engage in social networks. They are expanding quickly for many reasons, like their collaboration's solid, constructive nature, and it is accessible.[10].

Ties connect network nodes; the ties have relationships with one or many related connections, depending on the relation type[11]. Ties connect network nodes, and the links have a relationship with one or many related connections based on the types of ties. Information is disseminated globally via different networks. Analysis of how it disseminates is a massive research question. Hence, connectivity among the networks is a well-known catalyst to confirm the information will diffuse faster.

Graph theory is a popular method for network analysis [12]. Social engagements and content diffusion are visible on a time series-based social network graph [13]. A single social media platform is known as a "simplex social network."The term "simple social network" was coined in the middle of the 1950s by social psychologist Robert Freed Bales[14]. Bales developed the concept of a simplex social network to understand the communication patterns and group dynamics that emerged during small-group interactions [15]. A multiplex social media network will develop when multiple communication platforms are connected[16]. In most cases, homophily [17] and cascading[18] are general user engagements in multiplex social media networks.

A critical aspect of this study is "identifying the effect of cross-posting in a multiplex social media environment on trending social media content[19]." Under the context of social media, trending is based on the number of user engagements for a post in each time window.

1.7 Social media networks

Social media has transformed the way people interact with one another and how they consume and share information. While social media has its benefits, such as connecting people worldwide and promoting social causes, it also has its drawbacks, such as cyberbullying and spreading false information.

Social media has revolutionized the way people communicate with each other.

Platforms such as Facebook[1], Instagram[2], and Twitter[3] allow individuals to connect with others regardless of location or time zone. This has been particularly beneficial for long-distance/virtual relationships. Further those relationships allowed individuals to stay in touch with loved ones even when they are far apart. Additionally, social media has provided a platform for social activism[20], allowing individuals to share information about essential causes and organize protests and demonstrations. For example, the ”Black Lives Matter” incident [21]gained significant traction on social media, with users sharing information and resources to support the cause.

However, social media also has its drawbacks. One of the significant issues with social media is ”cyberbullying”. Cyberbullying is a form of online harassment where individuals use social media platforms to bully, harass, or intimidate others [22]. The anonymity provided by social media platforms can embolden individuals to say things they would not say in person. Cyberbullying can lead to severe emotional and psychological harm, particularly for young people, who are more likely to be targeted. Social media platforms have been used to spread false information and propaganda, which can have serious consequences[23]. For example, during the 2020 US presidential election, social media platforms were used to spread false information about the candidates, which may have influenced the election outcome[24].

Social media is one of the prominent domains in information diffusion. Especially one that can recognize people as they interact regularly. Due to the openness of social media, there are many ways to interact with several types of information. Society can interact with all social media content that is publicly available without many constraints like cost and time. It confirms that social media platforms play a crucial role in information diffusion by removing many traditional boundaries[25]. Hence, social media is a demanding platform in different contexts, such as information propagation.

Social media platforms have some standard features. These standard features include sharing user-generated content and user engagements, including reactions. Social media posts are the information chunks uploaded to any social media platform. This could be either a direct or indirect upload. Direct upload means ”the user directly uploads information to any social media platform.” Indirect upload means ”the original information is uploaded not to a social media platform but to a business website, personal website, or other non-social media platform.” After uploading the information to a non-social media platform, the user spreads it to social media platforms by sharing, uploading a link, or using any other method to propagate the information on

1https://www.facebook.com/
2https://www.instagram.com/
3https://twitter.com/

social media platforms. Hence, each piece of social media content has its own pace of information propagation. High-demand content receives high user engagement and spreads virally on social media platforms.

Social media content is trending upward or downward based on user engagement for assorted reasons, such as societal impact, personal importance, etc. In this research, ”trend” means the user engagements (i.e., views, comments, or any other means of digital engagement for the social media post) against time. Upon high user engagement, content has an upward trend; when users do not interact with the content, it has a downward trend. Information diffusion on social media platforms creates a trend over time due to user engagement. More user engagements pictorial the upward trend. Demand drives the trend. Once demand is lower, the trend line will move downward. As a result, it is critical to identify key features that drive demand for social media content.

Web browser cookies store information about the content that a user has visited or revisited. Hence, cookies are one of the key features available for identifying trends. Nevertheless, in this research, we consider a user as a node. Further, we focus on content instead of the nodes in the network. Additionally, trends depend on the location; most specifically, country- or region-wise trends are available. A life cycle exists for each content hosted on a social media platform. i.e., each social media post has some trend line after the initial post. Rather than engaging with simplex social media networks, multiplex social media network engagements are common in many user profiles. Information diffusion across multiple social media platforms is considered when focused on various platforms.

”Trending social media content” is one of the topics the researchers. The research community typically investigates in a simplex network or a more widely applicable multiplex context. Trend analysis is a fundamental research focus on multiplex social media networks. Short-term trends can be detected when a specific content trend over a short period, usually ending within a few days/weeks, or months. Moreover, each social media platform has a specific nature of information diffusion. For example, Facebook, YouTube, and Twitter are popular social media platforms that engage many users. Facebook has a common strategy of sharing information based on close relationships known as ”friends” or open groups to have members discuss topics of particular interest. Twitter is the largest microblogging platform, providing access to the entire community via hashtags. Twitter has followers and keeps a consistent feed on relevant topics. YouTube is an open community video-sharing platform that provides open access to information retrieval based on its search and home page utilities. During the past decade, researchers have focused on various aspects of trend detection and prediction,

including the scope of user engagements, topic mining, modeling, etc. Social media platforms are involved in identifying various social trends to identify user similarity, content similarity, and other socio-technical aspects.

YouTube keeps a record of live trending video content [4], and Twitter uses another set of features to recognize the trending hashtag [5]. Besides YouTube and Twitter, trending content identification is common across almost all social media platforms. Further, different trends are visible according to the platform, mainly focusing on business trends for advertising purposes.

Social media platforms play a key role in information dissemination by removing many traditional boundaries. Hence, social media content has its own pace of information propagation. More demand increases user engagement, spreading social media content at a different strength. Once the information spreads on social media platforms, depending on the user's interest in that information, users initiate engagements. If more engagements occur in less time, it depicts the viral nature of spreading information.

Moreover, each social media platform has its nature of information dissemination. Facebook, YouTube, and Twitter are popular social media platforms that engage in Sri Lanka. Facebook has a common strategy of sharing information based on close relationships known as "friends" or open groups for topics of particular interest. Twitter is the largest microblogging site and provides open access to the entire community via hashtags. Twitter has followers and posts consistent updates on relevant topics. YouTube is an open community video-sharing platform that provides access to information retrieval based on the search and home page utilities. Information diffusion on social media platforms creates an engagement trend based on user engagement over time.

1.7.1 Social media networks

There are a bunch of social media platforms available. Some of the most popular platforms are

- Facebook - A social networking website that enables users to connect with friends, relatives, and acquaintances, exchange information (such as images and videos), and become members of groups with like-minded people.

- Instagram - A social networking website where users may upload material, follow other users, and interact with postings through likes, comments, and direct

[4] Access to trending videos in a given country.https://www.youtube.com/feed/trending
[5] Access to trending hashtags in a given country.https://twitter.com/i/trends

messaging.

- Twitter - A social networking platform that lets users communicate with one another's tweets by liking, retweeting, and replying to them. These communications are known as "tweets."
- YouTube - A public Video sharing service/platform
- WhatsApp - Mobile-based social networking application

Facebook

Facebook was established in 2004 and has become one of the world's most popular and significant websites. Facebook has transformed how people interact, communicate, and share information online, with over 2.8 billion monthly active users.

Users of the site may create personal profiles, add friends, and exchange links, updates, images, and videos. Facebook also provides tools like groups, pages, and events that enable users to connect with others who share their causes, organizations, or hobbies.

One of its main advantages is that Facebook can link individuals worldwide and bring them together for a common direction for information diffusion. This has resulted in the development of several communities on the platform. Facebook groups, for instance, have been crucial in coordinating political action, bringing attention to social concerns, and uniting individuals with like-minded interests[26] .

Facebook has still come under fire for how it manages user privacy and data. The corporation has been under more attention recently for its role in disseminating fake information and propaganda and for how it has been applied to sway elections and thwart democratic processes[27].

Despite these difficulties, Facebook continues to be among the most effective and popular platforms in the world. Facebook will play an increasingly significant role in determining how people connect and engage with one another as people spend more and more time online.

Facebook has significantly changed how people interact and communicate online. Although it has encountered difficulties and controversy, its impact on the internet and society cannot be denied. Whatever your opinion of Facebook, it is evident that it will continue to play a significant role in our lives for a very long time.

YouTube

The YouTube platform has transformed the way we watch and share videos. From its modest beginnings in 2005, it has come a long way to become one of the biggest video-sharing sites in the world. People from many walks of life, including entrepreneurs and amateurs, utilize YouTube, which has completely changed how we watch and distribute videos.

YouTube has grown in importance as a tool for businesses, and many firms now use it for marketing[28]. Companies are embracing the video format to reach their target audience in an exciting and participatory way as a critical marketing mix component. Many businesses are using the platform's user involvement and reach to market their goods and services[29] and increase brand recognition[30, 31]. A vital tool for education[32] and generally, YouTube has also developed into a source of news and entertainment.

As a result of the expansion of YouTube, video material is now a lot more readily available and practical to distribute. The platform allows people to publish, share, and watch videos anytime and from any location. It has also created new opportunities for creativity and self-expression by enabling people to communicate their ideas, views, and experiences with a larger audience. Further, people may now connect with others who have similar interests and create communities around them.

Despite its many benefits, YouTube also has its challenges. One of the biggest challenges is ensuring that the content on the platform is appropriate and does not infringe on other people's rights. YouTube has implemented several policies and procedures to ensure that the platform remains safe and secure for all users and does not become a haven for hate speech, fake news, and other harmful content. Handling the enormous volumes of data created by YouTube's users presents another difficulty. This has grown in importance as the platform expands and draws more users. To manage this data and guarantee that the platform stays dependable and stable, YouTube has had to invest significantly in infrastructure and technology.

In conclusion, YouTube has altered how we consume and exchange information by becoming a crucial medium for viewing and disseminating video content. The platform's future is promising since it can expand and keep altering how we consume and exchange information. Despite its difficulties, YouTube continues to be a helpful tool for organizations, teachers, and people, and it will play a significant role in the digital world for years to come.

Twitter

Since its introduction in 2006, Twitter has grown to be one of the most well-known microblogging platforms in the world. Twitter, which has over 330 million active users monthly, enables users to communicate in 140-character ”tweets” about their thoughts, ideas, and experiences. The platform has developed into a potent tool for governments, businesses, and individuals to interact with and communicate with a worldwide audience.

Journalists, activists, and politicians have turned to Twitter as a powerful tool for reaching out to their followers, sharing their ideas, and participating in global dialogues. Twitter is sometimes called the ”pulse of the planet” and has established itself as a valuable platform for communicating information, ideas, and events as they develop in real-time. Natural catastrophes, political upheavals, and other significant events have all been covered by it, and it has played a crucial role in influencing public opinion and fostering societal change[33].

Despite these difficulties, Twitter continues to be among the world’s most effective and popular social media platforms. It plays a crucial role in influencing public debate and igniting social change. It will continue to be a prominent player in the social media landscape for many years.

Twitter is a powerful tool that has significantly changed how people interact and communicate. Twitter can connect people and inspire change in a manner that few other technologies can equal, whether you use it for personal or professional reasons. Whether you view Twitter as a force for good or evil, there is no denying that it has significantly integrated into our lives and will continue to do so for a long time.

Different social media platforms provide several content-sharing methods.

The table 1 describes the high-level features of the selected platforms.

As represented in table 1, the platforms have considerable common features.

social media platforms are the most extensive digital social engagement mechanism in the current era. These platforms enable social engagements to take place in a digital environment. Further, this research was based on social media data and the behavior of the digital society.

Furthermore, hate speech is detrimental to social growth and development. Hate speech initially arises due to multiple social diversities like politics, religion, and race. Language in social networks, on the other hand, considers various linguistic

Content sharing method	Facebook	Twitter	YouTube
Social networking	Yes	Yes	Yes
Discussion forums	Yes	Yes	Yes
Bookmarking	Yes	Yes	Yes
Image upload	Yes	Yes	Yes
Text writing	Yes	Yes	Yes
Live stream	Yes	Yes	Yes
Hashtag	Yes	Yes	Yes
Following	Yes	Yes	Yes

Table 1: Common features of social media platforms

dimensions. The language used in social networks frequently differs from traditional modes of communication, such as verbal or textual. Social media discussions use composite images and textual vocabulary in multiple languages simultaneously with minimal grammar consideration. Most of the sentences in the discussions could be informal and less informative. In the context of social media in Sri Lanka, it can recognize the following languages: Sinhala, Sinhala language words written in English text (known as ”Singlish”), Tamil, and English. The simplest example is the discussion of emojis, which express human emotions. They are acknowledged to use a simple form, ”hard to express,” using standard vocabulary.

On the other hand, social media platforms have minimal language restrictions. All major social media platforms support unicode characters for Sinhala, Tamil, and English. Hence, based on the end user’s comfort, any language, image, emoji, or other mechanisms can help interact with the discussion. The same information can spread in different contexts within social media platforms. Specifically, a model of communication by language has a wide variety, and it is coherent when it comes to hate speech-related discussions. The contexts of digitalized sociological hate speech range from minor social issues to major, complex social pathological debates, open-ended discussions, and the high strength of brainstorming to which we are accustomed. Contrary to direct hate speech, indirect hate speech in social media, such as offensive hate speech, is common.

Direct hate speech can focus on a single cluster of people or multiple clusters in different social regions. Any single human can interact with multiple social media platforms and engage with different social layers based on the intention of the information shared on social media platforms. A user can have multiple social media accounts on different social media platforms, and there is a possibility of sharing the same information chunk on different social media platforms by the content author or

discussion partner. Such information spreads across various social media platforms simultaneously, enabling user engagement on different platforms, and the number of engagements is changing over time. In other words, based on user behavior, a trend can be defined to identify the hate speech cross-posted by users in a single post. Trends on a single social media platform are based on user engagements on the given social media platform. The discussions are based on one cluster of people.

Contrary to trendy belief, cross-posted hate speech can have an impact on multiple social groups at the same time, which is evident in massively digitalized ”social unrest.” Hence, the most critical factor in the hate speech trend analysis is identifying trends in cross-posted hate speech. Identifying trends in social media platforms is essential, like blocking unwanted social interactions and digital sociological wars[34]. Hence, defining a proper cross-posted trend analysis is a critical factor, and it can help in numerous ways in critical digital social situations.

When social issues occur, the importance of identifying trends in cross-posts related to hate speech becomes viral; identifying the trends makes it easy to control or block whenever hate speech occurs. This research discusses the identification process, designing the appropriate algorithm for cross-posted hate speech trends, and analyzing and evaluating the algorithm for cross-posted hate speech in the context of social media platforms in Sri Lanka and involved in hate speech in Sinhala or Singlish.

1.7.2 Structure of a Social media network

The network’s behavior and the patterns emerging from the network create the network’s structure[35]. Mainly, the network structure is dynamic based on the interactions of the nodes and the content sharing in the network. An actor’s node in the network provides a view of the network’s influence nature and other factors affecting the node. Relationships are viewed as networks. Ties represent network relationships. The network’s behavior and the patterns emerging from the network create the structure of the network. Mainly, the network structure is dynamic based on the interactions of the nodes and the content sharing in the network.

As per the figure 1.2, A digital social network is a collection of nodes and edges connected to form a cluster.

1.7.3 Types of social media data

- Textual : Any text-based information produced on social networking sites, such as posts, comments, messages, and reviews, is included in this.

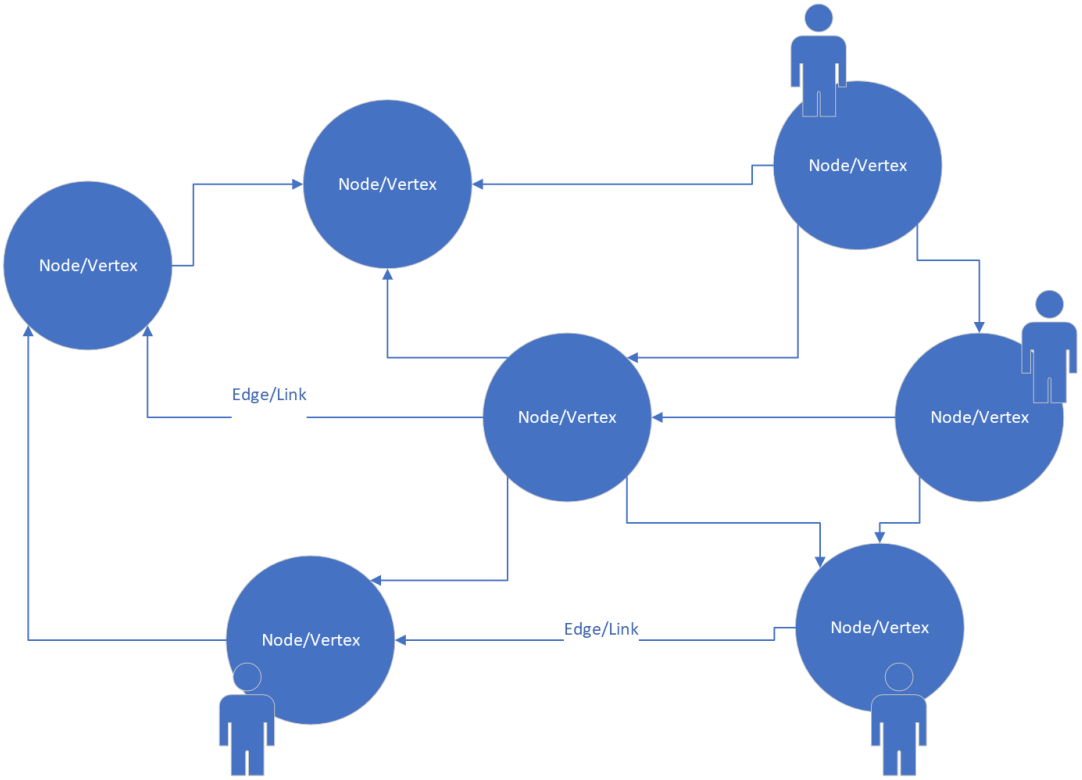


Figure 1.1: Social media network

- Graphical (Image/Video/Audio) : Images and videos shared on social media sites like Instagram and YouTube are frequently utilized to study visual trends and user behavior.
- Reactions: This covers all users' online social exchanges, such as likes, comments, shares, and clicks.

Commonly the data shared on social media can be viewed as mixed content. I.e., a mix of text and some graphics.

1.8 Social media analytics (Data science in social media context)

1.8.1 Social trends

Humans evolve with various aspects of change. Technological, environmental, religious, and other social changes significantly impact human evolution. Social media

platforms are the underlayer for the interaction of virtual social engagements.

Social trends are highly coupled with the environment. Due to different features, humans adapt to a new environment and eject some of the unnecessary components of their life. Humans evolving with different aspects of changes such as technological changes, environmental changes, religion, and other social-related changes favorably impact human evolvement. Social media platforms underlying the interaction of virtual social engagements.

1.8.2 Data Science

Data science is the science of processing a collection of raw data into meaningful insights, patterns, and design structures by using relevant algorithms and tools. Data science can involve many subject areas, such as statistics, computer science, mathematics, data cleaning and formatting, and techniques like data visualization. One of the reasons for the rise of data science in recent years is the vast amount of data currently available and being generated. This has created the perfect storm in which we have rich data and the tools to analyze it: rising computer memory capabilities, better processors, more software, and now, more data scientists with the skills to put this to use and answer questions using this data. In our context, any data shared among a social media network can define as ”social media content.”

1.8.3 Big data

There are a few qualities that characterize big data. The first is volume. As the name implies, big data involves large data sets, which are becoming increasingly routine. For example, YouTube has approximately three hundred hours of video uploaded every minute [36].You would have a lot of data available to analyze, but you can see how complex it is to wrangle all that data.

This brings us to the second quality of big data: velocity. Data is being generated and collected faster than ever before. In our YouTube example, new data is coming at every minute! In a completely different

The third quality of big data is variety. Different data types are in the examples I have mentioned. In the YouTube example, video or audio is a very unstructured data set. A database of video lengths, views, or comments would be a much more structured data set to analyze.

The context can evolves more with respect to the figure 1.2, the field of data science is comprehensively using large set of subject fields.

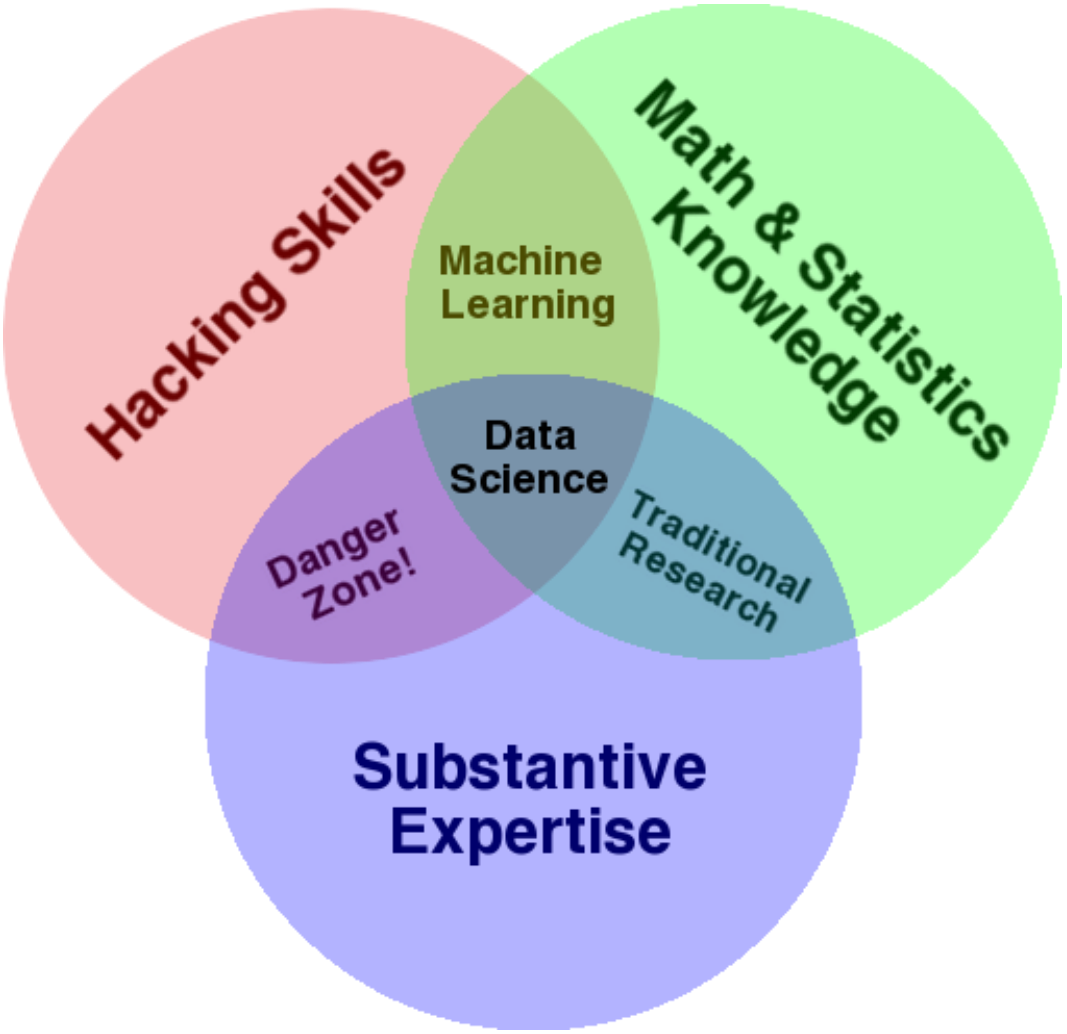


Figure 1.2: Drew Conway's Venn diagram of data science
[37]

Social media data analytics is one of the trending subjects of the current decade. Further, data science is the foundation of many modern technologies, such as machine learning and natural language processing, and it begins with a diverse set of data with various characteristics. Raw data sets are usually dirty and unprocessed, with varying attributes like missing values and outliers. It is possible to overfit or underfit at times. As a result, most data science projects prioritize defining only the required features. Due to the difficulty of the feature selection, many machine learning projects end up with unexpected results.

Generally, Data science has a set of pre-defined methods to use according to the situation. Each technique is discussed below.

- Predictive:To create a predictive model based on data and provide answers to unanswered concerns
- Prescriptive: Actions to perform
- Predictive causal analytics. *Sample Question: A company needs to develop a model that can predict user engagement for a social media post by the organization. Sample Solution: Use the customer's previous user engagement data to forecast future user engagement. Prescriptive analytics is used as a trained model. The model can adjust parameters according to the situation.*

1.8.4 More on analytical types

Larose and Daniel discussed more on the same context in more detailed[38]. Following sub sections are focusing on the same context.

Descriptive analytics

This is the historical data analysis used to define the trend. Concerning only the historical data. It mainly provides insights into the historical trend to understand trends over time. Statistics mainly involves determining social media trends using descriptive analysis.

Predictive analytics

Predictive analytics is involved in predicting future trends. The base for predictive analytics is descriptive analytics. Defining the different enhanced trend lines based on historical data and feature engineering can be helpful. This will predict the future of the same analytics case.

Prescriptive analytics

Prescriptive analytics produces a viable solution for the predictive analytics trend line. Key involvement is proper feature engineering to produce an optimal solution that helps define predictive analytics trends.

** Data may be separated into two groups based on its nature.

- Cross-sectional data : Several variables are included in a single input when cross-sectional data is gathered. These data are frequently applicable to supervised, unsupervised, or neural networking techniques.

- Time series data—information (both multivariate and univariate) gathered using time stamps. Data were gathered at regular periods.

The examination of the data points with time intervals may be done by combining cross sectional data with time series data.

1.9 Getting started with a data science approach

In most cases, the independent variable cannot be predicted just on the data order. Yet, the engagements in time series analysis is matter.

1.9.1 Data Extraction

Data extraction is essential for all enterprises in the age of big data. Data harvesting may provide businesses with many benefits, the most essential of which is that it will move the company into a highly competitive position. The company gains access to up-to-date information about the industry or any related topic through market research using data harvesting. Knowing what's happening in each platform allows your company to react quickly to any changes, reduce losses, and increase decision-making accuracy.

Web Scraping VS APIs

The most practical methods of data gathering are web scraping and API scraping.

Collecting data from a website or single web page is known as "web scraping.". Web analysis and analysis starts with extracting data from web pages [39] . An API (Application Programming Interface) is a set of methods and communication protocols that allow users to access data from a program, operating system, or other services.

API (Application Programming Interface) is a communication interface that allows one piece of software to communicate with another[40]. In simple terms, you can feed a JSON to an API, and it will return a JSON to you. There are always certain restrictions on what you can put in JSON and what it can return.

Web scraping is far more flexible, complicated, and limited by rules.With a web crawling and scraping system, you can obtain any data you observe on a website. When it comes to crawling data, you may use any approach available and are only limited by your imagination. You can create new ways to crawl data from websites with dynamically generated feeds if you have an experienced staff. However, as you can see, websites change their layout daily, so you'll need to update your scraping code occasionally to ensure that everything continues to work.

1.9.2 Feature Engineering

In 2005, ”feature engineering” was coined by Hanchuan Peng, Fuhui Long, and Chris Ding[41]. One of the three criteria in the article’s suggested feature selection approach, named ”Max-Relevance and Min-Redundancy” or ”feature engineering,” was based on mutual information and was one of three.Feature engineering is the core technique for defining valuable features for any novel design-related or data science project. Hence, it is only a complete data science project with feature engineering. On the other hand, feature engineering is necessary for precise and predictable results at the end of the project. Furthermore, the feature engineering process represents the ability to define, refine, modify, or extend required features.

Feature engineering is the process of identifying relevant features for any novel algorithm. Unstructured data is hard to fit for a machine-learning approach as it is. Feature engineering is the process of combining appropriate features with the machine learning model. It implies that feature engineering is the most challenging activity in building a novel algorithm. Feature engineering is mainly applicable when improving a model’s predictive performance by reducing computational or data needs and improving the interpretability of the results.

Raw data is any kind of data format (e.g., textual, graphic, or wave) that can be used to produce some productive output. As is, this raw data cannot be applied to anything. As a result, it must be cleaned. Clean data can be created from new data. Another set of data can be revealed after cleaning the data. The attributes required for the development of an algorithm can be referred to as ”features.”

1.9.3 Feature Engineering for social media data analytics

Social media trending content requires features that support the trend. Feature engineering plays a vital role in identifying the feature that provides a more robust output for trending content identification[42]. I have been involved in feature engineering to identify the feature space in trending social media content. In feature engineering, identifying features more supportive of a prediction and assessing their importance is considerable. This research mainly reveals the most impacted features to predict a social media post’s trending line. This mainly involves the identification of the starting point of the upward trend line and the trend line’s turning point (endpoint). The main focus of feature engineering and identifying mutual information is that significantly impacts the predictive model is revealing new features that support the upward trend.

Feature engineering is a mandatory process when developing intelligent applications[8].

Artificial intelligence model development, neural networks, and deep learning are some of the popular approaches for intelligent application development[43]. Many applications need more proper feature engineering to handle over- and under-fitting. As a result, identifying relevant features is critical. Furthermore, feature decomposition upon variation, encoding highly cardinally categorized content, and feature segmentation are heavily weighted in this study to identify upward-trending social media content.

1.9.4 Feature Engineering on social media information diffusion

Feature engineering is a mandatory process when developing an artificial intelligence model. Many applications experience overfitting and underfitting. [12]These issues are due to not employing feature engineering properly. Hence, identifying useful features is critical. Different social media content trends upward and downward in different trend lines based on user engagement. Data upload/social media posts, user interactions, and reactions are famous examples. Due to varying user preferences, multiplex social media network entries are typical in many user profiles. This information dissemination produces various trending content over time. All platforms are involved in the identification of various social trends.

Trends depend on the location. Most specifically, country- or region-wise trends. Finally, demand is a crucial factor in information diffusion. As a result, it is critical to identify key features for demand for social media content. This research focuses on feature space selection on upwardly trending social media content.

1.10 Graph theory and networks

Networks are based on graph theory, and fundamental concepts of social network analysis will apply to our context[44].

Actors and attributes

Actors are the nodes/vertices that interact with each other. A common actor is a person who uses social media.

Sample node attributes

1. Age
2. Gender

3. Religion

Ties and tie characteristics are the connecting links between the nodes. Further, there is no specific direction for the connectivity. The frequency of social network connectivity is a valuable component in network analysis. To calculate the frequency of the connectivity, we need the following features.

1. Degree: Number of connections it has to other nodes
2. Adjacent nodes: any two nodes connected by an edge

Adjacency matrix

There are mainly two types of representations available.

1. Node list
2. Adjacency matrix

1.10.1 Actors and network-level measures

Node-level measures

Node-level measures enable the identification of a node's significance.
Centrality is one of the measurement

- Total degree centrality (TDC)
- Betweenness centrality (BC)
- Eigenvector centrality (EC)

Following are the information diffusion research questions.

- TDC: Who are the most connected nodes? The direct connection with others
- BC: Who are the brokers (critical actors on the route's path)?
- EC: Who is the most potent actor? (with the most connections to other powerful actors)

There are open and closed communities on social media sites. Open groups have far more influence than private groups. Hence, evaluating public groups rather than community-concealed organizations is essential. User accessibility and platform support are other elements that influence information dissemination.

Graph theory analyzes information flow from one end to the other; time series analysis is used to gauge the rate of information dissemination; and event-driven architecture is used to process data in response to an event. Information diffusion theories are required to support the development of these theories.

A community is a bunch of nodes where the connections are within the network. I.e., the vertex set has a relationship between the nodes. During this research, I experimented with both inter-community sparsity and intra-community density.

1.11 Graph data science for social media analytics

To glean insights from complex network data, the field of research is known as ”graph data science[45].” It can support many applications, including social network analysis and machine learning. Social media platforms provide enormous network data that may be examined to comprehend social interactions, information flow, and user behaviour. Generally, social media analytics is a well-known application of graph data science.

Constructing a social network/ graph, where nodes stand for people or organizations and edges for social relationships or interactions between them. This is one typical method for social media analytics utilizing graph data science. Researchers may learn more about issues like community organization, influencer identification, and the spread of viral material by examining the structure and dynamics of the social network graph[46].

Social network graphs may be mined for useful information using graph algorithms like clustering, pathfinding, and centrality measurements. To find prominent members in a social network, for instance, centrality metrics like degree centrality, betweenness centrality, and eigenvector centrality can be utilised. Users that share similar interests or habits can be found using clustering methods like modularity and community identification. Modeling information diffusion and predicting the spread of information via the network may be done using pathfinding techniques like shortest route and random walk.

The development of recommendation systems, the detection of fraud and anomalies, the identification and mitigation of hate speech and other types of online abuse all use graph data science approaches.

1.12 Graph data science for measuring information diffusion on network

Measuring content diffusion on social media platforms has become the most popular platform for publishing user-generated content in recent years. Furthermore, social media platforms are rapidly replacing traditional sources of firsthand information. User engagement can affect information diffusion on social media platforms. Information diffusion can personalize output for a specific event, allowing people to interact with it more easily. Categorizing information diffusion is essential to understanding the nature of information diffusion. Segmenting information diffusion can help to understand the nature of a post, the trending nature of social media platforms, and many other variances. The most common attributes used for segmenting information are the time of the post, platform, lifestyle, and user behavior for a post. All these segmentations focus on behavioral data science. Interactivity is a well-known method for determining user engagement on social media platforms. This will produce a unique number for each post hosted on any platform. Historical and future data can help us understand the nature and effect of a social media post. That can be effectively helpful in determining the diffusion of information on social media platforms. Apart from generic social media data, social media network analysis produces essential values for the nodes in social media platforms. The two most well-known methods for analyzing a node in a social media network are centrality and betweenness.

1.12.1 Effect of node level features for information diffusion

Many features can affect node-level attributes. Some of them are,

- Text (Comments etc.)
- Images (emojis etc.)
- videos

1.13 AI in SM information diffusion context

To uncover patterns in data, data scientists use a technique known as machine learning. This technique enables computers to identify related data and forecast upcoming events, behaviors, and trends.

Explore big data

Humans are leaving digital traces everywhere.

Data fusion(messy and unfinished) (variety)—the data could not be on the same platform; for instance, not everyone uses Facebook. Simply, this is merging data from several sources to get more complex data. We only obtained a sample because the data is vast but biased.

(Q) In statistics, the word "sampling" describes the process of selecting a subset (a statistical sample) of individuals from a population in order to estimate the characteristics of the whole population. As previously said, "traditional surveys" refers to samples. Why? *Because a survey gathers data from a specific demographic. The methodological tool for contacting everyone is called a "census."*

If we consider data extraction from Facebook, this does not account for all users worldwide. The data that represents Facebook users is biased. Hence, 2 out of every 7 (nearly 30%) humans use Facebook. Facebook, however, needs to provide an exceptionally accurate sampling of the human population. This information is biased. Facebook, for instance, is well known for its age bias. It does not accurately reflect everyone in every age group. The following factors are heavily affecting sampling.

- Dynamic in real-time (strength)
- ML (automatic insights) (automated insights)

The information diffused on social media platforms is based on the user's interests. With user engagements, information may spread at different rates. If more engagements occur in less time, it depicts the viral nature of the spread of information.

Learning Methods

- Supervised Learning: Training data sets specify desired outputs by labeling the feature vector.
- Unsupervised learning: When the desired results are unknown or not included in the labeled feature vector in the training data sets. Finding the data's hidden structure is the goal.

- Reinforcement learning involves the learner interacting with the world through their actions and attempting to determine the best course of action regarding the rewards they will receive from their surroundings.

Training data for semi-supervised learning has a few desirable outcomes.

Supervised Learning Methods

Learning to anticipate a complicated output, such as a sequence or a tree, is called "structural prediction."

Learning inductively

Data deluge: Much data is being generated in real-time that researchers may exploit.

The exploitation can be more successful on scaling vertically, aka Step Up by following features.

- Increase computer size, storage capacity, and memory
- Place-physical limitations.AKA: Horizontal Scaling Scale-out
- Utilize a cluster of several tiny processors
- The core of big data and data science is scaled out.

Data hierarchy is a technique for determining if data is correctly structured and arranged.

The levels of machine learning tasks listed below are the most common research problems in information diffusion.

- Node-level prediction task for information diffusion
- Predicting Link level-task/Edge level
- Prediction of a given node/nodes (pair of nodes) is connected or not
- Graph-level prediction task

Under social media data, the nodes already have a set of attributes attached. A few of them are username, the unique annotated username (@username), and profile details that are shared publicly (Movies that you like, games that you like, stories, and personal information such as birth date, school, etc.)

The additional data that is required are the connections to the node. These connections can be friends, friends of friends, followers, etc. The other ways of having connections in a social network

- Groups
- Pages

These characteristics can be divided into two categories.

- Structural features: features that have a specific structure. E.g., Personal details, Close friend network
- Features that describe the attributes or properties

All of these features are initially used to develop the proposed algorithm.

1.13.1 Social media information diffusion process

Zhang and Philip defined social media information diffusion as the flow of information on different social media platforms in a given time frame[47]. The social media information diffusion process refers to the spread of information or content through social media networks. This process starts with the creation of content by individuals or organizations. The content is then shared through various social media platforms like Facebook, Twitter, Instagram, or LinkedIn.

Once the content is shared, it can be seen and interacted by the followers of the person or organization who posted it. If the content is engaging or relevant, followers may share it with their followers, leading to further dissemination of the information.

Various factors, including network size, content type, and user engagement, influence the speed and extent of information diffusion. For instance, controversial or emotionally charged information is more likely to be shared and spread rapidly than neutral or unengaging content.

Social media algorithms also play a significant role in the information diffusion process. The algorithms are designed to show users relevant and engaging content and maximize the time users spend on the platform. As a result, information deemed relevant or favored by the algorithms are more likely to be seen and shared by users.

Social media information diffusion process is a complex and dynamic phenomenon influenced by various factors, including user behavior, network structure, and platform algorithms. Understanding the process is crucial for organizations looking to communicate effectively with and engage with their target audience through social media.

Based on the information propagated, the trend is a critical factor that significantly impacts society positively and negatively.

Under the context of social media, determining which features are the most important with mutual information sharing, inventing new features in several real-world problem domains, encoding high cardinality categorical with a target encoding, creating segmentation features with k-means clustering, and decomposing a dataset's variation into features with principal component analysis are some of the vital research problems.

1.14 Algorithms used in social media information diffusion process analysis

Various algorithms are used to analyze the social media information diffusion process.

- Network-based algorithms: These algorithms are based on analyzing the structure and relationships of the network in which information spreads, such as identifying influential users and examining the spread of information through different types of ties, such as strong and weak ties. The most popular algorithms can be found under centrality analysis. E.g.:- Eigenvector Centrality [48],Centralises (Betweenness, Closeness etc.) [49], Louvain Modularity [50]

- Information cascade-based algorithms: These algorithms focus on analyzing the spread of information through the network over time, such as identifying the initial adopters and studying information transmission mechanisms. Most popular algorithm are,Independent Cascade Model[51],Epidemic models such as Susceptible-Infectious-Recovered (SIR) Model [52],

- Content-based algorithms: These algorithms focus on analyzing the content of the messages being spread, such as sentiment analysis (E.g., LSA (Latent Semantic Analysis)) [53], theme identification (e.g., Non-Negative Matrix Factorization (NMF) algorithm)[54], and topic modeling (E.g., Latent Dirichlet Allocation (LDA))[55].

- Hybrid algorithms combine elements from network-based, information cascade-based, and content-based algorithms to provide a more comprehensive analysis of the information diffusion process.Linear Influence Model [56], Threshold Activation Model [57]

These algorithms are combined with various data sources, including social media

data, network data, and user-generated content, to gain insights into the information diffusion process on social media platforms.

1.15 Time Series analysis

Time series forecasting for the identification of social media content using traditional machine learning methods is used to identify or predict the popularity of content. The traditional time series analysis is a statistical approach used to look at and analyze the patterns and trends in a data set over time. Time series analysis aims to get valuable conclusions and forecasts from the data. Numerous disciplines, including finance, economics, marketing, and engineering, frequently employ this study field.

Determining the data structure is the first stage in time series analysis. To do this, look for trends, seasonality, and strange patterns in the data. The next step is selecting the model that best explains the time series data. Time series models come in various forms, such as linear, non-linear, and Autoregressive Integrated Moving Average (ARIMA) [58] models.

The ARIMA model, or Autoregressive Integrated Moving Average, is one of the most widely used time series models. Moving average, integration, and autoregression components are used to create this model. The moving average component models the time series' error terms. The auto-regression component captures the linear relationship between the time series data and its lag values. The integration component modifies the time series data's non-stationary characteristics, such as trends and typical patterns.

Estimating the model's parameters is the next stage in time series analysis. To do this, the model must be fitted to the data, and the coefficients that best capture the connection between the time series data and its lag values must be computed. The model may forecast future values of the time series data once the parameters have been evaluated.

Analyzing time series also requires determining how well the model fits the data. To do this, residuals between the observed and expected values must be measured and compared to a statistical distribution. The model is considered well-fit if the residuals are near zero and randomly distributed about zero.

Time series analysis is valuable for determining the patterns and trends in time series data. It is essential for making decisions. Based on time series data, it is widely used in various industries such as finance, economics, marketing, and engineering. This kind of study helps researchers to forecast future values and offers an insightful understanding of the underlying dynamics of the data.

1.15.1 Elements of time series data

The ”trend” refers to the data set’s overall rising or downward tendency. The data covers a considerable amount of time, perhaps several years.

- Cycle: An upward and downward tendency is a cyclical movement.
- Seasonal: A certain day or hour affects the data’s creation. In the time series model, a specific upward or downward trend is thus displayed.
- Seasonal: Heights and valleys are seen Random or erratic data distribution with lingering variations. Generally, for a bit of time.

1.15.2 Time series analysis on social media information diffusion

In social media information diffusion process analysis, time series analysis is used to study the patterns and trends of information dissemination over time. Time series analysis can help researchers to identify key moments in the diffusion process, track the growth or decline of information spread, and identify influential factors that affect the diffusion of information on social media platforms.

One of the most commonly used time series models for social media information diffusion analysis is the Autoregressive Integrated Moving Average (ARIMA) model. This model is based on the assumption that the past values of the time series can be used to predict future values. Other time series models used in this field include Exponential Smoothing (ETS), Seasonal Decomposition of Time Series (STL), and Dynamic Linear Regression (DLR).

1.16 Trend analysis/ Diffusion Analytics

Trend analysis is a projection of patterns that occur due to varied reasons. A trend can be graphed, and/or the trend line can be seen as the main below.

- Uptrend - the trend line is getting an upward value compared to the previous value.
- Downtrend – trend line is getting a downward value in comparison to the previous value

In a real-world scenario, oscillation data will produce a trend. However, the overall trend can be upward or downward. Further, this can extend to short-term or long-term

trends.

The primary type of trend

- Time-based trending
- Location-based trending
- Short-term trend
- Mid-term trend
- Long-term trend

Further, categorize the same context:

- Discreet trend: Independence from representing a trend in content. E.g., user engagements.
- Continuous trend - A collective discreet trend is influencing the content. E.g., reply to a comment.

1.16.1 Techniques used in trend analysis

Statistical modeling of trend

Statistically, a trend can be defined as noisy sample data over a specific period.

- Standard deviation: This is the deviation of the current value from the predicted value.
- Mean: This is the average value of the whole data set.

Moreover, statistical features, AI features, and many other features can help to define the information diffusion trend.

Discussion

- Descriptive statistics: Consider linear regression. This will explain how things are progressing. If we use linear regression, it gives us an equation with a coefficient and co-relation. As a result, we may use it to explain our data, forecast the results of some independent variables, or offer an answer to a query like what the price of an element should be if it increases in value by 10%.

Auto-correlation

The auto-correlation is the similarity between the retrieved data as a function of the time lag between them.

1.16.2 Information diffusion patterns recognition and trend analysis

Identification of the pattern is the main activity in trend analysis. The core difference is that trends have a value and a direction, whereas patterns do not have a guide. A trend is formed when one direction is applied to a pattern. This is important, especially for social media content analyzers, content authors, and invigilators who need to identify the features of trending content to identify early detection of trending content. For example, if an author publishes hate speech-related content that quickly spreads throughout the community, such content should be identified immediately. Then the relevant authorized parties can take the necessary actions to block such content or report it to the appropriate authorities to keep society peaceful.

1.17 Simultaneous information diffusion/Cross-Posting

Cross-posting is one of the most popular techniques in information diffusion among different channels. Cross-posting is sharing identical information across multiple social media platforms. It has a trend line based on user engagement in a social media post over time. The upward trend has more user engagements over time, and there is a peak for each social media post in the upward trend. There are multiple reasons for the upward trend of social media posts. People are drawn to a specific content because of its rapid spread. That means we can partition people in a particular module according to the content nature. Hence, cross-posting is a crucial factor in information diffusion.

1.18 Event-Driven Microservices for the architecture of the system

Overview

Since the current trend runs in agile development concepts and rapid application development patterns, it is crucial to make sure application architecture is developed based

on microservices, and it makes sure the time for "ready to market" is too small. Currently, the monolithic architecture makes huge districts for this rapid development, and software architecture is changing for quickly developed services running their treads using microservices. Further, microservices are optimized for distributed solutions. They can help a large variety of software architecture patterns and frameworks needed for the subsequent era of application development by using automation and artificial intelligence.

The proposed architecture is supposed to run on artificial intelligence-based self-executables identified, analyzed, and predicted for the execution of specific microservice to ensure the services are up and running seamlessly. The proposed architecture should ensure it can run on a pluggable environment to ensure the software development time is quieter. In this case, many microservices components are developed and automated to acquire the architecture that is more robust and reliable architectural software development. Further explanation of the proposed architecture, it can run on the cloud environment to make sure it makes sufficient solid on enterprise-level applications to make the software packaging and deployment via automating by intelligence services; further, the artificial intelligence-based microservices can serve any level of software development components like user experience, site reliability engineering, software reverse engineering, etc.

This architecture can handle any self-evaluative perspective, and it can heavily depend on components requiring heavy data loads."Plug and play" software components are essential to ensure the minimum time to develop. It reduces the requirement of high-end technical requirements of the enterprise application development company, which are on medium and minor scales. In that perspective, the proposed architecture can run with the minimum human errors and minimize software bugs which can introduce current issues to the application. Microservices are already developed, and it is really to use by using a few lines to the applications. It can help to expand the software with other third-party software integration in a secure manner.

Concluding Remarks

Social media platforms play a crucial role in information dissemination by removing many traditional boundaries. Hence, social media content has its own pace of information propagation. More demand increases user engagement, spreading social media content at a different strengths. Once the information applies on social media platforms, users initiate arrangements depending on the user's interest in that information. If more engagements occur in less time, it depicts the viral nature of spreading information.

Moreover, each social media platform has its nature of information dissemination. Facebook, YouTube, and Twitter are popular social media platforms that engage in Sri Lanka. Facebook has a common strategy of sharing information based on close relationships known as ”friends” or open groups for topics of particular interest. Twitter is the largest micro-blogging site, providing direct access to the entire community via hashtags. Twitter has followers and keeps consistent updates on relevant topics. YouTube is an open community video-sharing platform that allows access to information retrieval based on the search and the home page utilities.

Information diffusion on social media platforms creates an engagement trend based on user engagement—more engagement in the upward and downward directions. Once demand is lower, the trend line will move downward accordingly.

The rest of the chapters are arranged as below.

- Chapter 2 - The background section focused on the existing literature related to the problem domain. In the current literature, the author focused on developing the research gap.
- Chapter 3 - Methodology and the various methods involved in this research work. High-level system design will be discussed in this chapter.
- Chapter 4 - The analysis section focuses on analyzing the data and optimizing the proposed algorithm. Overall implementation will be discussed during this chapter
- Chapter 5 - Evaluation will discuss during chapter 5. Mainly how the proposed algorithm is strong to handle real-world implementation will be discussed in this chapter.
- Chapter 6 - The conclusion section provides a conclusion of the research work.

Chapter 2

Literature Review

2.1 Social media networks

A wide range of methods are used to calculate information diffusion, and the following research areas were identified as relevant literature. Topic-based information diffusion, Social influence, Herd behavior, and Information cascade. Apart from these standard methods, time series analysis was considered to calculate the adjacent speed of the information diffusion.

Graph theory provides the interrelation among the diffusion, whether it is an open social network/ platform or a closed social network/platform. Moreover, graph theory represents the adjacent information diffusion. Finally, the background on event-driven architecture (Event-driven microservices) provides the computational architectural model for the implementation of the identification of information diffusion.

2.1.1 Characteristics of ties

Direction-based networks (semantic networks)

There are primarily two types of networks:

- Directed networks
- Undirected network

Inbound and *outbound* network connections are available on these networks.

Ties are the connecting links between the nodes; connecting links do not show a specific direction. Nevertheless, there is a specific direction to information diffusion in the context of information diffusion. As a result, this research considered all diffusion networks to be directed networks.

Many metadata and node attributes are available for any given social network. Under network connectivity, the following key features were considered:

- Similar user types with a defined boundary are called ”user similarity.”
- Location: the user is in the same particular temporal space, etc.
- Membership: users in the same memberships, groups, etc.
- Attributes: same likeliness, same gender, etc.

Moreover, social relations/social connections can be described below.

- Kinship
- Friend
- Affective: like the same content.
- Cognitive: a reaction to the content
- Interactions, Recommendations, Suggestions, etc.
- General flows. E.g. - Information flow, Resource flow, Cognition, Interaction flow

The frequency of social network connectivity is a valuable component in network analysis.

2.1.2 Social Network Sites

Social network graphs are the graphical representation of nodes and their interactions. These are digital platforms that connectivity providers in the digital environment have many engagements like adding new friends and novel designs proposed for visualizing the social network data. These are highly dynamic and novel ideas for dynamic social graphs based on SNSs. Mainly SNSs followed a standard set of methods as below.

They established traditional social science methods, such as surveys and interviews, to gather information and process data in real-time.

APIs and AI are used to keep connectivity with the physical entities in the digital environment (keep tie strength firmly). Dyad connections are the minor social groups that connect and novel design, suggesting social media tie strength in common aspects.

Nohuddin et al.[59] discussed a novel framework for analyzing a social network based on frequent pattern-mining techniques. Significantly, the trending behavior of

the uses on epochs. Self-organizing based maps are a novel idea required to update the necessary trending behavior in trend analysis, known as "Total-From-Partial (TFP)." Specifically, TFP maps can identify the trending data to seamlessly manage a large cluster of data sets. TFP can help to cluster the data sets and related informative high volume of data to split and update the trends accordingly.

2.1.3 Propagation of information on social media platforms

Social media platforms use AI, and in most cases, each social media platform tracks user behavior, and based on the user interactions, it will provide recommendations in its platform. A novel "double layer" recommendation system based on density clustering was introduced. Moreover, it is a standard method to keep posting synchronized on several social media channels.

Avetisyan et al. [60] added an overview of information diffusion models. The authors of this study investigated the patterns of information diffusion on social media platforms and the factors that impact user-sharing behavior.Stefan Stieglitz and Linh Dang-Xuan [61]investigate the role of emotions in information distribution on social media platforms. This review article discusses the role of emotions in information spread on social media platforms. The authors discuss contemporary studies on emotional contagion, emotional arousal, and emotional valence in the context of information diffusion.Li et al. [62] investigate the impact of source trustworthiness and message valence on social media information dissemination. They research Weibo, a popular Chinese social media platform, to investigate the relationship between user characteristics, message attributes, and information diffusion. In this study, the authors look at the impact of source trustworthiness and message valence on information transmitted on social media.

Uses and Gratifications Theory

Blumler and Katz's [63] uses and gratifications theory proposes that media consumers contribute and have a role in the media. The user has four simple needs

- Diversion (escape from everyday life problems)
- Personal Relationships (using media for emotional and other interactions)
- Personality identity (Reflecting yourself in texts, learning behaviors and values from articles)
- Surveillance (Articles/information could be helpful for a living)

To examine the motivations for social media use and information sharing, ”Uses and Gratifications Theory” approaches have been explored. It’s also been utilized to look at the impact of content quality on information distribution [64].

2.1.4 Software-based Social network analysis

Often, software-based applications like Networkx and Gephi are widely used in computations of SNA to identify and analyze the graph-based social networks running in highly varied environments. Besides the analysis of SNA, it is mandatory to have proper visualization since the majority of the case is easy to explain with visualization techniques, and a comprehensive analysis conducts regarding software applications that are running with SNA done by Naeem[65].

2.2 Tree structure and Graph theory in information diffusion context

2.2.1 Tree structure for information diffusion analysis

The paper ”The data-driven null models for information dissemination tree in social networks” [66]provides a novel method for analyzing information diffusion in social media using a tree data structure. The authors argue that the proposed technique provides a more efficient and precise way of analyzing information diffusion compared to existing methodologies.”A note on modeling retweet cascades on Twitter” is an interesting discussion [67]. The authors of this article propose an approach for predicting Twitter retweet cascades based on a diffusion model based on trees. This study uses a tree structure to describe the dissemination of information on social media as an epidemic process[68], with each node on the tree standing for a possible or actual recipient of the information. A method to forecast information cascades in social media, this study suggests a tree-based technique, where each node in the tree represents a user who may or may not be impacted by the cascade [69]. This study uses a tree-based approach to rumor identification and tracking, where each node in the tree represents a piece of information that may or may not be a rumor [70]. Another study proposed a tree-based method to measure information dispersion in social networks, where each node in the tree represents a user who may or may not have received or supplied the information [71].

In this study [72], the authors propose a tree-based paradigm for analyzing knowledge propagation in online social networks. Each node represents a user who may

or may not have participated in the diffusion process [73]. This research describes a tree-based approach for detecting prominent players in social networks, with each node in the tree representing a user who is either a key player or not. In another article [74], the authors propose a tree-based technique for analyzing information diffusion on social media. Each node represents someone who may or may not have shared the information.

In another study [72], the authors propose a tree-based approach for analyzing the dissemination patterns of social media communications. Each node represents a user who may have shared the message [75]. This study proposes a tree-based modeling and analytic approach for examining information distribution in social networks, with each node representing a person who received or gave the information [76]. In this study, the authors propose a tree-based approach for analyzing information diffusion on social media. Each node represents someone who may or may not have shared the information.

Another study [77] proposed an effective tree-based diffusion model for measuring information propagation in social networks. This study introduces an effective tree-based diffusion model for analyzing information dissemination in social networks. Each node in the tree represents a user who received or passed on the information. Another study [78] present a tree-based technique for assessing viral propagation on social media. The authors offer a tree-based technique for evaluating viral diffusion in social media, with each node representing a user who may or may not have shared the viral material. Another study [79] create a tree-based model to better understand the dispersion of information in social networks. The study proposes a tree-based model for understanding information propagation in social networks, with each node representing a person who received or gave the information.

2.2.2 Graph theory for information diffusion analysis

Another study [80] provided a graph-based approach for identifying influence diffusion networks and community structure in complex networks. The authors provide a graph-based method for inferring the influence of diffusion networks and community structure in complex networks, where nodes represent users and linkages represent information dissemination. Another study [81] describe a graph-centric approach to finding and analyzing communities in social networks that may be used in studies on information dispersion. This paper describes a graph-centric approach to recognizing and analyzing communities in social networks, in which nodes represent users and edges indicate interactions between them. The authors also show how this approach may be applied

to analyze knowledge propagation in social networks. A study [82] described a random graph walk-based method for identifying significant nodes in social networks, which may be used to analyze information diffusion. In this study, the authors describe a random graph walk-based technique for locating significant nodes in social networks, which might be used for information dissemination analysis. They show that their technique beats earlier centrality-based approaches.

Another article [83] describes a graph-based approach for modeling information transmission in social networks over time. The authors propose a graph-based approach for modeling information transmission in social networks across time in this article. Using graph partitioning techniques, they analyze the evolution of communities and find influential nodes [84]. Further presented a graph-based algorithm for investigating social impact and dissemination in complex networks. The authors present a graph-based paradigm for investigating social effects and dissemination in complicated networks. The authors identify critical nodes and investigate network information flow using centrality measures [85]. This study the diffusion of information in social networks using a graph-theoretic technique. The authors of this study analyze information diffusion in social networks using a graph-theoretic framework. They show that network characteristics such as degree distribution and clustering coefficient can significantly influence the rate and reach of information diffusion[86]. Further proposes identifying opinion leaders in Twitter networks using graph theory. In this article, the authors propose a graph-theoretic approach for finding opinion leaders in Twitter networks. They use centrality metrics to identify nodes that significantly influence information dispersion in the network [87]. This explores information propagation on Twitter using a graph-theoretic technique. This study investigates the diffusion of information on Twitter using a graph-theoretic framework. The writers analyze the features of the Twitter network and identify the most significant members[88]. This paper describes a graph-theoretic approach for detecting relevant members in social network knowledge spread. In this study, the authors describe a graph-theoretic approach for detecting relevant actors in information transmission in social networks. They use centrality metrics to identify the network's most influential nodes.

2.3 Time series analysis

Time series analysis has four components. The level is the horizontal history of a product. In other words, the level is the pattern if there is no trend, seasonality, or noise. The trend is a trend of increases or decreases over a specific period. Further, the

trend is the main component of forecasting. Seasonality is a repetitive pattern with a specific set of factors, and patterns can have a linear pattern or a curved design. These are commonly known as cycles in the pattern[89].

Noise is a random fluctuation. Though this is an unexpected opportunity, this can be explained by some qualitative techniques. Only the situations that would not occur in the past. Dynamic data analysis is one of the significant concepts in time series analysis. Based on time series analysis and different sociological modifications, and based on time series analysis, there is a high ability to forecast the future. Based on the time series analysis, it is mandatory to visualize the data set. One method is subgrouping based on the data sets. Once the deviation completes, the only movement is required to conduct the expected time series analysis to forecast the value[90].

2.3.1 Time series techniques

Time-series techniques can categorize into mainly two parts. The objective of these techniques is how they identify and define the main patterns in time series analysis. [91]

- Open model time series techniques
- Fixed model time series techniques

Open model time series techniques (OMTS)

OMTS involves the following steps.

1. Analysis of the patterns
2. Build a model to the time series (This model is mainly involved with projecting the patterns in the future)
3. Forecast time series

Fixed model time series techniques (FMTS)

FMTS has a predefined equation. The equation derives from a prior definition of the time series component. Upon losing assumptions for the definition of FMTS, it might need to be revised in some situations.

Attributes of FMTS

- Simple

- Inexpensive
- Less data quantity required

Since FMTS requires only a small amount of data, FMTS can use for short-term forecasting models.

FMTS forecasting models

Average based models

The forecasting involves the average value of the current data set. It applies to data sets containing level and noise and can be expressed below [92].

$$y\hat{t} + 1 = \frac{1}{n} \quad i = 1^{n} y_{t-i+1} \tag{2.1}$$

Explanation

The above equation, $y\hat{t} + 1$ based on the average. This equation reflects the prediction for the following time of the most recent n observations ($yt - i + 1$). In notation, $\sum_{i=1}^{n}$ confirms that we're summing up the most recent n observations, giving the output by dividing n. Hence it gives us the average value. The advantage is removing the noise and the forecast level of the values set. In contrast, the main disadvantages are lagging once the dataset grows and assuming no trend or seasonality.

Moving average

Moving average calculates the most recent average values of the data set. The forecasting can depend on the number of components. Usually, the period calculates for three periods or four periods. Nevertheless, this can extend to more periodical values like five or six periods to N number of periods. Once N is the total number of periods of the data set, it will convert to the mean[92, 93, 91].

$$y\hat{t} + 1 = \frac{1}{n} \quad i = 1^{n} y_{t-i+1} \tag{2.2}$$

Explanation

This equation, $y\hat{t} + 1$, represents the forecast for the next period. (i.e., it averages the value for the next value). The notation $\sum_{i=1}^{n}$ confirms that we're sum-up the nearest n observations, and dividing by n gives us the average value. Further, It is based on the average value of the latest n observations ($yt - i + 1$).

Disadvantages

- Losing the seasonal components once the period equals the sequence number (n)
- All the periods are getting the same weight
- Assume no trend or seasonality

Exponential smoothing average

The exponential smoothing average is the technique that adds weight to the average moving technique. Each datum is distributed with a specific weight(α). Based on α, the total distribution can be calculated more accurately than the simple moving average. Once $\alpha = 0.5$, it calculates the values' meaning accordingly[94, 93].

Adaptive smoothing

To calculate the α, calculate the absolute error of the previous value, and then the error can use to forecast the value as the following data element. Hence, this can note as the below function. The following equation is for adaptive smoothing.[91, 94]

$$\alpha_{t+2} = \left|\frac{F_{t+1} - S_{t+1}}{S_{t+1}}\right| = |PE_{t+1}| \tag{2.3}$$

In this situation, the required values can adapt to α. —PEt+1— *can have a wide range of values. Hence the following rules should apply to get the proper output.*

$$\begin{cases} \alpha_{t+2} = 0.99999 if |PE_{t+1}| \geq 1.0 \\ \alpha_{t+2} = 0.00001 if |PE_{t+1}| = 0.0 \end{cases}$$

Exponential smoothing with trend

Adaptive smoothing can enhance by adding a trend for forecasting and usually smoothing, having a well-defined smoother than the average adaptive smoothing methods[95].

2.3.2 Software support for time series analysis

Predicting social media usage with a time series is an important aspect.

Facebook introduced Prophet[1] for the analysis of time series. This tool is highly confident with trends based on the time series. All the above methods mainly focus on data sets with a limited scale, and "Prophet" is mainly targeted for a large-scale data set with vertical scaling methods. Further, the Prophet has well-defined statistical methods

[1] https://facebook.github.io/prophet/

and other restructuring methods for real-time forecasting. Tensorflow [2] is another well-defined framework for time-sensitive model development, and tensor flow has inbuilt libraries for timestamp-based analysis. Keras [3]provides a framework for time series data processing [4]. Further, the extended data calculation with data preprocessing ability integrates with Keras. Hence Keras is a well-defined framework for adapting a time series analysis model.

For identifying personalized search result generation techniques for semantic information retrieval, topic modeling based on latent Dirichlet allocation [96] is combined with topic-driven community detection methods.

Ontologies emerged as an alternative to databases in applications requiring a more ”enriched” meaning [97].

2.4 Time series analysis in information diffusion context

Zhang et al.[98] use time series analysis to analyze the diffusion of information in social media networks. The authors of this study use time series analysis to evaluate information propagation via social media networks. They propose a new diffusion model based on the SIR (Susceptible-Infected-Recovered) epidemic model and investigate a time series of information propagation. Time series analysis to investigate the dissemination of information on Twitter during the Euro 2012 football tournament. Alperin et al. [99] used time series analysis to investigate the spread of information on Twitter during the Euro 2012 competition. By analyzing the temporal patterns of retweets, the authors use time series models to predict the dissemination of information. Hong et al.[100] provide a time-series approach to modeling and forecasting social network diffusion processes. The researchers propose a time series approach for modeling and forecasting diffusion processes in social networks. They use autoregressive integrated moving average (ARIMA) models to analyze information transmission time series and estimate future trends [101] . Further, they used time series analysis to study online communication trends in a Parkinson’s disease support group. In this paper, the online communication patterns in a Parkinson’s disease support group are explored using time series analysis. The authors use time series models to estimate future activity by analyzing the temporal patterns of message posting. Using time-series analysis[102]. Moreover, explore Twitter data related to the 2012 U.S. presidential election. The authors of this paper use time series analysis to look at Twitter data related to the 2012

[2]https://www.tensorflow.org/
[3]https://keras.io/
[4]https://keras.io/examples/timeseries/

U.S. presidential election. They use time series models to estimate future trends by analyzing tweeting and retweeting tendencies across time [103]. Further, they used time series analysis to analyze the propagation of political memes on Twitter during the 2016 U.S. presidential election. This study used time series analysis to examine the propagation of political memes on Twitter during the 2016 U.S. presidential election. The authors use time series models to estimate future trends by analyzing the temporal patterns of retweets.

2.5 Machine Learning algorithms

Naive Bayse (NB) Classifier

The naive Bayes classifier, one of the most often used classifiers in the sentiment analysis process, is one of the Bayesian classifiers based on Bayes theory. It also features a probabilistic classifier that can look at pre-categories or examined data sets to understand the pattern and process. A phrase or passage's likelihood of being positive or negative can be assessed using the NB classifier. Assuming that the existence of one feature in a class is unconnected to the existence of other characteristics, this classifier also estimates the probabilities of each feature. NB classifiers are easy to use, very efficient, and only require a modest amount of data to implement. Nevertheless, the data's accuracy could be higher since the classifier is based on assumptions and contains dependencies.

The Naive Bayes Classifier has been used to examine information dissemination trends on social media. It effectively forecasts information spread and identifies notable users [104].

Support Vector Machine (SVM)

Support vector machines, a supervised machine learning method, may train and categorize reviews using a well-known sentiment polarity model (SPM) [105]. SVM separates n-dimensional spaces into subspaces, groups the subspace's vectors according to their shared polarity or properties, and then creates hyperplanes. SVM produces faster and more reliable results without much training data.

The SVM algorithm has been used to investigate how information spreads on social networking platforms. It effectively forecasts information spread and identifies notable users [106].

K Nearest Neighbor (KNN)

Because it is simple and quick to build, the K Nearest Neighbor (KNN) [107]supervised machine learning algorithm is frequently employed in classification and regression forecasting problems. The KNN technique is known as a ”lazy learning algorithm” since there is no training phase, and training is done concurrently with classification. KNN is sometimes called a non-parametric algorithm since it uses no underlying assumptions or data. A KNN case will be categorized according to a point system that examines similarities between new and old occurrences.

The K Nearest Neighbor algorithm was used to investigate information distributed on social media platforms. It is effective in anticipating information spread and identifying notable users[108].

Decision Tree

The decision tree used in the nonparametric supervised machine learning approach helps deal with classification and regression problems. The model contains a root node that symbolizes the start of the decision tree. These internal nodes reflect features obtained from data sets, branches that indicate the decision rules, and leaf nodes that denote the conclusion, similar to a tree-structured classifier. The class or polarity of the target variable is also predicted using a set of predetermined conditional criteria and the decision tree approach. These conditional rules comprise a succession of understandable if-then-else phrases [109].

The diffusion of information on social media platforms has been studied using decision trees[110]. They help identify the primary factors glorifying the spread of information.

Maximum Entropy Algorithm

A well-liked technique for estimating the chance that a data set’s characteristics will be distributed in a particular way is the maximum entropy algorithm [111]. Moreover, it employs a categorized weightage approach to assign a weight to each feature found in the given text document before categorizing it according to the chance that the feature will appear throughout the material.

The maximum entropy approach has been used to analyze content disseminated on social networking platforms[112]. The algorithm is beneficial in anticipating the spread of information by considering a range of criteria.

2.5.1 Residual Analysis

Residuals occur mainly due to variations that the model does not explain. There are two reasons for variation that are not explained.

- Pure random noise: pure random noise cannot be eliminated. Hence, this is unpredictable.
- Identification of random noise: Upon satisfying the following factors, the model can describe a random noise
- Linearity: The model has residuals that randomly distribute.
- Independence – All residuals are performing an independence
- Normality – All residuals are distributed as a Gaussian distribution
- Equal variance – variance is constant among residuals

2.5.2 Trend analysis and forecasting

The key output of the trend analysis is defining the future behavior of the same data sets. Forecasting is mainly considered a baseline to detect upcoming occurrences in the same context.

An artificial intelligence-based application for a simplex social network using an artificial neural network (ANN) is a general research problem in social media data analytics. While expanding artificial intelligence more towards attempting conjunction with deep learning techniques, the novel features are considered towards more accurate prediction models.

2.5.3 Feature engineering

Extracting the most relevant features is a critical concern in feature engineering. Critical importance is the nature of the feature, like continuous or discreet features, and a selection of the most appropriate according to the situation. Further discussion is conducted regarding discretizing the continuous data under supervised and unsupervised learning. A novel model was introduced based on the constrained-based feature selection process and some other key factors and how those are properly feature-engineered using the novel framework. Feature crossing is highly impactful in high-dimensional social media data analysis. Hence, large clusters of social platforms required analysis based on feature crossing. Furthermore, deep learning concepts are highly impactful for

the optimized feature engineering process because the automated feature engineering process is explicitly implemented in “deep learning” methodologies.

In social media platforms, feature engineering has been utilized to improve the accuracy of information dispersion studies. The algorithm can better predict the distribution of information and identify essential people by extracting and choosing significant portions[113].Using temporal parameters in the model can assist in increasing the accuracy of information dissemination analysis [114]. By analyzing the period of information propagation, the software may better capture the dynamics of information distribution on social media platforms.

2.5.4 Artificial neural networks

Artificial neural networks are the leading software implementations that are required to develop a solid model to predict social media trends. Chatting is one of the discussion methods in social media networks, and early identification of threats is essential to take suitable preventive action and reduce the impact. A novel framework based on SVM and NN to identify predatory conversations. Deep neural networks are another crucial factor for the identification of trends. Popularity increment with time, has an upward trend, and mainly consider in a trending context. Based on feature engineering concepts, a novel framework introduces the identification of popularity based on multi-variant attributes. The identification of cyberbullying is considerable in social media networks. A deep learning-based identification framework introduces by addressing critical questions related to cyberbullying. Social networks are high domination rapid data-generating platforms, and data extraction is mainly based on APIs, which were developed and exposed to the environment by respective social media platform developers to involve other developers to extract data in real-time for their applications to use those data in a helpful manner. Moreover, these non-Euclidian networks generate a massive amount of data, and adequately feature engineering and embedded into low dimensional data is another aspect that is required to evaluate a social media platform, and these drawbacks can solve by using reinforcement learning. Reinforcement learning techniques are applied to identify social reputation/ social rewards and how the interactive social media posts help to perform well recognition in a social media context. Selected articles from many journals and papers. Then check the usage of different data mining algorithms. Provided a scientific method to perform a review article/literature survey.

In the era of social media, deep learning plays a vital role in predicting user engagement trends.

On top of ANN techniques, some of the most recent research innovations propose

novel hybrid architectures for social media data predictions.

High trends in user engagements, like hate speech, can involve social unrest, and there can be trends that can be analyzed based on burst analysis techniques. To predict the bursts of multiplex social networks, an LSTM model was introduced. These designs are more supportive of multiplex social media platforms.

A deep learning-based user engagement model was developed to predict user engagement-based trends.

According to Zhang et al.[115], artificial neural networks may be used to predict information dissemination in social media.In a study by Naeem et al., artificial neural networks were utilized to predict information propagation on Twitter [65].

2.5.5 Statistical analysis

2.5.6 Techniques used in trend analysis

The modeling of trends mainly considers the following factors:

- Standard deviation
- Mean

Standard deviation

Standard deviation is the deviation of the current value from the predicted value. Forecasting can adjust with the standard deviation.

Mean

Mean is the average value of the entire data set. Statistical forecasting for trend analysis Statistical forecasting is mainly concerned with identifying patterns in history and using these patterns to predict the future. The critical assumption is that “the trend has a seasonality.” Demand patterns focus on stationary and completely random fluctuations around the mean level. In this case, the residuals fluctuate among the data. Data values distribute without proper patterns.

Cycles

Cycles are a combination of seasonality and trend. The trend line has a set of curves that have similar points.

Regression analysis

Regression analysis is a statistical model that can predict a dependent variable or multivariable, and regression analysis is the most widely used in forecasting a trend. Regression analysis uses a dependent variable) with an independent variable, and the independent variable is commonly known as the explanatory variable. Further, regression analysis is the primary tool for identifying causal relationships.

Single linear regression

Only one independent variable is used in linear regression to forecast the dependent variable.

Multiple linear regression

Predicts a single dependent variable with multiple independent variables.

2.5.7 Error handling methods

Least square method

This method uses to minimize the error of the trend line by minimizing the sum of squared errors. It calculates the difference between the observed value and the predicted value. Then it squared and found the best possible solution. Research on YouTube video audience, view count, likes and emotion . Furthermore, it was established that there is a significant association between stay duration and unfavourable feelings. One of the pervasive issues in the area of machine learning is classification. Moreover, logistic regression and random forest are tried-and-true techniques for binary classification. [Logistic Regression vs. Random Forest for Binary Classification]

2.6 Information diffusion models

Topic-related information diffusion The objective of topic-related information diffusion is to keep society open to interacting[116]. Under this context, information diffusion focuses on a specific topic. Many platforms provide information diffusion aspects such as hashtags and mention to specify the content[117]. A popular example is ”Diffusion of innovation theory”[118]. Hence, this type of diffusion is visible in open communities[119].

2.6.1 Herd behavior

Social behavior occurs when a group does the same activity without necessarily disregarding their information messages[120]. Further, a study related to herd behavior describes the ties and how the strength of ties affects herd behavior[121].

2.6.2 Graph learning

Graph learning is the best implementation for analyzing the information diffusion in a given network[122]. Graph learning provides a mathematical representation of information diffusion[123]. Especially implementations like neural graph networks are beneficial for creating a more robust and reliable algorithm to define information diffusion[124].

2.6.3 Information Cascade

People in a social network adopt the information based on the followers and the information generated by the primary node. Sharing the information among the close network is evident. This behavior is known as information cascading[125]. Information cascading has a more significant influence on sharing information. Among their close networks, the information diffusion rate is relatively higher than other methods. Moreover, a study was conducted on the Twitter platform, and a novel prediction method[126].

2.6.4 Network-based algorithms used in social media information diffusion process analysis

Social media information diffusion process analysis uses information cascade-based methods to examine how information or influence spreads within a network. These algorithms are based on a ”cascade,” which describes the orderly transfer of knowledge or power from one node (person or thing) in a network to another. Analyzing information diffusion processes in social media using network-based algorithms is standard practice. These algorithms concentrate on the network’s links between persons and groups and how information moves via them. Several network-based methods are often employed in social media information dissemination process analysis.

One of the earliest and most basic models for information transmission in social networks is the Independent Cascade Model [127, 128]. It assumes that a node only has an opportunity to activate its neighbors after activation. The Independent Cascade Model simulates the diffusion of knowledge or influence as a random process, where

each node has a chance of activating and disseminating knowledge to its neighbors. According to the Linear Threshold Model, [129, 56], each node has a threshold value, and information can only reach a node if more or equal to that node's threshold value of neighbors is active. Each node in a network has a threshold in the linear threshold model. If the impact of its neighbors reaches the threshold, the node becomes active and begins disseminating information. The susceptible-Infected-Recovered (SIR) Model is a classical epidemiological model widely used to model information diffusion processes in social networks[130]. It considers the status of nodes as susceptible, infected, and recovered, and the spread of information follows the SIR process. Spread-based Centrality[131] method assesses a node's significance based on how far its information has been disseminated. Impact Maximization [132] is a well-known approach for identifying the group of nodes with the most significant potential for information dissemination. This technique aims to identify the network nodes with the most tremendous potential for initiating the dissemination of information. The condition of its neighbors influences the node's threshold in the non-linear threshold model. This model considers the interconnectedness of network nodes[133]. In the weighted linear threshold model[134], each node in a network has a threshold and an influence weight, and the effect of its neighbors is weighted through these influence factors. The effect from a node's neighbors is a weight matrix of the numerous criteria each network unit possesses according to the multi-linear threshold model[135].

2.6.5 Content-based algorithms for information diffusion analysis

The main topics of the content-based algorithms employed in social media information diffusion process analysis are the message's substance and how it distributes over the network. Content-based algorithms include, for instance:

- Hashtag-based analysis: This technique uses hashtags to monitor how information spreads on social media sites like Twitter[117, 136].
- This method, known as Latent Dirichlet Allocation (LDA) [96], is used to determine the subjects covered in a collection of documents, including social media messages [137]. LDA can aid in understanding the communications' substance and how they are disseminated.
- Text classification: In this technique, text-based information—like social media posts—is categorized into predetermined groups. The categories may then be used to spot patterns and trends in disseminating knowledge.

– Sentiment analysis: This technique includes categorizing social media messages using algorithms into good, harmful, or neutral categories depending on their feelings or ideas.

2.6.6 Hybrid algorithms

Another method is the employment of hybrid algorithms in the investigation of the information dissemination process in social media. In order to analyze the dissemination of information via social media, hybrid algorithms integrate elements of both network-based and content-based algorithms. Popular hybrid algorithms include the following:

- Edge-weighted information cascades (EWIC): In order to analyze the spread of information, this technique employs edge weights that are based on both network structure and content.[138]
- Hybrid diffusion model (HDM):HDM employs network topology, user behavior, and content attributes to forecast information propagation in a social network[139].
- Network-content integration (NCI) Network structure and content data are both used by NCI to examine how information spreads through social media[140].
- Social network analysis and topic modeling (SNATM): This hybrid technique combines topic modeling and network structure analysis to examine how topics spread via social media[141].

2.7 Techniques for Text Preprocessing

NLP trends

Consumer feedback is obtained in an unstructured, free-text manner; as a result, the data must be translated into a structured format before being categorized and examined. This method is known as ”text preparation.”

Text can be preprocessed using various techniques, including stemming and lemmatization, tokenization, part-of-speech tagging, and stop-word removal. The following sections will go into great detail on each of the tactics mentioned above.

The authors conduct a literature review on integrating natural language processing with social network analysis for information dissemination studies. They begin with an introduction to NLP and SNA, followed by a review of studies [142] that employed

both approaches to investigate various aspects of information dispersion, such as recognizing notable persons, detecting rumors and fake news, and anticipating information dissemination patterns. The authors also look at the issues and promise of merging NLP with SNA and future research directions in this subject.

Speech Element Tagging

The practice of categorizing parts of speech into linguistic categories based on how they are used in sentences is known as ”parts of speech tagging.”

Text Mining Algorithms

Naive Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME), K Nearest Neighbor (KNN), Decision Tree, and Linear Regression are only a few of the algorithms and techniques used in text mining. A few widely used algorithms are explained in depth in the following section.

Text mining technologies such as topic modeling and sentiment analysis have been used for information dissemination analysis [143]. Li et al. [144] examined the content of online reviews using text-mining technologies to study the impact of word-of-mouth marketing on product sales.

2.7.1 Summary of Research Areas, Findings, Methods, and References in Social Media Analysis

Research Area	Findings and Methods Used	References
Social Network Sites	Traditional social science methods like surveys and interviews used for data gathering and processing. Utilization of APIs and AI to maintain tie strength. Dyad connections and novel designs for visualizing social network data. Novel framework based on frequent pattern-mining techniques for social network analysis. Total-From-Partial (TFP) concept for trend analysis and data clustering.	[59]
Propagation of information on social media platforms	AI-based recommendation systems for user behavior tracking and content recommendations. Introduction of a ”double layer” recommendation system based on density clustering. Study on patterns of information diffusion and factors influencing user-sharing behavior. Investigation of the role of emotions in information distribution on social media platforms. Research on the impact of source trustworthiness and message valence on social media information dissemination.	[60, 61, 62]
Uses and Gratifications Theory	Blumler and Katz’s theory identifies four simple needs driving media consumption behavior: Diversion, Personal Relationships, Personality Identity, Surveillance. The theory is applied to examine motivations for social media use and information sharing. It’s utilized to study the impact of content quality on information distribution.	[63, 64]

Table 2: Summary of research areas, findin5g5s, and methods used in the context of social media analysis.

Existing tools

A program called a ”sentiment analysis tool” may automatically and without the need for technical knowledge analyze customer reviews in a more straightforward form for its users to understand. It can also determine the sentiment, feedback, or opinion underlying the reviews. Several review analyzers are available on the market, but some of the more well-liked ones include Brand4, Quicksearch, BrandWatch, Repustate, and MonkeyLearn. The traits, advantages, and disadvantages of some of the applications mentioned above are discussed in the following sections.

According to social networks and related social issues, the evolution of terminology is another considerable aspect. Diachronic analysis (evaluating linguistics evolving with time) is critical in social media’s linguistic trend analysis. Last decade, social media imposed several restrictions on text, especially on words harmful to society or words or phrases used as ”dirty words.” The restrictions mentioned above are destructive to some social media uses. Social media users generate novel words by using Sinhala or based on the textual representation in Sinhala but written in English words. In each version of the social media platform, several words/phrases are prohibited by the owners of social media platforms. However, some words can be unrestricted due to their non-dictionary and non-meaningful nature in traditional society. Due to this glitch, several social media users are using alternative language from time to time, and diachronic analysis helps identify those linguistically morphological expansions.

On the other hand, synchronic analysis (the lexical analysis regardless of time) will define the inherited words from the Sinhala language. These are the words used in the Sinhala language as synonyms but have different meanings. These words do not evolve with time. Nevertheless, it has no proper meaning in the Sinhala dictionary too. Somehow, those words are available. Open discussions are one of the main features of social media networks. Based on sentiment analysis, they keep the critical factors required to adapt to social network data, as opinions are essential.

According to Kim et al.[145], software frameworks may be used for efficient and scalable processing of large-scale text data for NLP trend analysis. Apache Spark is a well-known software framework for NLP trend analysis used in a wide range of NLP workloads [146] In their study, Khan et al. [147] performed NLP trend analysis on social media data using a range of software frameworks, including Hadoop and TensorFlow.

Software frameworks for NLP trend analysis

Much software is available for textual analysis, and most software packages support more than one natural language, including but not limited to Apache OpenNLP and NLTK. These software packages are supported for numerous features like non-destructive tokenization, Named entity recognition.

However, the Sinhala language has yet to consider any of this software. Hence, the software is highly demanded for the analysis of Sinhala language text processing, including morphological analysis. Wu et al. proposed a novel framework for analyzing the NLP based on the specific text with benchmarking and precisely analyzing the textual representation in tweeter data. Statistical analysis.

Computational mathematics is another branch of deriving social network behavior analysis and trend analysis. Further, a novel social media experimental feature matrix is introduced to handle ideal social networks that work in a dynamic environment. Another aspect of trend analysis is the relationship between two quantitative variables. Primarily, the trend is based on the variance of the hypnotized variables. Statistical modeling in trend analysis is another crucial aspect. Statistically, trends can be defined as noisy sample data over a specific time.

2.8 Event driven architecture

The event-driven architecture is one of the most sophisticated software architectures that provides streaming, event-driven data processing. Especially when it comes to event streams, the most robust solution is event-driven microservices.

EDA is a technique for creating software systems that respond to events or messages rather than the traditional request-response architecture. EDA uses events to start software components, creating a highly scalable and loosely coupled system design [148].

EDA's ability to handle real-time event processing is one of its key advantages, making it a popular choice for systems that demand speedy and continuous data processing. Another benefit is that it is modular, allowing for adding or removing components as needed without affecting the rest of the system.

EDA is commonly used in a microservice architecture, where each microservice is responsible for processing certain events [149]. This enables the development of highly specialized services, which can then be readily coupled to construct sophisticated applications.

Nevertheless, EDA has its own set of issues, such as guaranteeing data integrity

across dispersed systems and processing events in the proper order [148].

Despite these obstacles, EDA has grown in popularity in recent years because of its capacity to manage enormous amounts of data in real time. EDA is predicted to be essential in developing modern software systems as more firms embrace event-driven architectures.

2.8.1 Event-driven micro-services

Event-driven microservices focus on creating micro-services with event triggers. These events are executed according to the trigger microservices and execute events. The execution speed is strong evidence for the information diffusion speed.

An event-driven microservice architecture is a type of microservice architecture that uses an event-driven technique to handle communication and data exchange between microservices. In this design, microservices communicate by exchanging events, which are frequently short packets of information containing data and metadata about a specific event or action that has occurred.

One of the benefits of event-driven microservices is their ability to expand and manage vast volumes of data and events [150]. It states that the event-driven strategy allows for a more scalable and fault-tolerant architecture than traditional request-response systems. The event-driven microservices' loosely coupled design also allows for greater flexibility and faster service deployment and maintenance.

Another essential characteristic of event-driven microservices is message brokers, who act as middlemen between services and allow events to flow. These message brokers can manage massive amounts of data and include features like message durability and guaranteed delivery, ensuring that events are processed and sent to the appropriate services constantly. Two typical message brokers used in event-driven microservices are Apache Kafka and RabbitMQ [151].

Event-driven microservices offer a scalable and customizable architecture for handling huge volumes of data and events. Using an event-driven architecture and message brokers, microservices may communicate with one another in a loosely connected and fault-tolerant manner.

2.8.2 Event driven architecture in information diffusion analysis

Woo et al.[152] describe an event-driven micro-services system for social media information dispersion analysis that uses Apache Kafka as a message broker and Apache Flink for real-time data processing. The system is designed to collect and analyze so-

cial media posts and their distribution patterns, providing valuable data for marketing and social media analytics.[153]. It proposed a system for sentiment analysis of social media data based on event-driven microservices and machine learning algorithms. The system can process massive volumes of social media data in real-time, providing rapid and trustworthy insights into customer sentiment. Moreover, event-driven microservices were used to analyze sentiment in social media data.

2.9 Model error evaluation

Testing and evaluating information dissemination analysis approaches is necessary to determine their efficacy [**?**]. Diffusion models may be evaluated using accuracy, recall, and the F1 score [154].

There are several reasons to perform an error evaluation for any model. For instance, an error evaluation can capture variables the model may have missed during development. Additionally, it can help identify issues that might have arisen from using the wrong methodology or errors due to random noise. Furthermore, an error evaluation can help assess the model's validity, particularly when residuals are present.

Further following testing methods are related to information diffusion analysis

- Cross-validation evaluates the performance and generalizability of machine learning models by splitting data into subgroups for training and testing [155].
- A/B testing is a controlled experiment in which the efficacy of two or more treatments or interventions is compared [155].
- ROC analysis is a graphical method for evaluating the efficacy of binary classification models by presenting the actual positive rate vs. the false positive rate [156].
- Precision-Recall (PR) curve analysis is a graphical method for evaluating the performance of binary classification models by plotting precision vs. recall [157].
- The F1 score is a statistic that combines accuracy and recalls into a single value to assess the performance of binary classification algorithms [158].

Concluding Remarks

This research project focuses on developing an information diffusion analytic approach for multiplex social media platforms. The study provides an overview of available

techniques and tactics for measuring social media information transmission, such as network analysis, subject modeling, and machine learning approaches. The research focuses on the limitations and constraints of these approaches. It presents a framework that integrates them to comprehensively analyze information dispersed across multiplex social media sites.

The proposed solution emphasizes the need to consider the dynamic nature of social media platforms, such as the diversity of user behaviors and preferences, the different structures of the social network, and the numerous communication routes. According to the findings, the proposed framework might be a valuable tool for academics and practitioners interested in learning more about the complex information diffusion processes in multiplex social media platforms.

Chapter 3

METHODOLOGY

3.1 Data extraction in social media platforms

3.1.1 Overview

Large-scale real-time social media analytics enables the development of prediction models under certain conditions. Users are used as training and test instances, and when they converse in online forums, lexical traits and network elements about those users' interactions are gradually made public. We provide a range of handling methodologies, from traditional batch training and testing to incremental bootstrapping and active learning to estimate latent user qualities from this dynamic data.

In some cases, large-scale, real-time social media analytics allow for the development of prediction models. While they chat in online forums, individual users serve as training and testing subjects for language patterns. We also look at the relationships between a substantial sample of users in an online social network who exhibit various predicted user traits, opinions, and emotions.

Initially, I contrast user tweets' emotional profiles with user demographics and personality. The relationships between the anticipated user characteristics and the user-environment dynamic contrast assessed over a range of neighborhoods, including friends, retweets, and mentioned persons, are then examined. Last, we assess and contrast how latent user attributes, emotions, and interests effectively predict the portrayal of arrogant behavior and self-promotion.

The extraction of data from social media platforms is one of the major concerns for researchers. Each social media platform has many application programming interfaces (APIs) to extract this data. This research discusses the likelihood of a given post being distributed across social media platforms. This research focused on Facebook, Twitter,

and YouTube for a single social media post. Put, conduct a data analysis across multiple social media platforms. The most common scenario is based on API-based extraction. All social media platforms offer a set of APIs that allow data to be extracted from the back end. To keep people interacting, it is necessary to consider the combined impact of various social media platforms. An automated algorithm can find out the total effect of a single post. This might have an impact on other social groups. Semantic search strategies can enhance the total search results.

Social media users often use different platforms to propagate information among their networks. According to the social impact, there will be a series of times and uses in which people are interested. It is a widespread practice to post to all social networks simultaneously. Then the users will react accordingly. This is a common scenario for information propagation. The demand analysis used by different models to have streaming data analytics and no SQL-based database querying can help find the social impact of a social media post. In today's world, many applications require a large amount of streaming data to make decisions on various aspects of different social media platforms that work on the impact of each social media platform. Social media platforms use many mechanisms to add impact to any post. The number of likes generated in a network of clusters.

Users in the same cluster are curious about the impact of the posts. Authors are taking the sub-instances of things like arguments, discussions, and opinions while ignoring the context of the problem. In other words, arguments and viewpoints can be expanded to include political and regional perspectives on the main topic.

3.2 Research design

A mixed methods study methodology is frequently used in social media information distribution analysis to provide a comprehensive understanding of the phenomenon [159]. Mixed methods research combines quantitative and qualitative research methodologies to address research questions that cannot be adequately addressed by either methodology alone [160]. In social media information diffusion analysis, quantitative approaches are used to evaluate the patterns and characteristics of information distribution. In contrast, qualitative methods investigate the meaning behind these patterns and characteristics [161].

Sequential explanatory design is a popular mixed methods research design in social media information diffusion analysis, which involves collecting and analyzing quantitative data first, followed by collecting and analyzing qualitative data to provide a more

detailed explanation of the quantitative results [160]. Researchers can use quantitative analysis to find patterns and trends in data and then employ qualitative analysis to explain why these patterns and trends exist. Other mixed methods research designs commonly used in social media information diffusion analysis include concurrent triangulation design, which collects and analyses quantitative and qualitative data concurrently to provide a complete understanding of the phenomenon, and sequential transformative design, which changes the research design based on the initial findings.

3.2.1 Solution Overview

The following image 3.1 represents a high-level overview of the application. As it depicts, the entire application is designed based on four main pillars.

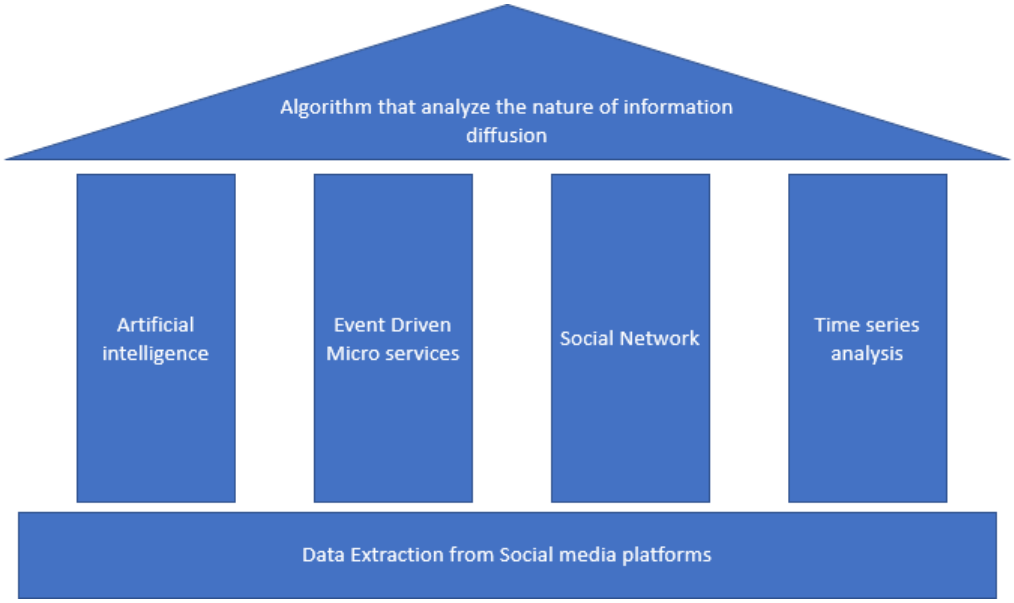


Figure 3.1: Design Overview

Explanation of the solution overview

As 3.1 depicts, the proposed solution is based on mainly four factors. A short explanation of each factor is given below.

- Social Network - A social network for social media information diffusion analysis can assist us in understanding how material flows and is consumed on social media platforms. By examining social networks, we can identify key influencers and

assess the impact of various types of content on consumers. This information may be used to develop more effective marketing, advertising, and communication strategies in the digital age.

- Time series analysis - This is essential for social media information dispersion research because it allows us to track and understand how information flows over time. By evaluating time series data, we may discover patterns and trends in how information travels and how people engage with it. This information might be used to develop more effective strategies for monitoring and regulating material published on social networking platforms.

- Machine learning- Social media information diffusion analysis is crucial because it allows us to manage and analyze vast amounts of data swiftly and precisely. We can use AI to discover patterns and trends in how information moves and how people engage with it, allowing us to develop more effective strategies for regulating and managing the material distribution on social media platforms. AI can also help us automate procedures like sentiment analysis and content classification, allowing us to analyze social media data more efficiently.

- Event-driven architecture - It is essential for social media information analysis since it allows for real-time processing and analysis of social media data. We can detect events and trends with Event-driven architecture and respond quickly and effectively to emerging problems. This is especially true in crisis management, where timely and accurate information is critical for public safety and decision-making.

3.3 Data collection methods

3.3.1 YouTube live data extraction

Initially, the YouTube API for data extraction should be enabled. Then the data should be extracted as an HTTP response.

Steps you need to follow to extract data from YouTube API

Steps

- Enabling YouTube data API: All YouTube data extraction can be performed via Google Cloud. To access the Google Console [1]

[1] https://console.cloud.google.com/home/dashboard

- Then, a new project needs to be created.

3.3.2 Extraction of Data from YouTube Channels in live streaming

1. Link the Google Collab with the Google Drive
2. Mount the Google Drive on Google Collab
3. Install the required libraries.
4. Import the packages
5. Set the YouTube Parameters
6. Snippets
7. Statistics of the Channel
8. Content Details of the Channel
9. ‘Playlist Items
10. Analyzing the YouTube extracted Data

Steps to activate API from Google console

1. Create a New Project in the console.
2. Login to Google, make an account if you don’t already have one, and then log in.
3. Create a new project from the top of the page, as seen in the image below, and then select ”Select a project” from the top:
4. Now Click on “New Project.”
5. You’ll be instantly directed to the Google APIs dashboard after you’ve finished.
6. The next step is to activate the YouTube API, which can be found in the side panel under API & Services.
7. Then, at the top of the page, select Enable API & Services.
8. Select YouTube Data API v3 after searching for YouTube.
9. Then, as seen in the image below, enable the API by clicking the Enable button.

Channel statistics extraction from YouTube

One of the other methods is a direct HTTP call to the service end point2. The result will be an HTTP response. The relevant JSON object will return if the call is successfully executed. Following is a sample response JSON object.

- You are required the channel id
- The key is the YouTube API key

  Then an API key needs to be generated. A sample API key can view as follows.

  API key - AIzaSyCTH8yqSkChtMmnEmKdBsVm4n3BxSNrsLg


```
{
  "kind": "youtube#channelListResponse",
  "etag": "iFqqp8g3ykhTB-ixqnqWINQ0KD4",
  "pageInfo": {
    "totalResults": 1,
    "resultsPerPage": 5
  },
  "items": [
    {
      "kind": "youtube#channel",
      "etag": "poNfGiWWth5IOgKL9y3vkvMrOQ4",
      "id": "UC5IDHX2sg9Hg_FPkXGUj_gw",
      "statistics": {
        "viewCount": "60939730",
        "subscriberCount": "476000",
        "hiddenSubscriberCount": false,
        "videoCount": "63"
      }
    }
  ]
}
```


Figure 3.2: Response JSON Object

Authentication using YouTube

The developer should have a Google Cloud Project host to get API keys. Then the user can create a new Google Cloud Project and activate YouTube Data API Version 3. Once the data API is activated, the required API keys can be generated.

Under this research work, I have extracted data from trend lists in 25 countries.

2https://www.googleapis.com/youtube/v3/channels?part=statisticsid=[vidoeID]]key=[Your API Key]]

Country selection

YouTube provides an API for extracting trending YouTube videos. A sample API response is below. The ten most popular countries, ten least popular countries, and ten average countries were chosen. Country list is generated with ISO 31663 4

Using the YouTube Data API, this application looks for ”python programming” videos and gathers details, including the video’s title, description, thumbnail, and URL. The collected data is stored in a list of dictionaries and may be further processed or written to a file as necessary.

Steps followed

- Create a new YouTube project in the google cloud console Then it will create a new project (You may get the project information)
- Go to APIs overview
- Click enable APIs
- Select YouTube data API V3
- Click Enable button to enable the API
- It directs to the overview screen
- To use the API resources, it is mandatory to create credentials. Hence click and enable the credentials. Select YouTube API
- Cloud will generate a key as below.
- The application will redirect to the following screen.

3.3.3 Limitations in data extraction

Some of the countries do not have specific YouTube trending list YouTube is banned in some countries, such as South Sudan and Iran.

Note - Besides authentication, this code also needs the Google-auth and google-api-python-client libraries installed.

3https://www.iso.org/iso-3166-country-codes.html
4https://www.iso.org/obp/ui/#search

Limitations in YouTube Data API

Since data extraction depends on platform-specific limitations, YouTube Data API also provides specific limitations for data extraction.[5]

3.3.4 Twitter data extraction

Twitter data extraction can perform mainly in two methods

- Twitter API
- Web Scraping

Scrapping or API

Twint[6] provides Twitter scraping service without the Twitter API. Mainly focus on solving the limitations provided by Twitter API. Hence that is one of the options for data extraction from the Twitter platform.

Twitter provides an extensive API-based service for data extraction and backend system updates. Hence, I have yet to use scraping throughout the research project. Only used API service to perform all the required activities.

Steps to create a Twitter bot

1. Sign up for a Twitter developer account and create a new app.
2. Generate the necessary API keys and access tokens for your app.
3. Use a library or package for your chosen programming language to interact with the Twitter API.
4. Write code to automate tweeting, following, or liking tweets, etc.
5. Test your bot and make any necessary adjustments.
6. Deploy the bot to a server or hosting service so it can run continuously.

Before building your bot, read and comprehend the tight regulations and standards Twitter provides for bots. A bot's ethical implications must also be considered, as must its compliance with all applicable legal and regulatory requirements.

[5] https://developers.google.com/youtube/v3/getting-started
[6] https://github.com/twintproject/twint

How to get these authentication tokens in the Twitter platform For Twitter, users should have a developer account. Use the following URL to get the developer account[7]. Twitter provides the following keys to consuming resources.

- Consumer Keys
- Authentication Tokens

3.3.5 Facebook data extraction

Developers must access all resources though Authentication service Authenticating Facebook To consume resources hosted on the Facebook platform, developers should have a developer account[8][9].

To get authenticated to this platform, the user should have the following authentication tokens from Twitter, Facebook, and YouTube. The authentication service is the immediate service call in the application[10].

3.3.6 Facebook graphs API

The Graph API is the primary way to call a Facebook account[11]. Graphs explore[12] one of the simplest methods to perform differently: https://developers.facebook.com/tools/explorer/ aspects. Conversely, developers can use Graphs Explore as a sandbox application.

Apply the knowledge of API research and analyze the alternative solutions similar to the proposed system that a suitable API could enhance. Design an application that will utilize a range of APIs for the proposed solution and justify the design choices used.

- Data extraction through scraping for YouTube
- The process of importing information from a website into a spreadsheet or a local file saved on your computer is known as data scraping. It is also known as web scrap. It is one of the most effective ways to collect information from the web and, in some situations, to send that information to another website.

Here are the main steps to achieve this,

7https://developer.twitter.com/en
8https://developers.facebook.com/
9https://developers.facebook.com/docs/pages
10https://developers.facebook.com/docs/pages/getting-started
11https://developers.facebook.com/docs/graph-api/overview
12https://developers.facebook.com/tools/explorer/

1. ”Go To Web Page”- to open the targeted web page
2. Create a ”Loop Item”- to loop, enter searching keywords
3. Dealing with infinitive scrolling
4. Create a ”Loop Item” -to loop extract each item
5. Extract data - to select data, you need to scrape
6. Run extraction - to run your task and get data

Google offers a variety of APIs, each of which has a specific application in many areas. This makes work easier when developing mobile applications, websites, and other projects. Google’s YouTube Data API v3 is one such API.

3.3.7 Extract using API - general information

Although each product is new, the following features set the gold standard for scraper APIs:

- It uses a headless browser to create JavaScript and access the HTML code behind dynamic webpages.
- Has a large proxy pool, ideally in the hundreds of thousands, of data center and residential proxies;
- Rotates proxies automatically while offering the user the option of static proxies.
- To blend in with ordinary views, the application employs anti-captcha features.
- Data is delivered in JSON format.

Web scraping APIs are excellent solutions for organizations with extensive software architectures and smaller designs as long as the users have some coding skills. Companies that rely on price intelligence and product data will benefit the most from data extraction.

Eventually, I realized one of the simplest methods yet ubiquitous for data extraction is using *hashtags*

Moreover, platform-defined trending is added as an additional feature to determine whether it is in the trending list. As an example, youtube trending list[13].

13https://www.youtube.com/feed/explore

3.4 Social network analysis

The Social Networks and their attributes will address social network analysis topics related to examining health behaviors, such as,

- Data collection
- Ego network analysis
- Diffusion and peer influence
- Communities in networks
- Respondent-driven sampling
- Network visualizations
- Statistical Models for networks (ERGM, AMEN, SOAM)
- Agent-based modeling

Furthermore, cross-promotion targets customers of one item or service with comparable advertisements for another.

3.4.1 ML algorithm overview

Image 3.3 describes the databases-related analytics. The entire design was focused on Redshift, a cloud-based data warehousing and analytics system provided by Amazon Web Services (AWS). It allows users to store and query huge amounts of data in an economical, scalable, and secure manner. Redshift is designed for online analytical processing (OLAP) and can execute complex SQL queries on massive datasets. Organizations and organizations of all sizes widely use it for data storage, business intelligence, and advanced analytics. Redshift has data compression, columnar storage, automatic backups, and the flexibility to scale up swiftly or down the number of nodes in a cluster.

3.4.2 Contagion approach for information diffusion analysis

While developing this algorithm, the author wanted to adopt a contagion approach in social media. According to the contagion method in social media, information, sentiments, and behaviors may spread swiftly and extensively across online networks, much

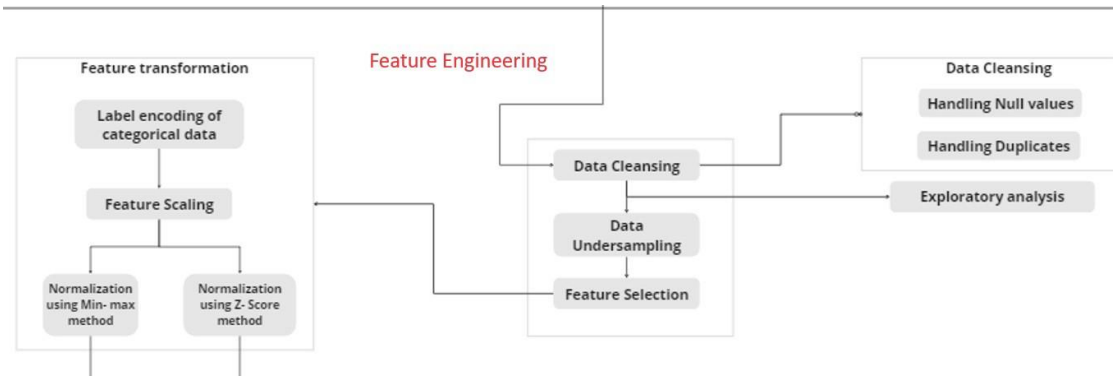


Figure 3.3: High-level feature engineering process for data analytical service (Selected based on the key factors

like a contagious illness. This approach views social media users as interconnected nodes in a network, each of which has the power to influence and be influenced by others.

As per the figure 3.4, an experimental data set was used to perform contagion, and that is an excellent approach to identifying how the information diffusion works. A schematic of the social media contagion approach is displayed in the image. The figure's central node, which stands for a social media user, is surrounded by additional nodes representing their connections or followers. The central node is labeled "User X," while the peripheral nodes are labeled "User Y," "User Z," and so on.

By displaying the influence or flow of information between users, the pathways that link the nodes illustrate how information or behaviors may spread via the network. The arrows on the lines indicate the direction of the effect, with some pointing from User X to User Y and others pointing in the other direction.

3.5 Architecture the system for Event-Driven

The primary methodology of the proposed solution is based on artificial components which can straightforwardly handle microservices. Mainly developing artificial bots to integrate large-scale applications simply. Hence, it will get pluggable micro-services to ensure it runs on an independent platform and can scale in and out vertically and horizontally. This target methodology is the primary path for developing structured microservices-based applications to run with acceptable content by using artificial intelligence. Finally, it will develop microservices like (plugins) for the architecture to integrate with upcoming microservices components to ensure the application architecture is robust enough to handle upcoming technologies. Docker is a platform as

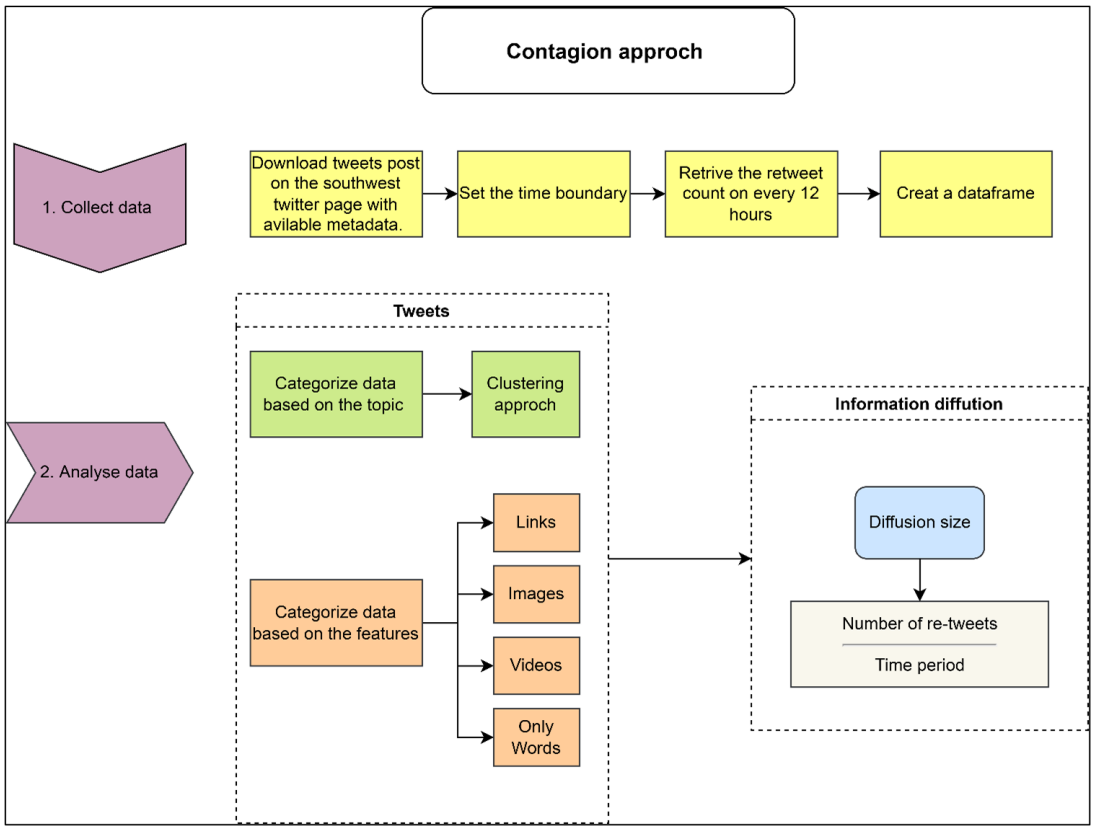


Figure 3.4: Adopting to contagion approch

a service for virtualization. I can get some software packages, develop them on local machines, and then provide deployments in different containers according to the requirements.

Figure 3.5: High-level overview of the proposed system architecture

Figure 3.5 describes the overview of this event-driven architecture. A sample algorithm (LSTM) has been added to explain the overview of this application. The entire application was focused on many features to define the event trigger. When an event occurs on time t=0, architecture is designed to keep track of the event from t=1 onwards until the trigger ends. Further, figure 3.6 describes how this system was focused on multiplex social networks.

RUL, or Remaining Useful Life, finds applications in various fields, with maintenance and reliability engineering being the most common [162]. It signifies the estimated operational lifespan of a component or system, especially in the context of an object's validity within a data structure, such as a graph. RUL prediction primarily aims to determine the remaining lifespan before the object is removed from the data structure. This is similar to how a garbage collector functions in many software programs, optimizing memory usage. This information can be leveraged to improve overall data structure organization.

Figure 3.6: An overview of proposed RUL situation

Figure 3.5 describes a situational overview of the entire application that takes url as a sample and focuses on how it considers the structure of the layered network generating an output for the system. As per this layered architecture, The information flowed in the most controlled method.

3.5.1 Relationship with the microservices and proposed algorithm

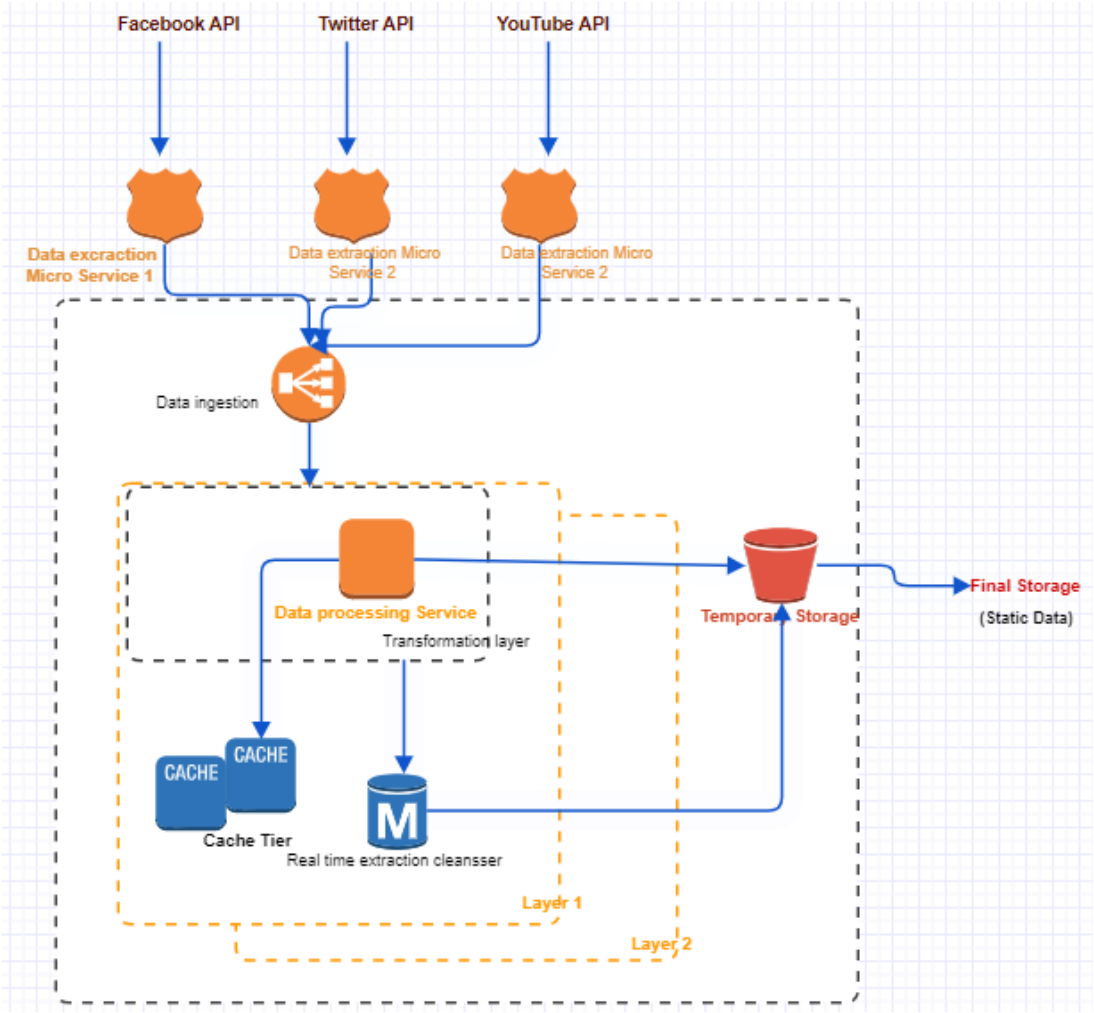


Figure 3.7: An overview of proposed event-driven microservices architecture

Figure 3.7 describes the primary usage of event-driven architecture. In the proposed system, individual social media networks operate as separate, independent entities, each providing specialized services for data extraction. A unique service call is created for each network to extract data from these networks. This approach necessitates the implementation of an event-driven microservices architecture to streamline the process.

The fundamental advantage of employing an event-driven architecture lies in its

ability to selectively activate services based on events occurring within a specific social media platform. Only the corresponding service is triggered when an event transpires on a particular platform, optimizing efficiency and resource utilization. Additionally, this event-driven architecture offers the flexibility to seamlessly extend the system to accommodate other social media networks. The same method can be employed to integrate and interact with diverse social media platforms, ensuring a consistent and streamlined data extraction process across the entire system.

Concluding Remarks

The process of gathering and analyzing data from different social media platforms, including Facebook, Twitter, Instagram, LinkedIn, and others, is known as data extraction on social media platforms. Market research, social listening, sentiment analysis, and other business-related uses frequently employ this technique.

Using specialized tools, software, APIs, or web scraping to collect data from social media networks automatically is the norm for data extraction. User-generated material, such as posts, comments, and reviews, as well as user demographic data, such as age, gender, and location, can be included in this data.

Data extraction from social media platforms may be a laborious and complicated process that calls for knowledge of both data analysis and social media marketing. Nonetheless, it may offer insightful information about customer behavior and preferences, aiding firms in making more knowledgeable choices.

It is crucial to remember that, to prevent legal problems, data extraction from social media platforms must be done per the platform's terms of service and local data privacy laws.

Chapter 4

ANALYSIS

Algorithm Design

The algorithm focuses on the data ingested in a cloud repository to deal with published and user engagement time.

4.1 Proposed Algorithm

Information Diffusion Analysis framework for multiplex social media platforms

4.1.1 Overview of the algorithm/Design an algorithm

Design the proposed algorithm that depicts the information diffusion in multiplex social media platforms. Critical components in the algorithm,

- A short-term trend analyzer is a component that analyzes the information diffusion less than three months of the period. Further, this case considered the seasonal factor as another trend component.
- Mid-term analyzer analysis of the information diffusion happens more than three months of the period but less than nine months, and this will help to identify the residual analysis of the project.
- Long-term analyzer focuses on an analysis of the information diffusion more than nine months to several years

Implement the algorithm The proposed algorithm will implement in a live environment and identify the information diffusion in relevant occurrences.

An array of social media postings P with specific identification, a timestamp, and several tags are used as the input. The result is a collection of diffusion trees T, where

each tree represents the diffusion of a single piece of data and comprises nodes for postings and edges for diffusion paths.

Discussion

- Collecting data entails learning about user profiles, connections between individuals, and information flow on various social media sites.
- Network construction: Create multiplex social networks using the data acquired, considering the numerous connections between and information-dissemination patterns of each platform.
- Information propagation in multiplex networks should be modeled, considering interactions between different platforms and different types of linkages.
- Extract community structures, temporal trends, and centrality metrics from multiplex networks to understand how information is spread.
- Analysis and interpretation: Examine the retrieved traits, looking for patterns in communication and interpersonal ties.
- Validation: Verify the analysis' conclusions using additional data sources, such as surveys or interviews, and change the framework as necessary.

This algorithm uses several platforms and user relationships to comprehend how information is disseminated across numerous social media platforms.

Data gathering

Social media data analytics is crucial in identifying and defining proper social media usage. Especially without a well-defined feature engineering process, some critical factors are missing accordingly. Moreover, collecting high-quality social media data is considerable. Hence, defining features and related data sets provide considerable social media usage statistics and impact accordingly. Moreover, data science and artificial intelligence are underlying architectures to analyze and output data results adequately. Moreover, Define the social media content trends using descriptive analytics to analyze historical trends and other social media trends that can be used in predictive analytics and artificial intelligence to create future trends. Further, a well-defined architectural framework in cross-data employment requires different social media platforms Sri Lankan people use. Finally, prescriptive analysis deploys to identify future requirements and removing social media data to avoid social splits in Sri Lanka. Various online

repositories provide social media data for research purposes. Some of the popular ones include:

- The SNAP (Stanford Network Analysis Project) [1]
- Kaggle [2]
- UCI Machine Learning Repository [3]
- Harvard Dataverse Network [4]
- Microsoft Research Open Data. [5]
- The Web Data Commons Project [6]
- Figshare [7]

These archives include a wide range of social media data from platforms like Twitter, Facebook, YouTube, and others that may be utilized for various research projects, including sentiment analysis, information dissemination, and more.

procedure DiffusionTree(P)
    Sort P by timestamp
    $T \leftarrow \emptyset$
    for each post p_i in P **do**
        Start a new diffusion tree T_i with a single node for p_i
        for each post p_j in P with a date later than p_i **do**
            Calculate the similarity of posts between p_i and p_j
            if similarity is greater than a predetermined content-specific/platform-specific threshold **then**
                Add a directed edge from node representing p_i to node representing p_j in T_i
            end if
            if p_j has already been added to a diffusion tree T_k in T **then**
                Merge T_i with T_k
            end if

1http://snap.stanford.edu/data/index.html
2https://www.kaggle.com/datasets
3https://archive.ics.uci.edu/ml/datasets.php
4https://dataverse.harvard.edu/
5https://msropendata.com/
6http://webdatacommons.org/
7https://figshare.com/

end for
Add diffusion tree T_i to T
end for
return T
end procedure

Sample implementation

Listing 4.1: Sample implementation of diffusion tree algorithm

```
def diffusion_tree(posts):
    # sort posts by timestamp
    posts = sorted(posts, key=lambda post: post.timestamp)
    # initialize an empty set of diffusion trees
    trees = []
    for i, pi in enumerate(posts):
        # start a new diffusion tree with a single node for pi
        Ti = {pi}
        for j in range(i + 1, len(posts)):
            pj = posts[j]
            # calculate the similarity of posts between pi and pj
            similarity = calculate_similarity(pi, pj)
            # if the similarity is greater
            if similarity > threshold:
                # Add a directed edge
                Ti.add(pj)
            # check if pj has already been added
            for Tk in trees:
                if pj in Tk:
                    # merge Ti with Tk
                    Ti = Ti.union(Tk)
                    trees.remove(Tk)
                    break
        # add diffusion tree Ti to T
        trees.append(Ti)
    return trees
```

Each piece of information in the collection of posts is given its diffusion tree, which is how this method operates. A diffusion tree is a directed tree in which every node

corresponds to a post and every edge to the diffusion channel between posts. Before iterating through each post in sequence, the algorithm sorts the posts by timestamp. It creates a fresh diffusion tree for every post and looks for posts with similar tags but with later timestamps. A directed edge is added from the node representing the earlier post to the node representing the later post if a similar post is discovered. The current diffusion tree is combined with the previous diffusion tree if the later post has previously been added to one. The method then returns the collection of diffusion trees.

Following equation satisfies the given requirements.

Let P be the collection of social media posts, where each post p_i has a timestamp(i.e. a published/reacted time) t_i and a set of tags (Under this research we are focusing only hashtags) T_i. Let T be the set of diffusion trees (i.e. each social media platform will generate a single diffusion tree), where each tree/each social media platform T_i represents the diffusion of a single data point and it is developed based on nodes for postings/time stamps and edges/vertices for diffusion paths.

Lemma

The enumerated conditions are satisfied by the following equation.

Sort P by timestamp: $P \leftarrow \text{sort}(P, t_i)$
Initialize an empty set T: $T \leftarrow$
for each post p_i in P **do**
Initialize a new diffusion tree T_i with a single node representing post p_i: $T_i \leftarrow p_i$
for each post p_j with a timestamp later than t_i **do**
Compute the similarity between the tags of p_i and p_j: similarity $\leftarrow$ similarity(T_i, p_j)
if similarity is above a certain threshold **then**
Add a directed edge from the node representing p_i to the node representing p_j in T_i: $T_i \leftarrow T_i \cup (p_i, p_j)$
end if
if p_j has already been added to a diffusion tree in T **then**
Merge T_i with the diffusion tree containing p_j: $T \leftarrow T¥T_j \cup \text{merge}(T_i, T_j)$
end if
end for
Add the diffusion tree T_i to the set of diffusion trees T: $T \leftarrow T \cup T_i$
end for
Return the set of diffusion trees T: T

In here, sort(P, t_i) is the base function that is responsible for sorting and the use

posts P by timestamp t_i, and similarity(T_i, p_j) is a base function that is responsible for compute the similarity between the hashtags of the given post p_i and post p_j in the diffusion tree T_i. The merge(T_i, T_j) function is merges diffusion trees T_i and T_j by adding required all nodes and edges from T_j to T_i.

A sample implementation can find in Appendix.(A.1). An overview can find below.

```
def diffusion_tree(P):
# Sort P by timestamp
P = sorted(P, key=lambda x: x.timestamp)

# Initialize an empty set T
T = set()

for i, p_i in enumerate(P):
    # Initialize a new diffusion tree T_i with a single node
    p_i
    where, T_i = {p_i}

    for p_j in P[i+1:]:
        # Compute the similarity between the tags of p_i and p_j
        similarity = compute_similarity(p_i.tags, p_j.tags)

        if similarity > threshold:
            # Add a directed edge from the node
            T_i.add((p_i, p_j))

        for T_j in T:
            if p_j in T_j:
                # Merge T_i with the diffusion tree containing p_j
                T.remove(T_j)
                T_i = merge(T_i, T_j)
                break

    # Add the diffusion tree T_i to the set of diffusion trees T
    T.add(T_i)

# Return the set of diffusion trees T
```

return T

4.2 Adding temporal aspects using time series analysis.

We can propose an algorithm that deals with temporal situations by adding time series analysis to the above base algorithm. Some of the specific use cases are seasonality. The way we compute the similarity between posts must be changed to incorporate time series analysis into the prior approach. We must also consider the postings' timestamps in addition to just their tags. Using a function that gives more recent postings a higher weight is one method.

Algorithm with embedded time series analysis

Require: Social media posts (P), similarity threshold (s), and weight function (w)
Ensure: Diffusion trees (T)
function EmbeddedTimeSeriesAnalysis(P, s, w)
Sort posts P by timestamp
Initialize an empty set T of trees, where each tree represents a simplex social network
for each post p in P **do**
Initialize a new tree T_p with a single node representing p
for each post q with a timestamp later than p **do**
Calculate the similarity between p and q as $w(p, q) \times sim(p, q)$, where $sim(p, q)$ is the similarity between the tags on p and q, and $w(p, q)$ is the weight assigned to q based on its timestamp relative to p
if the similarity is greater than the threshold s **then**
Add a directed edge from the node representing p to the node representing q in T_p
end if
if q has already been added to a tree in T **then**
Merge T_p with the diffusion tree containing q
end if
end for
Add tree T_p to the set of diffusion trees T
end for
return trees T
end function

The algorithm performs an enhanced diffusion tree analysis on social media postings by including a time series analysis. The algorithm is fed a collection of social media posts P, each with a timestamp, a similarity threshold s, and a weight function w. The approach generates a collection of diffusion trees T, each representing a simplex social network. The algorithm first sorts the postings in P by their timestamp. It then generates an empty tree collection, T. For each post p in P, the approach generates a new tree T_p with a single node. The algorithm computes the similarity between p and q as $w(p, q)timessim(p, q), where$sim(p,q) represents the similarity between the tags on p and q, and $w(p, q)$ represents the weight assigned to q based on its timestamp relative to p. If the proximity exceeds the s criterion, the algorithm adds a directed edge in T_p from the p node to the q node. If q has already been added to a tree in T, the procedure merges T_p with the diffusion tree that contains q. The procedure then adds the tree T_p to the collection of diffusion trees T. The method is carried out by the function EmbeddedTimeSeriesAnalysis, which returns a set of diffusion trees T. The program can evaluate social media data to find patterns and trends in distributing information or ideas among users.

We incorporated a weight function to account for each post's date relative to the current post, as well as a similarity criterion to filter out weak correlations in this technique. We made some minor formatting changes to improve the algorithm's readability.

Please refer to the appendix for a sample implementation.

Combined algorithm

Next, a more robust combination of a time series with social media attributes will be discussed.

Require: A collection of social media posts P with a timestamp, a unique identity, and other tags

Ensure: A collection of diffusion trees T

Sort P by timestamp;

$T \leftarrow \emptyset$;

for each post p_i in P **do**

Start a new diffusion tree T_i and add one node for p_i to it;

for each post p_j in P with a timestamp later than p_i **do**

Compute the similarity of the tags on p_i and p_j;

if similarity is above a predetermined threshold **then**

Add a directed edge from the node representing p_i to the node representing p_j in T_i;

if p_j has already been added to a diffusion tree T_k in T **then**
Merge T_i with T_k;
end if
end if
end for
Add the diffusion tree T_i to T;
end for
return T;

This algorithm aims to generate a set of diffusion trees based on social media postings. The algorithm employs social networking concepts to enhance the diffusion tree construction process.

The input to the algorithm is a collection of social media postings P with a timestamp, a unique identity, and other tags. The approach generates a set of diffusion trees T.

The algorithm first sorts the collection of posts P by timestamp to ensure that posts are treated chronologically. Then it generates an empty set T to hold the resulting diffusion trees.

The script then runs over each p_i post in P. Each post p_i generates a new diffusion tree T_i and adds a node representing p_i.

For each post p_j with a date later than p_i, the algorithm computes the similarity between the tags on p_i and p_j using a weight function w. If the similarity reaches a specific threshold, the algorithm inserts a directed edge in T_i between the nodes representing p_i and p_j. If a diffusion tree T_k from T has already been added to p_j, the procedure combines T_i and T_k.

The set of diffusion trees T, the approach includes the diffusion tree T_i. Once all of the posts have been processed,

The algorithm creates a series of diffusion trees from social media posts, each reflecting the spread of a particular subject or concept over time. Considering the links between postings and the weight of those ties over time, including social networking concepts, increases the diffusion trees' accuracy.

This solution includes a phase that computes tag similarity for each pair of posts and a conditional statement that merges diffusion trees if two posts are similar and already in the same diffusion tree. We made some minor formatting changes to improve the algorithm's readability.

Using the diffusion trees T, identify the social network's prominent users. A user who has started or forwarded several diffusion trees and whose postings have received

a lot of exposure is considered an important person.

Network centrality measures, including degree, betweenness, and eigenvector centrality, should be calculated for the prominent users. These metrics can provide light on the social network's structural characteristics as well as the relative significance of the important people.

Examine the propagation patterns, diffusion rates, and information cascades in the diffusion trees T to analyze the temporal dynamics of the diffusion process. The major elements that affect whether information dispersion in the social network is successful or unsuccessful may be determined using this study.

This research proposes a new algorithm that links a new social peer in MSN to make hate speech more neural. The key element is arguing that a peer has a strong enough impact to stop hate speech. More important than ever, connecting to a new cluster that supports the removal of hate speech is a fantastic social innovation platform for any given social media platform.

The novel algorithm considers essential elements like the Multi-view SVM meta-classifier for categorizing a particular node's hatred level in the social graph. The graph connectivity between the social nodes is then calculated using closeness centrality and betweenness centrality. These elements are necessary when recommending a new peer since "Trust" is very important. The influencing element must determine the ego strength to prevent hate speech by merging it with another person once the excellent connection has been calculated. Use "Multiple criteria decision making," a powerful tool once the situation is appropriate for a solution, to complete this assignment. As a result, the influence factor aids in eliminating hate speech on MSN. Lastly, collaborative matrix factorization filtering maintains peer recommendations. The characteristics listed above.

In order to fine-gain and optimize the unique method, connection measurements, Bias parameters, and Regularization tests were performed on it. Lastly, the outcomes demonstrate that the unique algorithm considers hating speech removers while linking. There is no proof that the same algorithm would be used to connect and distribute hate speech since hate speech and user interactions pair. The authors are willing to create a unique algorithm based on natural language processing to self-connect peer recommendation networks that can eliminate hate material from social media networks as an extension of this experiment.

4.3 Derivation of an equation

4.3.1 Analyze the algorithm behavior

Architectural solution for prescriptive analytics with the evolutionary aspects of the proposed algorithm.

Upon the above algorithms, the following mathematical equation can derive

$$y(t) = \frac{k}{1 + Ae^{-r(t-t_0)}} \tag{4.1}$$

where:

$y(t)$ is the number of people affected by the information at time t K is the maximum number of people who will be affected by the information A is the initial number of people who were affected by the information at time t_0 r is the growth rate of the information diffusion t_0 is the initial time when the information was first introduced.

Calculating the maximum number of people affected.

Here is the equation that relates the maximum number of people affected (K) and the growth rate of the diffusion (r) to the parameters of the model:

$$K = \frac{k}{1 - A}$$

Where k is the maximum potential size of the population being exposed to the information, and A is a parameter related to the initial conditions of the diffusion process.

To express r in terms of the parameters of the model, we can take the natural logarithm of both sides of the original equation:

$$\ln y(t) = \ln k - \ln \overset{t}{1 + Ae^{-r(t-t_0)}}) \tag{4.2}$$

Lemma

The derivative of both sides (considering to time t, we get:

$$\frac{d \ln y(t)}{dt} = -\frac{Ae^{-r(t-t_0)}}{1 + Ae^{-r(t-t_0)}} \tag{4.3}$$

Applying the chain rule to the equation

$$\frac{d \ln y(t)}{dt} as \frac{dy(t)/dt}{y(t)} \tag{4.4}$$

$$\frac{dy(t)}{dt} = y(t) \cdot \frac{Ae^{-r(t-t_0)}}{1 + Ae^{-r(t-t_0)}} = ry(t) \quad 1 - \frac{y(t)}{K} \tag{4.5}$$

Where we have used the expression for K that we derived earlier, this final equation relates the growth rate of the diffusion r to the maximum number of people affected K and the dynamics of the diffusion process.

4.4 Software and tools for the research

For analyzing social media information dispersion, numerous methods and methodologies are available. NodeXL[8] is a popular free and open-source platform for gathering and analyzing social media data [163]. Another popular tool is Gephi[9], a free, open-source application that contains several network research and visualization features [164].

Other social media information diffusion analysis software tools include Radian6[10], Netlytic [11], and Brandwatch[12].

Radian6 is a high-end tool for tracking and analyzing social media data across several platforms. Netlytic is a web-based tool that uses natural language processing algorithms to analyze social media data. Brandwatch is a paid service that provides social media monitoring and analytics.

It is critical to note that each program and tool has advantages and disadvantages and that the software selected is influenced by the study topic and data sources. Additionally, the ethical implications of using these data collection and processing tools must be examined [165].

As mentioned above, there are certain limitations in each product. Hence, I have decided to develop the prototype.

Python is a popular choice for social media research because of its simplicity of use, readability, and the availability of various libraries, such as Pandas[13], NumPy[14], and NetworkX[15]. R is also widely used in social media analysis, with packages such as igraph [16] and statnet[17] providing useful network analysis features. Java and C++ are

8https://nodexl.com/
9https://gephi.org/
10https://www.socialstatus.io/radian6-and-social-status/
11https://netlytic.org/
12https://www.brandwatch.com/
13https://pandas.pydata.org/
14https://numpy.org/
15https://networkx.org/
16https://igraph.org/
17https://statnet.org/packages/

preferred for larger-scale studies and building high-performance algorithms for complex data structures.

Popular packages involved with the research project

- NetworkX is a Python package for creating, manipulating, and investigating complex networks' structure, dynamics, and functions.
- PySNA[18] is a Python toolkit for social network analysis.
- Scikit-learn[19] is a Python machine-learning package with clustering and classification capabilities.
- PySpark[20] is a Python API for Apache Spark, a high-performance platform for massive data processing.But, I haven't used it during the research.
- tweepy[21] is a Python package for accessing the Twitter API and gathering Twitter data.
- Facebook-SDK[22] is a Python package that allows you to connect to the Facebook API and get data from Facebook.

4.5 Validity and reliability

Social media information diffusion analysis demands trustworthy and valid techniques for accurate results. Triangulation is a well-known strategy for strengthening these systems' validity and reliability. Triangulation is the act of combining many sources or approaches to provide a more accurate and comprehensive understanding of a phenomenon. In the context of social media information dispersion study, triangulation may imply combining many data sources, such as social network analysis, text mining, and machine learning approaches, to provide a more complete view of the dissemination process.

Some studies have employed triangulation to improve the validity and reliability of social media information dispersion analysis methodologies. For example, Kumara [166] used text mining, network analysis, and machine learning approaches to construct

18https://pypi.org/project/pysna/
19https://scikit-learn.org/stable/index.html
20https://spark.apache.org/docs/latest/api/python/index.html
21https://www.tweepy.org/
22https://developers.facebook.com/docs/graph-api/guides/our-sdks

a comprehensive strategy for identifying influential people and anticipating information dissemination on Twitter. In another study, Li [167] used a combination of network analysis, content analysis, and machine learning algorithms to identify the factors influencing information transmission on Sina Weibo.

Nonetheless, it is essential to note that the validity and reliability of the algorithms used in social media information dispersion studies are still restricted. These limitations can be ascribed to inadequate or biased data collection, algorithm design or implementation flaws, and the complexity and unpredictability of human behavior on social media platforms. As a result, while performing and analyzing their investigations, researchers must recognize and overcome these limits.

Lastly, triangulation in social media information diffusion studies can improve algorithm validity and reliability. However, constraints must be addressed to provide an accurate and comprehensive study.

Member checking

Member checking is a qualitative research method used to increase the reliability and validity of findings. It consists of presenting the findings or interpretations to all or a subset of the participants and asking for feedback, recommendations, and corrections. Member checking allows academics to confirm the validity and trustworthiness of their analyses and interpretations by comparing them to the participants' perspectives and experiences.

However, regarding social media information dispersion research, member checking may not always be feasible or appropriate. This is due to the large volume of data and the anonymity of most social media users, which makes it challenging to identify and contact specific individuals for feedback. Further validation techniques, such as inter-coder reliability, expert review, or comparison with other data sources, may be necessary in such cases. These tactics might help to validate and rely on the algorithm and analytic outcomes.

Furthermore, member checking can help to address the ethical considerations that are inherent in social media research, particularly in terms of respecting participants and their viewpoints [168]. By including participants or stakeholders in the study process, member checking can promote transparency and open communication, which are key principles in ethical research methods.

inter-coder reliability

One way to evaluate an algorithm's validity and reliability is "inter-coder reliability," which involves multiple coders independently coding a section of the data and then comparing their findings for consistency.

Inter-coder dependability has been demonstrated to improve the validity and reliability of algorithms employed in social media information dissemination analysis [169, 170]. Using several coders reduces the possibility of mistakes and biases in the coding process, resulting in a more accurate and trustworthy data analysis.

Additionally, using inter-coder reliability allows for discovering any discrepancies or inconsistencies in the data, which can then be addressed and corrected to improve the overall quality of the research.

As a result, employing inter-coder dependability is an essential step in determining the validity and reliability of an algorithm for a social media information dispersion study.

4.6 Research ethics

One of the most critical ethical challenges in social media information dissemination analytics is obtaining informed permission from participants. Researchers must inform participants about the nature of the study, the aim of their involvement, and the potential risks and benefits. This information should be provided in straightforward language, and participants should be allowed to leave the study anytime. Failure to get informed consent may have negative consequences for participants and jeopardize the validity of the research findings [171].

While analysing data in social media information distribution analytics, ethical considerations should be taken into account. Researchers should take care not to expose their participants' identities or include any identifying information in their study findings. Additionally, researchers must follow data protection rules and protocols while analysing and reporting data [172].

Lastly, research ethics is a critical component of social media information distribution analytics. Researchers must get informed consent from individuals, respect their privacy, avoid deceiving them, and adhere to ethical standards while analysing data. These ethical issues are necessary to protect both the participants and the conclusions of the study.

4.7 Limitations

4.7.1 limitations in social media data extraction.

Social media data extraction involves collecting and analyzing data from social media platforms. While it provides helpful insights into user behavior and communication patterns for academics, it has severe downsides.

One problem is the need for more data management. Social media firms hold the data and can block access or change their terms of service at any time. This can make repeating research and creating reliable databases difficult.

Another disadvantage is the accuracy of the data. Users on social media may only sometimes provide correct information or may misrepresent themselves, making reliable inferences from data difficult. Moreover, the methods used to extract data may only sometimes capture all relevant information, resulting in biased or incomplete datasets.

Privacy concerns also limit social media data mining. Researchers must be cautious not to infringe on user privacy rights or reveal sensitive information. Dealing with data from public profiles can be incredibly challenging since people may not expect their information to be used for research purposes.

Lastly, the sheer volume of available social media data might be a restriction. With millions of users and billions of interactions occurring every day, identifying and extracting valuable data may be challenging. As a result, analyzing massive datasets can be time-consuming and resource intensive. Thus, while social media data extraction is a crucial research tool, researchers must be aware of its limits and take precautions to ensure the quality, privacy, and relevance of the data they gather.

Analytical issues

The social media analysis sample may be representative of only some of the population. If the study is based on Twitter data, for example, it may not completely reflect the beliefs and actions of non-Twitter users.

Contextual knowledge: Knowing the context in which the data was produced may make it possible to analyze social media data. When the context is neglected, words and images can be easily misinterpreted. Data dependability and validity: Data dependability and validity might be issues in social media analysis. Data collection, coding, and interpretation mistakes might result in inaccurate results.

Access to social media data may be restricted due to privacy concerns or data-

sharing rules. This has the potential to limit the analysis's scope and depth.

Limited analytical approaches: Because social media data can be complex, specialized analytical tools are needed to understand it. However, many researchers may need more knowledge or resources to carry out these processes, which limits the scope of the research.

Chapter 5

Evaluation

5.1 Data Preparation

5.2 Exploratory data analytics

In this section, the essential methods applied in the research are discussed. The initial data set consisted of the following features.

- Channel features
  - Subscriber's count
  - Number of currently published videos
  - View a count of the last three videos published by the channel
- Video features
  - Length
  - Published day
  - Is holiday
  - Is weekend
  - Published time
  - Topic
  - Number of words
  - Language
  - Special characters in the topic

	Column	Non-Null Count	Dtype
0	video*id*	175191 non-null	object
1	title	175191 non-null	object
2	publishedAt	175191 non-null	object
3	channelId	175191 non-null	object
4	channelTitle	175191 non-null	object
5	categoryId	175191 non-null	int64
6	trending*date*	175191 non-null	object
7	tags	175191 non-null	object
8	view*count*	175191 non-null	int64
9	likes	175191 non-null	int64
10	dislikes	175191 non-null	int64
11	comment*count*	175191 non-null	int64
12	thumbnail*link*	175191 non-null	object
13	comments*disabled*	175191 non-null	bool
14	ratings*disabled*	175191 non-null	bool
15	description	166669 non-null	object

Table 3: Description of the features

Generally, the entire data set consists of the following data types

- Boolean features = 2
- int64 features = 5
- Object features = 9

Additional information related to Table 3 columns

- videoid - A unique video ID is given to each video
- title - The title given by the content author
- publishedAt - Original publication time
- channelId - Unique channel ID
- channelTitle - Unique channel title
- categoryId - The category of the end users' usage
- trendingdate - When it added to the trending list
- tags - If there are any unique tags, those are mentioned

- viewcount - number of views for the video
- likes - Number of likes for that video
- dislikes - Number of dislikes for that video
- commentcount - Number of comments added for that video
- thumbnaillink - URL for the thumbnail of the video
- commentsdisabled - Is the content author disabled the comments
- ratingsdisabled - Is the content author disabled ratings
- description - What is the description given by the author

List-wise deletion: This method mainly involves removing all cases with missing values.

Mean, mode, or median imputation: All missing values can be filled with a suitable statistic value. In most cases, the means is used to fill the data set. For missing-value imputation, the median and mode can also be found. The benefits of this method include its simplicity and speed and its suitability for small data sets. The disadvantages of this method are as follows: We are not concerned with the correlation between factors. Working only at the column level, mediocre performance for reprised categorical variables, Less accuracy with a high amount of data space

End of distribution imputation: For a broad distribution, when the tail of the distribution is too far, the end of the distributed data can be filled with a N/A. The major disadvantage is that it ”distorts the original distribution.”

Random imputation: Random sampling imputation involves inserting a random value within the range of the given feature. That means the observations of minimum and maximum values within the feature can be extracted. Hence, an arbitrary value can fill the missing data spot. The main advantage is that it is easy to add a value, and the main disadvantage is the high probability of being error-prone. Distribution can be modified accordingly.

Arbitrary value imputation: An ”arbitrary” number imputation involves any amount. The main advantage is that it is effortless to input. The main disadvantage is the high possibility of outliers and High-error-prone,

Outliers Anomaly detection (outlier identification) identifies significantly different data from the standard distribution. Outliers deviate the entire data set from the expected output. Some of the effects of outliers are that the proposed algorithm does not work correctly, adding noise to the data set.

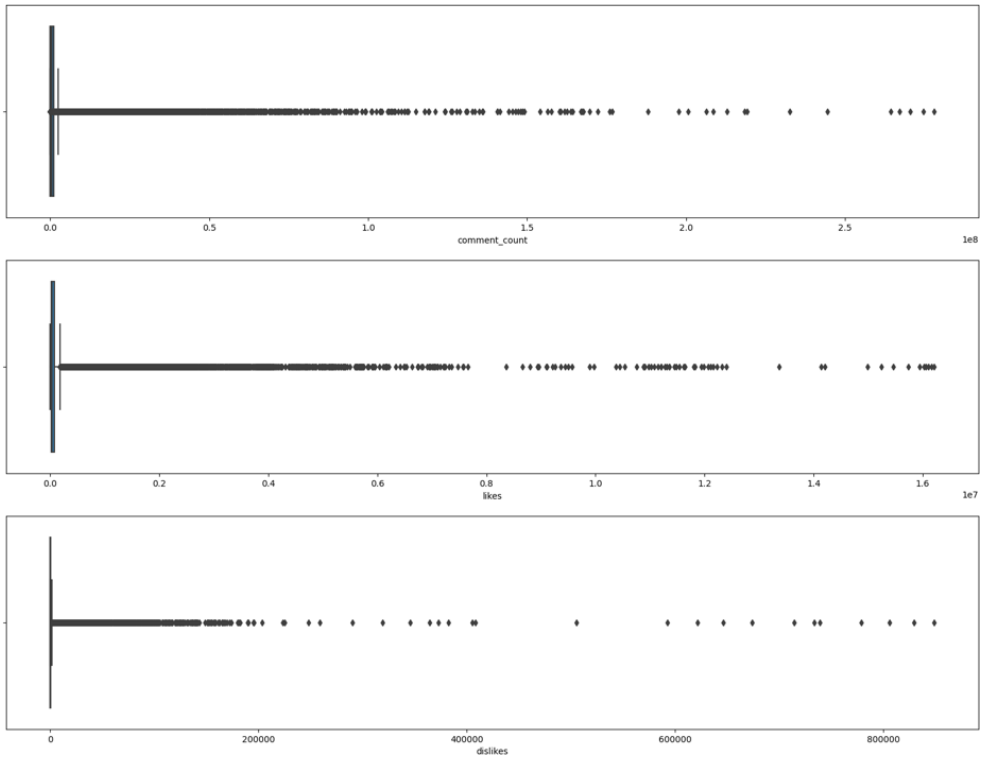


Figure 5.1: Box plot for analyzing the outliers

As shown in figure 5.1, the image depicts outliers in a long-tailed data distribution. Hence, the data set was initially required to remove the outliers for a more accurate result. The base axis for the feature and the Y axis represents the distribution. During this research study, I evaluated the numerical features (i.e., like count, dislike count, and view count).

Log transformation As a variance-stabilizing transformation, this method is employed. Log transformation is used in exploratory data analysis and is often only relevant for ”positive numeric values” data.

The log transformation technique converts a dataset to a logarithmic scale. With this method, a new, converted dataset is constructed by taking the logarithm of each data point. Log transformation can be helpful in various situations, including minimizing the impact of extreme values. Some datasets could have extreme values, which would distort the study.

Implementation

Apply log transformation to the column

df[’logvalues’] = np.log(df[‘feature’])

Label encoding For each job, continuous data may be readily labeled and encoded. The categories must remain sequential to function well in this situation. The labeling

process might then have a set sequence.

In data preparation, the label encoding technique transforms category variables into numerical variables. In label encoding, the dataset's original categorical value for each category is changed to a unique integer value. In some situations, label encoding provides a quick and effective technique to deal with categorical data, but it has several drawbacks. Integer values, for instance, can provide a sense of scale or order among categories where none previously existed. Furthermore, the resulting numerical values may need to be scaled due to their size if there are several distinct categories. Software like Python's scikit-learn module may be used for label encoding, frequently with additional data preparation methods like one-hot or target encoding.

Implementation

Create a sample DataFrame df = pd.DataFrame('Dataframe')

Create a LabelEncoder object and fit it to the 'color' column le = LabelEncoder() le.fit(df['feature'])

Transform the 'color' column using the LabelEncoder df['colorencoded'] = le.transform(df['feature'])

End of distribution imputation

When the tail of the distribution is too far away for broad distribution, the end of the scattered data may be filled with a N/A.

5.3 Feature Engineering

All three social media platforms provide APIs for the identification of features in limited conditions. Well-identified features are the mandatory factor for a successful algorithm. Hence mainly involved with the feature engineering process that requires time-based (Short- and long-term trends).

The number of variables used to represent a dataset determines its dimensionality. Regularization might aid in lowering the danger of overfitting.

Feature Using extraction methods can also result in several other benefits, including:

- Improvements in accuracy.
- Decreased likelihood of training speed overfitting.
- More effective data visualization.
- Improvement in our model's capacity to describe things

By generating new features from the current ones, feature extraction attempts to decrease the number of features in a data set (and then discard the original features of the data set).

Feature selection is a method to minimize the number of features in a data set.

The objective of feature selection instead of feature extraction is to rate the significance of the already present characteristics in the data set and exclude less significant ones (no new features are created).

5.3.1 Feature Extraction

Principle Components Analysis (PCA)

One of the most popular methods for reducing linear dimensions is PCA. We use our original data as input and look for a combination of characteristics that can most effectively summarize the original data distribution to minimize its original dimensions [173]. By focusing on pair-wise distances, PCA may do this by maximizing variances and decreasing the reconstruction error. Each orthogonal axis projected from our original data is ranked in PCA according to its significance. As an unsupervised learning technique, PCA is primarily concerned with variation and needs to be more concerned with the data labels. In some instances, this might result in incorrect data categorization.

Independent Components Analysis (ICA)

Independent Component Analysis is a statistical technique used in signal processing and machine learning to separate independent, non-Gaussian signals from a linear mixture of signals[174]. It works by linearizing non-Gaussian data and finding a set of statistically independent components that can be used to reconstruct the original data. ICA is a linear dimensionality reduction technique that uses a variety of independent components as input data and seeks to accurately identify each one of them (deleting all the unnecessary noise)[175].

ICA has been applied in various fields, such as speech processing, finance, and image analysis. One of its applications in social media is to analyze the diffusion of information in social media networks, where ICA can be used to identify the independent sources of information in the network.

Linear Discriminant Analysis (LDA)

Applying LDA to non-Gaussian data may result in poor classification outcomes since it is anticipated that the input data, when using LDA, follows a Gaussian distribution (as in this example) [176]. In this demonstration, we use LDA to condense our dataset to a single feature, evaluate its precision, and visualize the outcomes.

5.3.2 Locally Linear Embedding (LLE)

LDA operates by finding a linear combination of features that maximizes the separability between different classes[177]. It reduces the number of variables in the data by projecting it onto a new set of orthogonal dimensions to each other and capturing the most significant variation in the data. In other words, LDA is used to find the linear combination of features that separates the classes with maximum separation. LDA works by assuming that the classes have Gaussian distributions with equal covariance matrices and then computing the linear combinations of features that result in maximum separability between the classes. The computation of the LDA model involves the following steps:

- Computation of class means and covariance matrices.
- Computation of the between-class scatter matrix and the within-class scatter matrix.
- The eigenvectors and eigenvalues of the between-class and within-class scatter matrices are computed.
- Selection of the eigenvectors corresponding to the largest eigenvalues to form a linear discriminant function.

In multi-class scenarios where classic linear classifiers may need to perform better, LDA is a powerful data visualization and classification tool. It has been extensively employed in many areas, including bioinformatics, speech recognition, and image processing.

Projecting the data in a lower-dimensional way while keeping neighborhood-specific distances is the goal of locally linear embedding (LLE). It is possible to compare it globally as a collection of local Principal Component Analyses to identify the best non-linear embedding.

It is now time to look at how to handle non-linear scenarios as we have already looked at approaches like PCA and LDA that work best in linear connections between the different features.

A dimensionality reduction technique based on manifold learning is called locally linear embedding. A D-dimensional object that is entangled in a higher-dimensional setting is called a manifold. To avoid displaying this object in an unnecessarily wider space, Manifold Learning aims to make it representable in its original D dimensions.

A typical example of manifold learning in machine learning is the Swiss Roll Manifold. We can unroll the data and reduce it to a two-dimensional space when given data in three dimensions with a distribution that matches a roll. Manifold learning techniques include Isomap, Locally Linear Embedding, Improved Locally Linear Embedding, and Hessian Eigen mapping.

A dimensionality reduction technique based on manifold learning is called locally linear embedding. A D-dimensional object that is entangled in a higher-dimensional setting is called a manifold. To avoid displaying this object in an unnecessarily wider space, Manifold Learning aims to make it representable in its original dimensions.

A typical example of manifold learning in machine learning is the Swiss Roll Manifold. We can unroll the data and reduce it to a two-dimensional space when given data in three dimensions with a distribution that matches a roll. Manifold learning techniques include Isomap, Locally Linear Embedding, Improved Locally Linear Embedding, and Hessian Eigen mapping.

5.3.3 Feature Preparation

Identifying relevant data for the machine learning problem is known as ”feature identification.” In this case, some unwanted features are removed, editing features are added to make that data more valuable, and new features might be required. Hence, all data collection is used for the complete feature identification process.

Recursive feature elimination (RFE)

One of the most often used methods for finding the best features through feature reduction. The main emphasis was placed on feature significance, which may explain the effect of a social media post on trends. Additional factors were taken into account while choosing the best algorithm. On each cycle of the feature removal, each component was taken into consideration to delete a characteristic.

Autoencoders

Using the family of machine learning techniques known as autoencoders is one way to reduce dimensionality. The main contrast between autoencoders and other dimensionality reduction techniques is that autoencoders use non-linear transformations to project data from a high to a lower dimension. Denoising autoencoders, variant autoencoders, convolutional autoencoders, and sparse autoencoders are only a few of the several types of autoencoders.In this experiment, we'll build a basic Autoencoder first. The two main components of an autoencoder's fundamental architecture are as follows:

- Encoder: This step compresses the input data to remove any noise or extraneous information. The output of the encoder step is frequently described as latent space or a bottleneck.
- Decoder: attempts to reproduce the original Autoencoder input using just the compressed form of the encoded latent space as input (the encoded latent space).

If all input qualities are independent, it will be extremely difficult for the Autoencoder to encode and decode data to input into a lower-dimensional space.

Autoencoders may be implemented in Python using the Keras API. In this instance, we indicate in the encoding layer the number of characteristics to be subtracted from our input data (in this case, 3). As seen in the code example below, autoencoders employ X (our input features) as our features and labels (X, Y). For this experiment, I utilized Softmax for the decoding phase and ReLu as the activation function for the encoding phase. If I hadn't utilized non-linear activation functions, the Autoencoder would have attempted to perform a linear modification to reduce the input data (therefore giving us a result similar to if we would have used PCA).

Feature split

The majority of data sets have organized datasets that offer numerous data values. Because of this, the main purpose of feature splitting is to divide the data set in the most effective way. The importance of each feature frequently determines feature splits.

Binning

Fine details are made clearer by binning. Granular data removal primarily aims to classify numerical values into various groups. Continuous variables are split into categorical bins using this technique. Moreover, found outliers have a detrimental

effect on the ML model. Binning can minimize the impact of this undesirable outlier on the model.

Feature engineering in a traditional AI environment

Binning in supervised learning methods

Transform the numerical variables into ”categorical” components, and those components choose discrete points based on the target class information. The biggest disadvantage is the expense of the performance, while other positives make the model more reliable and avoid overfitting.

Binning in unsupervised learning methods

Convert the numerical variables into different categorical components; specific elements do not use the target class information. Some techniques involved are ”equal width binning” and ”equal frequency binning.”

Entropy-Based binning

Entropy may be estimated based on class labels and is mostly used in the splitting approach. Finding the entropy based on ”best dividing into bins” is the conventional method. To obtain as much information as you can after the split is the key objective.

A statistical method called entropy-based binning divides related data points into discrete categories, or ”bins.” Moreover, data preprocessing frequently employs entropy-based binning, particularly when analyzing categorical variables. The method is beneficial for lowering the data’s dimensionality, which can enhance the effectiveness of machine learning algorithms. Entropy-based binning aims to reduce the number of necessary bins while increasing the information value of the generated bins. Entropy may be calculated using class labels, and the splitting strategy mainly uses this estimate. The traditional approach finds the entropy based on ”best partitioning into bins.” The primary objective is to collect as much data as possible following the separation.

The entropy metric is used to express how random or unpredictable the data are. When a variable is uniformly distributed, its entropy is maximum; when it is extremely concentrated, it is lowest. Entropy-based binning divides the variable of interest into bins in a way that maximizes the entropy of the resulting groups.

When using the entropy-based binning procedure, the number of bins is often increased iteratively until the entropy is optimum. By grouping comparable data points

into bins, this method reduces data noise and makes it possible to conduct a more insightful analysis.

log transformation

This technique is used as a variance-stabilizing transformation. Applicable only for ”positive numeric values” data, log transformations are used in exploratory data analysis.

Label encoding

To function well in this situation, the categories need to remain in sequential sequence. The labeling process might then have a set sequence. So that binning may be used effectively, labels might have a weighted ordering scheme. For every activity, labeling and encoding continuous data is simple.

One hot encoding

For every activity, labeling and encoding continuous data is simple. The categories need to remain in sequential sequence to function well in this situation. The labeling process might then have a set sequence. So that binning may be used effectively, labels might have a weighted ordering scheme.

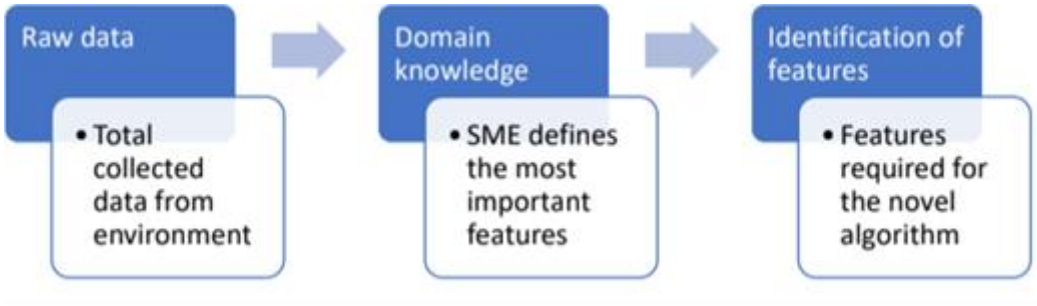
Multicollinearity

Multicollinearity issue: removing specific recently introduced columns. Following a ”one-hot encoding,” freshly added columns could exhibit a strong correlation. It offers the chance to anticipate any of the other recently developed categories. The ability to use the model to anticipate other variables must be preserved when eliminating a column.

Learning with counts

A data set has a wide range of various values or categories. Thus, a ”one-hot encoding” would provide many high-dimensional features. As a result, categorization based on the frequency of each trait is simple. The number of additional columns may then be kept to a minimum. A distinct subgroup is produced for each category at the end of the procedure. Based on their densities, heat maps make it straightforward to see each category.

Figure 5.2: Feature section process



Scaling

There is often a wide distribution in data space. The scaling method takes restrictions within a particular range into account. Data space is restricted to a specific limit after ”scaling”. Thus, it is simple to build various data clusters. Scaling is one of the key operations in KNN or K-Means algorithms because of this.

Moreover, the scaled values are distributed as below.(This is only a selected sample) array([[-0.90068117, 1.03205722, -1.3412724 , -1.31297673], [-1.14301691, -0.1249576 , -1.3412724 , -1.31297673], [-1.38535265, 0.33784833, -1.39813811, -1.31297673], [-1.50652052, 0.10644536, -1.2844067 , -1.31297673], [-1.02184904, 1.26346019, -1.3412724 , -1.31297673]])

The following is a high-level overview of the feature section process used in this study.

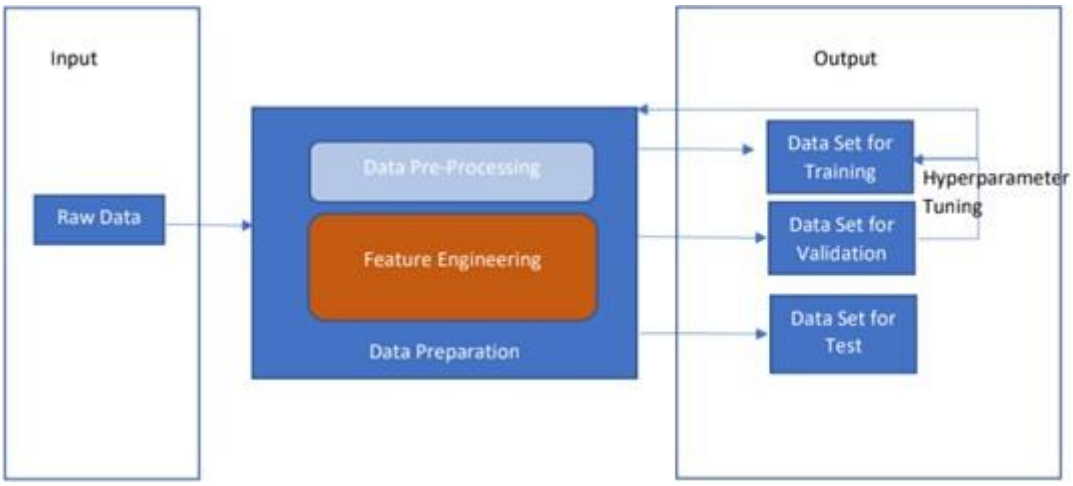
After that, I will talk about the scaling techniques used to create the model. scaling techniques.

- Normalization
- Min-Max Normalization
- Standardization

Normalization

The treatment of skewness and the aggregate of characteristics are the core concerns of normalization. The key aim is to consolidate all scalable characteristics into a single data space range.

Figure 5.3: Overview of feature engineering



Min-Max normalization

The fundamental objective is to centralize the data space neatly.

$$\frac{x - \min(x)}{\max(x) - \min(x)} \tag{5.1}$$

Discussion

In this equation, x is the variable to be normalized, $\min(x)$ is the minimum value of x, and $\max(x)$ is the maximum value of x. The result of this equation is a normalized value between 0 and 1.

Standardization

The data set's standard deviation is mainly considered through standardization, a standard statistical technique known as z-score normalization. In the data domain, standardization lessens the impact of outliers.

The following diagram represents an overview of the feature engineering process.

5.3.4 Issues with the number of features

The model will only work correctly if the data sets get big. The model could be biased if there are few features. Also, the variance could be more significant since there are fewer characteristics. The primary reason is that further details are needed. The amount of features needs to be fixed.

High variance error is getting increasingly dangerous for unobserved data. The model's ability to understand the noise and forecast overfit is the primary cause. On the other hand, bias decreases, and variance increases as the number of characteristics increases. According to the graph, underfitting would be problematic if the model train had fewer features chosen. The model is overfitting if it is greater than the break-even threshold.

5.4 Applying more techniques for fine tuning

5.4.1 Feature crossing

[178]

A method for linearly optimizing classifiers. By mixing connected characteristics, feature crossover develops a new feature factor. It produces a function that has a Boolean result. It puts an end to feature crossing mostly because it gives a thorough insight into how the crossed features interact.

5.4.2 Hashing

[179]

A collection of vectors must be created before hashing can be performed. When features are operating without a defined lexicon, hashing is a fantastic strategy. Nonetheless, it offers the best method for having maximum features. Hashing, thus, makes the best use of memory.

5.4.3 Embedding

[180]

A word's meaning can be expressed by embedding. Embedding is frequently used to compress a vast category feature space. The most effective method for deep learning models is embedding in most circumstances.

5.5 Defining feature matrix

The feature matrix represents the features chosen for the machine learning method. A square bracket denotes these connected traits.

User	*Likes*	*Shares*	*Comments*
UserA	10	5	2
UserB	20	8	4
UserC	5	2	1
UserD	12	6	3
UserE	8	4	2

Features are selected for the ML problem, commonly known as feature vectors. Example:

$$A_1X_1 + A_2X_2 + A_3X_3 = Y$$

where A_1, A_2, and A_3 are the model parameters, X_1, X_2, and X_3 are the feature vectors, and Y is the predicted value.

According to the above feature matrix, X_1, X_2, and X_3 are Likes, Shares, etc.

At the end of the complete process, a clear set of feature vectors is defined that optimally (not more than features or less than features) fit with the novel algorithm.

A feature selection procedure that is automated is provided by deep learning. As a result, it is the most reliable algorithmic solution available. Even with deep learning, feature engineering is occasionally necessary.

The PCA algorithm receives input from the preprocessed data. Data processing is primarily concerned with decreasing computation power.

The LSTM model is then used to apply the data. With social media channels, society may be impacted by both hate speech and other social news. A social media post with a high slope of an upward trend for a debate linked to hate speech may have a big influence on society. This study primarily focuses on the increasing tendency due to social influence. First, the issue may be reduced to a binary categorization: has the economy been growing? Non-upward trending posts were classified as class label 1, and the classes of trending articles were given class label 0.

5.6 Statistical analysis

5.6.1 Univariate analysis

A univariate analysis is performed for a single variable to provide descriptive information regarding a single feature. In other words, a set of data has only one column. Individual column data can be used to get an idea of a singular overview of the whole data set. A bar plot or a box plot is the most popular visualizing method used in

univariate analysis.

5.6.2 Bi-variate analysis

The bivariate analysis compares descriptive statistics for two or more variables. After performing a bivariate analysis, it represents a correlation between two or more variables. A scatter plot and a heat map can visualize the bivariate analysis.

5.7 Statistical forecasting for trend analysis

Statistical forecasting is mainly concerned with identifying patterns in history and using these patterns to predict the future.

- The trend has a seasonality
- Demand patterns
- Stationery and completely random fluctuations around the mean level

In this case, the residuals fluctuate among the available data. Data values are distributed without a proper pattern.

Cycles

This is the combination of seasonality and trend. The trend line has a set of curves that have similar points.

5.7.1 Regression analysis

Regression analysis is a statistical model. It can predict a dependent variable or multiple variables. This is the most widely used in forecasting a trend. Regression analysis uses a dependent variable with an independent variable. An independent variable is commonly known as an explanatory variable. The primary involvement is the identification of casual relationships.

Single linear regression

The dependent variable is forecasted using a single linear regression with only one independent variable.

Multiple linear regression

Predicts a single dependent variable with multiple independent variables. Steps that are followed in regression analysis

- Select one or more x variables to describe y

- Visualize (Plot a scatter plot to the potential relationships between x and y)
- Select a model and estimate (model parameters based on a sample of data)
- Test the significance of the model by R2 and inferences of the slope
- Perform a residual analysis (Validate the model by analyzing the errors of the model)

Experiment

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes
0	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	17158579	787425	43420
1	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	1014651	127794	1688
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146035	5339
3	d380meD0W0M	17.14.11	I Dare You: GOING BALDI?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095828	132239	1989

Figure 5.4: Sample data set

As shown in graph 5.4, the image depicts a sample of the data set. As it represents, the overall data set has 40881 rows × 16 columns for the analysis.
Next starts the analysis of the dataset to fit for the regression.

Plotting for missing values

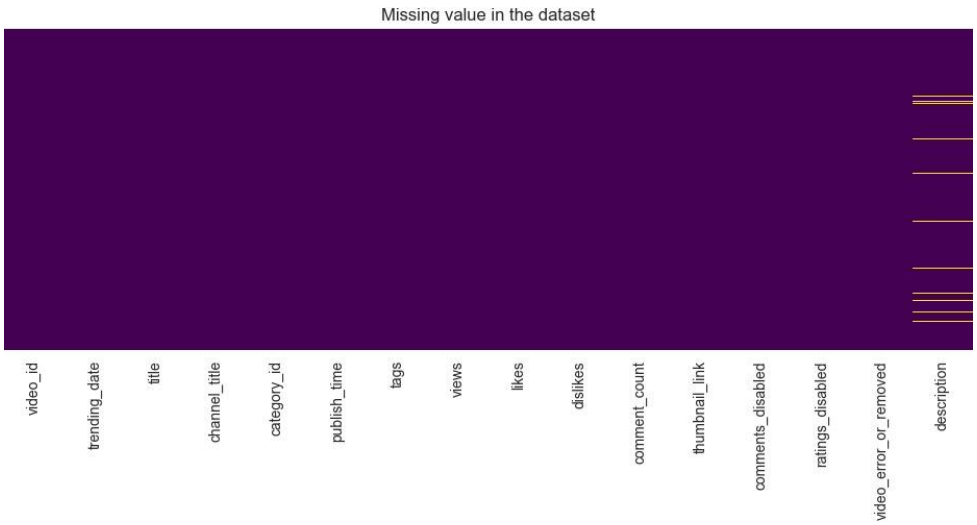


Figure 5.5: Missing value identification

As shown in graph 5.5, the data sets contain missing values in the comments section. Since the regression was focused on numerical data, the missing values do not affect a lot, and I skipped the comments feature till the NLP analysis came into the context.

correlation mapping

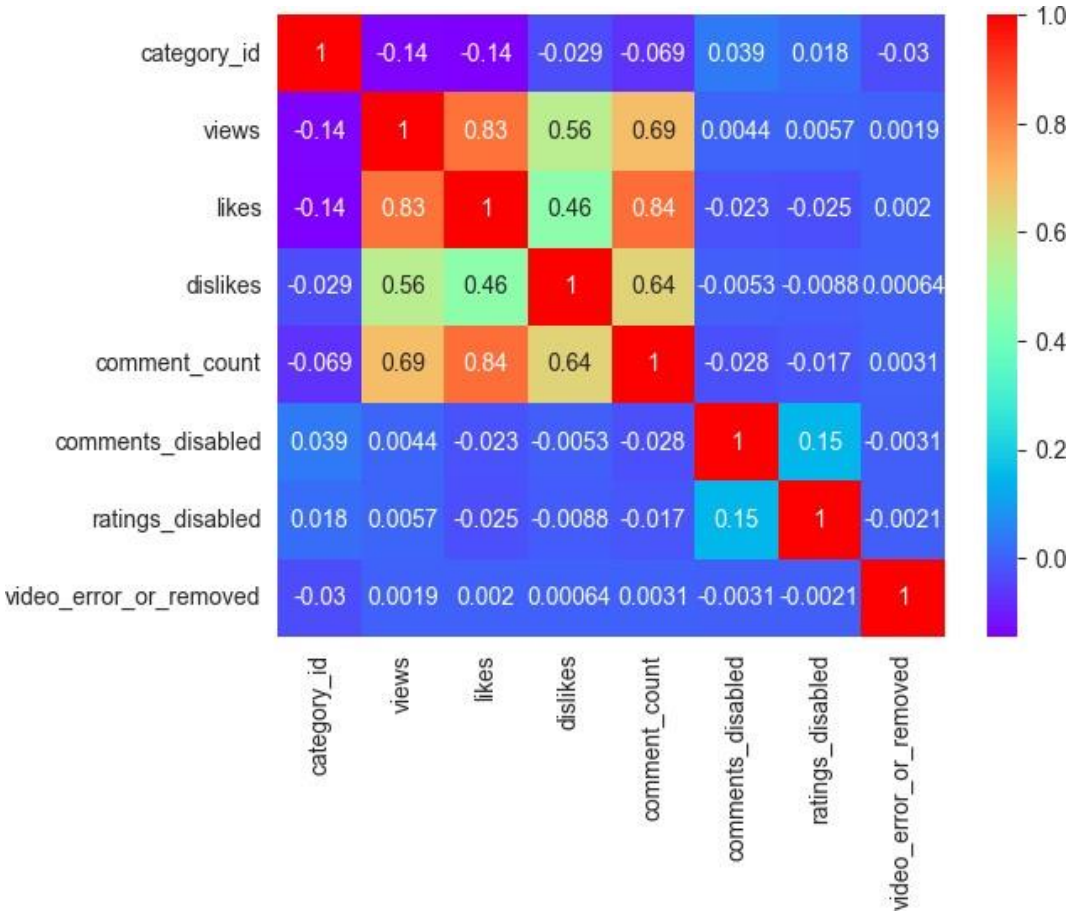


Figure 5.6: Correlations analysis

The correlations are strongly coupled with the feature engineering and the rest of the analysis. It was essential to create a heat map for feature correlation analysis. As shown in Figure 5.6, there is a strong relationship between the view count and the like count. Hence, both features have similar values when developing the algorithm. Hence, I have eliminated likes from the analysis. However, I occasionally kept a ”Likes” count to confirm the algorithm’s abilities.

	category	view	likes	dislikes	cMtCoun	cMtDis	ratDis
category	1	-0.1396	-0.1444	-0.0287	-0.0688	0.0387	0.0183
view	-0.1396	1	0.829	0.5576	0.6931	0.0044	0.0057
likes	-0.1443	0.8289	1	0.4604	0.8366	-0.0231	-0.0247
dislikes	-0.0287	0.5576	0.4604	1	0.6434	-0.0052	-0.0087
cMtCoun	-0.0688	0.6931	0.8366	0.6435	1	-0.0281	-0.0166
cMtDis	0.0387	0.0044	-0.0231	-0.0053	-0.0281	1	0.1479
ratDis	0.0184	0.0057	-0.0247	-0.0088	-0.0166	0.1479	1

Table 4: Correlation of the selected attributes

Theoretical aspect of correlations

During the analysis, the author considered a general correlation analysis to identify relationships. The author wasn't interested in specific correlations like Pearson or Spearman.

The Pearson Correlation measures the strength and direction of linear relationships between two continuous variables assuming normal distributions. Since the data did not perfectly follow a normal distribution, and the author was not interested in linearity for building the algorithm, this method was eliminated.

On the other hand, Spearman Correlation calculates monotonic correlations between data without assuming linearity or normality. It assesses whether variables tend to move together (positive Spearman correlation) or in opposite directions (negative Spearman correlation). However, there was no specific reason for conducting a monotonic correlation analysis, so this method was also eliminated.

Deal with views

The "view" distribution is highly right skewed as shown in graph 5.7. Nearly 95% of the data set consists of clusters; the rest have a low view count. Somehow the tail is considerably lengthy. Hence, when developing the algorithm, I kept an eye on the modifications to the distribution of the views.

cumilative distribution function for the distribution

When the cumulative distribution function is applied to the "views," it generates a perfect graph (5.8) for mapping the distribution. That also provides a perfect hit to develop the algorithm which is based on view count.

Moreover, likes distribution was considered with 50 bins of distribution. Similarly, the correction also generated the same distribution plot for the "like" count. The following plot 5.9 provides a great visualization of the "likes" distribution.

Descriptions of columns for table 4

- category - This is a video category given by the platform. A unique category for

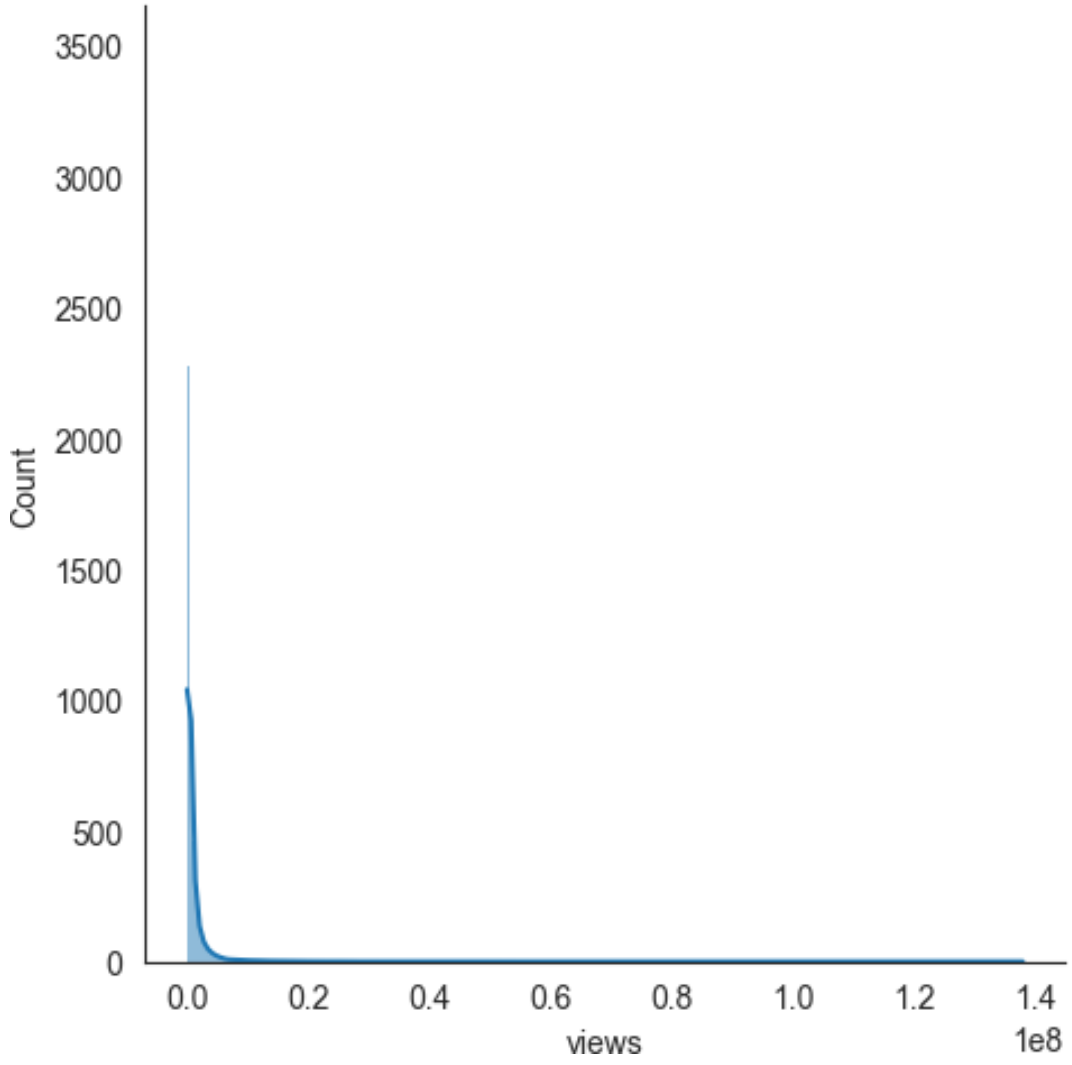


Figure 5.7: Distribution of the number of views

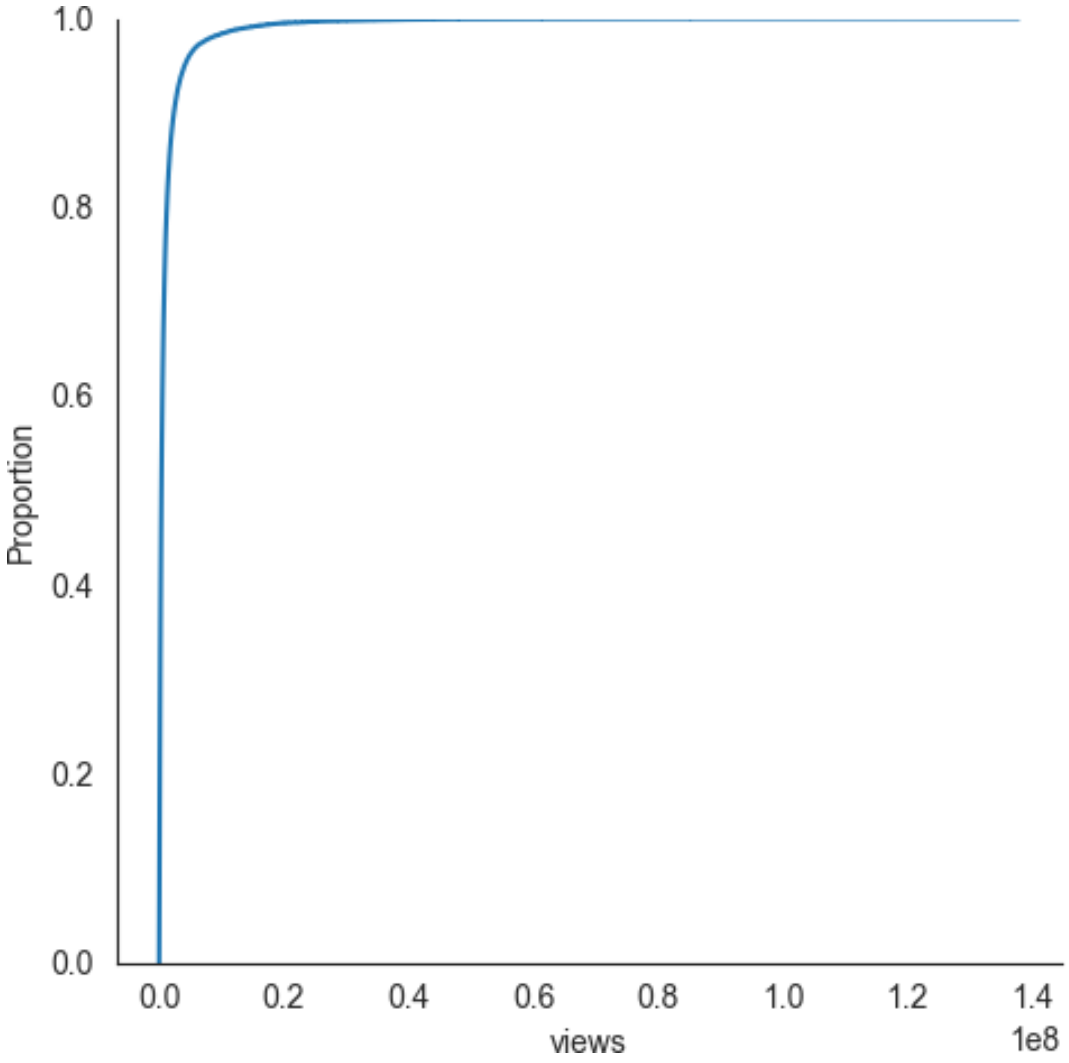


Figure 5.8: Cumilative distribution of number of views

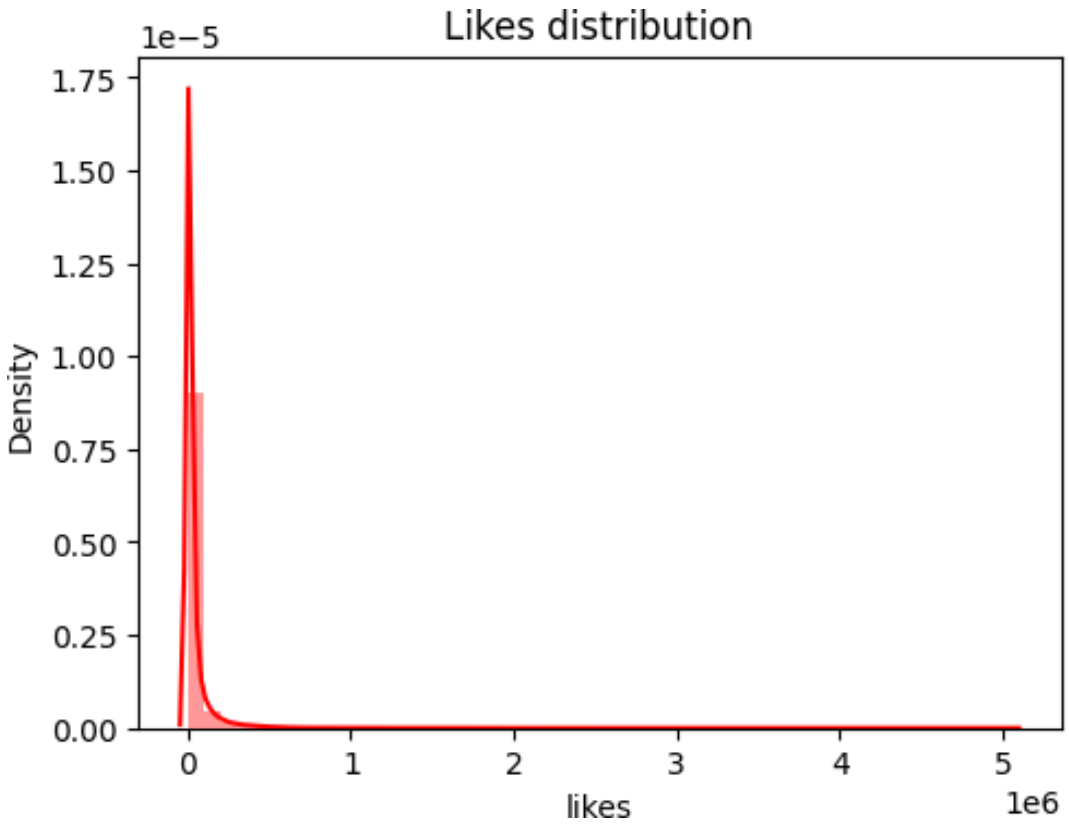


Figure 5.9: Distribution of number of likes

Parameter	Output
coefficient	0.0319
Intercept	3017.0109
R^2	70.66%
Mean Square Error(MSE)	5507.5
Root Mean Square Error(RMSE)	74212.5565
mean squared log error	2.69
Root mean squared log error	1.64

Table 5: Regression Model evaluation

each content.

- view - View count for the given content
- likes - like count for the given content
- dislikes - dislike count for the given content
- cMtCount - number of comments count
- cMtDis - Is that given content rating disabled by the content author
- vidError - Is there any error in the link. (Eg - The original author has removed the content)

As you can see in Table 4, there is a high correlation between the views and likes, which is more than 85% correlated. Hence, the analysis confirms the usability of these feature vector correlation mappings.

Model evaluation

Explanation

The intercept and coefficient of the provided regression model are 3017.0109 and 0.03193, respectively. The model's independent variable(s) account for 70.66% of the variation in the dependent variable, according to the R-squared value of 70.66%. An R-squared result of 70.66% can be regarded as moderately positive. However, this ultimately relies on the analysis's objectives and unique circumstances. It could be enough for some uses but might not be appropriate for others, particularly if more accurate forecasts are required. The MSE value of 5507.5 is comparatively high, which raises the possibility that the model may not nicely fit the data.

Moreover, a related lmplot is given in 5.10. It represents the nature of the regression in views and likes.

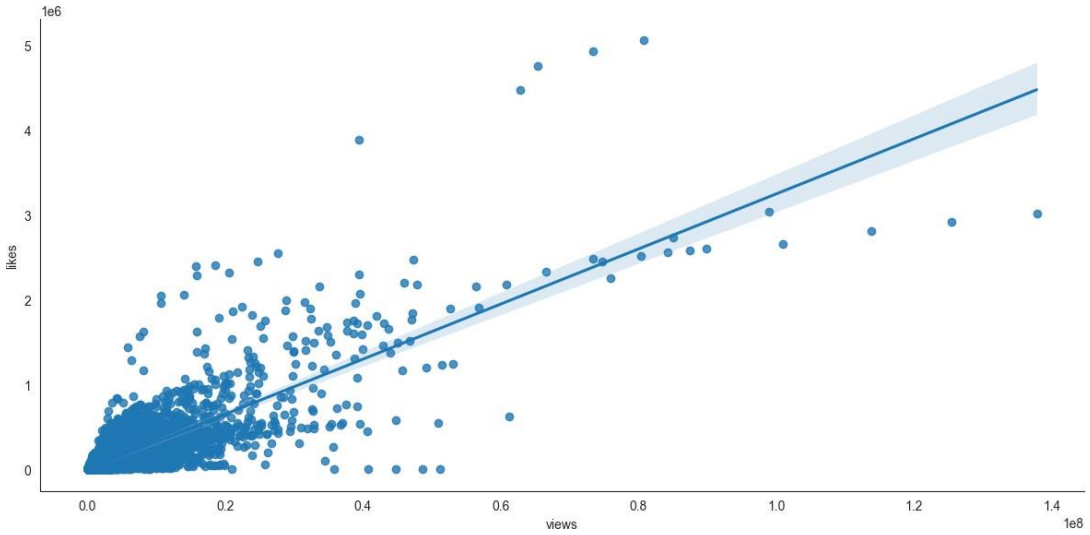


Figure 5.10: Implot of number of likes vs. views

5.7.2 Check for Multicollinearity

I have to figure out if there is any multicollinearity in the given data set. Hence I have employed the ”Variance Inflation Factor” to figure it out. The VIF score of 3.4088 indicates that there is significant collinearity between the independent variables in the model.

FacetGrid for category wise comment analysis

Following 5.11 depicts that only category ten has a proper value distribution of comments.

5.8 Working with Node attributes

Natural Language Processing

Social media platforms are vital in information propagation in the current digital era, and many people interact on social media platforms regularly. The COVID-19 outbreak impacted Sri Lanka and many other countries, resulting in many social and economic changes.

While the country was affected tremendously, social media has stupendously worked as information propagation media. Due to the openness of social media, different topics are being discussed virtually. This research aimed to scope down the comments for a

Figure 5.11: category-wise comment distribution

social media post on popular social media platforms, namely Facebook, Twitter, and YouTube, during the outbreak. Sinhala language and words are written in English, and English text was considered for sentiment analysis. The trend depends on many factors, including the context of the post. Each comment is performed based on the disseminated content performing time series forecasting and focused on collective sentimental trends. Forecast models are based on seasonality, trend, cycle, and related factors. This research used forecast models for seasonality, and the trends were modeled based on exponential smoothing. Time-sensitive trends analysis and forecasting were performed using the Prophet framework. The high velocity of interaction in the first half of the post was identified, and the latter part content consists of the outer cycle of the comment's initiators. Finally, this paper discusses a trend line of hate speech posts on social media platforms for any outbreaks or short period related social media engagements.

Morphology is the term used in linguistics to describe the study of words, their structure, and their links to other words in the same language. It looks at the structure of words and their parts, including stems, root words, prefixes, and suffixes.

Polarity analysis assigns a numerical value to each choice in a text. Instead, polarity analysis evaluates a word, phrase, or passage. The words contained in that text chunk define its polarity value. A text chunk's polarity value is set to 1 if it has a negative connotation. A term like "chuck" has a polarity rating of 1, which is highly favorable. So, polarity is the best method for converting qualitative language into a quantitative value.

Text analysis is one of the significant components of social media content analysis. The remaining sections of the document describe how a firm makes decisions while considering sentiment analysis. One of the popular packages is Textblob[1].

Checking the polarity of a sentence.

To install the required Python packages, run the following commands:

```
pip install textblob
python -m textblob.download_corpora
```

To check the subjectivity of a sentence using TextBlob, you can use the following code:

```
from textblob import TextBlob

text = 'Text for the analysis goes here'
```

[1]https://textblob.readthedocs.io/en/dev/

```
blob_text = TextBlob(text)
sentiment = blob_text.sentiment

print(sentiment)
```

This will output the sentiment polarity and subjectivity, for example:

```
Sentiment(polarity=0.0, subjectivity=0.0)
```

Note that TextBlob may not be able to provide both subjectivity and polarity for certain languages or text types.

```
[('I', 'PRP'), ('had', 'VBD'), ('an', 'DT'), ('awesome', 'JJ'), ('day', 'NN')
```

For the above instance, the system would generate the polarity and subjectivity as below.

```
Sentiment(polarity=0.5, subjectivity=0.7)
```

5.9 Time series analysis

Time series analysis is a statistical technique used to analyze time-dependent data. In recent years, this field has witnessed significant advancements, including developing new algorithms, integrating machine learning techniques, and incorporating time series analysis into broader data analysis tools. Below are some recent implementations in time series analysis.

One common application involves examining historical data to forecast trends over time. We can make predictions based on similar attributes by analyzing these historical patterns. This process, known as forecasting, helps us understand how patterns evolve.

The time series' primary elements include,

- forecasting techniques
- Qualitative approaches

Optimal when there is little information and uncertainty, recurring events, new products, and new technology

5.9.1 Statistical techniques

It is best when historical data is accessible, and things are calm. For instance, predicting the trends for washing machines in the future.

5.9.2 ARIMA model analysis

The following factors were considered for the analysis

Testing Period - 24

Box-Cox lambda transformation parameter - 1

Degree of non-seasonal differencing - 1

Degree of seasonal differencing - 1

Seasonal period - 1

AR(p) order /MA(q) order, SAR(P) order, SMA(Q) order to value 0

Where, Y[t]: The time series' observed value at time t. F[t]: According to the model, the predicted value of the time series at time t. 95% LB: The anticipated value's lower bound is within the 95% confidence range. 95% UB: The anticipated value's upper bound within the 95% confidence interval.

Further, Multivariate ARIMA analysis generates the following results.

Where,

time: The observation's time index.

Y[t]: The time series' value at time t.

Y[t] = F[t], where Y[t] is the time series' observed value at time t and F[t] is the series' predicted value at time t, is the p-value used to test the null hypothesis H0.

P(F[t]¿Y[t-1]): Given the expected values of the time series up to time t-1, the likelihood of witnessing a value of F[t] larger than the observed value Y[t-1] at time t-1. Given the projected values of the time series up to time t-s, P(F[t]¿Y[t-s]) is the likelihood of detecting a value of F[t] larger than the observed value Y[t-s] at time t-s. Given the predicted values of the time series up to time 8, P(F[t]¿Y[8]) is the likelihood of witnessing a value of F[t] larger than the observed value Y[8] at time 8.

Model - 2

5.9.3 Autoregressive Time Series Modelling

After performing autoregressive time series modeling following results were generated.

Mean X(t) - 53.6046512

Variance X(t)- 952.1925365

Table 6: Univariate ARIMA Extrapolation Forecast

Time	$Y[t]$	$F[t]$	95% LB	95% UB
7	89	-	-	-
8	84	-	-	-
9	74	79	-5.181	163.2
10	33	74	-114.2	262.2
11	12	69	-246	384
12	90	64	-397.1	525.1
13	72	59	-565.3	683.3
14	60	54	-749	857
15	84	49	-947	1045
16	51	44	-1158	1246
17	21	39	-1382	1460
18	19	34	-1618	1686
19	29	29	-1865	1923
20	72	24	-2122	2170
21	92	19	-2390	2428
22	92	14	-2668	2696
23	84	9	-2955	2973
24	37	4	-3252	3260
25	27	-1	-3558	3556
26	84	-6	-3872	3860
27	93	-11	-4195	4173
28	14	-16	-4526	4494
29	10	-21	-4865	4823
30	94	-26	-5212	5160
31	25	-31	-5567	5505
32	18	-36	-5929	5857

Table 7: Univariate ARIMA Extrapolation Forecast

time	Y[t]	p-value	P(F[t]>Y[t-1])	P(F[t]>Y[t-s])	P(F[t]>Y[8])
7	89	-	-	-	-
8	84	-	-	-	-
9	74	0.454	0.454	0.454	0.454
10	33	0.335	0.5	0.5	0.459
11	12	0.361	0.589	0.589	0.463
12	90	0.456	0.588	0.588	0.466
13	72	0.484	0.461	0.461	0.469
14	60	0.494	0.483	0.483	0.471
15	84	0.473	0.491	0.491	0.473
16	51	0.495	0.474	0.474	0.474
17	21	0.49	0.493	0.493	0.475
18	19	0.493	0.506	0.506	0.476
19	29	0.5	0.504	0.504	0.477
20	72	0.483	0.498	0.498	0.478
21	92	0.476	0.483	0.483	0.479
22	92	0.477	0.477	0.477	0.48
23	84	0.48	0.478	0.478	0.48
24	37	0.492	0.481	0.481	0.481
25	27	0.494	0.492	0.492	0.481
26	84	0.482	0.493	0.493	0.482
27	93	0.481	0.482	0.482	0.482
28	14	0.495	0.481	0.481	0.483
29	10	0.495	0.494	0.494	0.483
30	94	0.482	0.495	0.495	0.483
31	25	0.492	0.482	0.482	0.484
32	18	0.493	0.492	0.492	0.484

Autocorrelation - 0.2886475

First order serial-correlation- 0.0683844

Second order serial-correlation - -0.3162497

First partial serial-correlation - 0.0683844

Second partial serial-correlation- -0.32243397775915933

The following graph represents the distribution of average values in a given diffusion tree.

As shown in Figure 5.12 and the values received, we can decide the following things based on the information provided in the image. X-axis represents the ”temporal Order” of the series. The x-axis is instrumental in establishing the chronological order of your data, a critical aspect of time series analysis. It enables you to comprehend how past observations are interconnected with future ones.The y-axis represents the

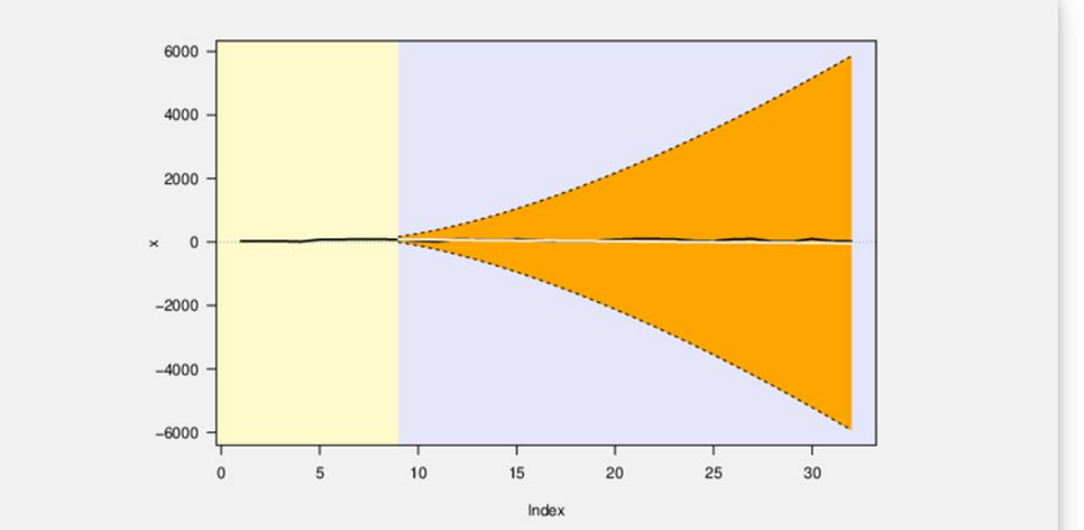


Figure 5.12: ARIMA with Regression

unit of analysis, which may or may not be a physical unit. In this case, it specifically represents the total user involvement.

The median value of $X(t)$ is 53.6, and this number may be used as a benchmark against which to measure subsequent data. The fact that $X(t)$ has a variance of 952.2 indicates that the data are quite variable. According to the autocorrelation value of 0.29, there may be a slight positive correlation between observations made at various times. There is just a slight positive correlation between neighboring data, according to the first-order serial correlation value of 0.07. The score of -0.32 for the second-order serial correlation implies a somewhat negative correlation between observations made at different times. The first partial serial correlation value equals the first-order serial correlation value, showing that the intervening observations have little to no impact on the correlation between neighboring observations. These results can be a reference for additional time series data analysis and modeling.

5.9.4 Techniques of Quantitative Forecasting

- Deep Learning Techniques: Deep learning techniques, such as Long Short-Term Memory (LSTM) networks, have been applied to time series analysis and have shown promising results in predicting time series data.
- Ensemble Methods: Ensemble methods, such as Random Forests and Gradient Boosting, have been used to improve the accuracy of time series forecasting.
- Integrating Time Series Analysis with Text Analytics: Text analytics has been

integrated with time series analysis to analyze unstructured data, such as social media posts, and extract insights into time series data.

- Using Transfer Learning for Time Series Analysis: Transfer learning, a machine learning technique, has been applied to time series analysis to improve the accuracy of time series forecasting by leveraging information from related time series data.

These are just a few examples of recent implementations in time series analysis. The field is constantly evolving and new techniques are being developed and tested regularly.

There have been numerous studies on the information diffusion of social media posts, and several factors have been identified as contributing to the spread of information on these platforms. These factors include:

- User characteristics: User characteristics, such as the number of followers, the level of engagement, and the level of influence, play a role in the spread of information on social media.
- Content characteristics: The content of the social media post, including the type of information, the level of novelty, and the level of emotional impact, also affects the spread of information.
- Social network structure: The structure of the social network, including the number of connections and the level of homophily, affects the spread of information.
- Timing: The timing of the social media post, including the day of the week and the time of day, affects the spread of information.
- Platform characteristics: The platform characteristics, such as the design of the interface and the algorithms used to recommend content, also affect the spread of information.

5.9.5 Selecting time series packages

An open-source library for categorizing time series is called Pyts[2].

This adaptable toolkit includes several methods documented in the literature, preprocessing options, and a tool for importing data sets.

Pyts uses the usual scientific Python libraries numpy, scipy, scikit-learn, joblib, and numba, all of which are BSD-3-Clause licensed. An example of how information is

[2]https://github.com/johannfaouzi/pyts

exchanged. Greater distance means there are no content problems. But, when there is less room between nodes, the speed must be decreased. There will be a scale from 0 to 1 used for each time period.

Generally, information diffusion on social media is influenced by various factors, including user characteristics, content characteristics, social network structure, timing, and platform characteristics. Understanding these factors can help researchers and practitioners design and implement effective information diffusion strategies on social media platforms.

Data retrieved using a particular technique (as discussed early using APIs) and with a uniform distribution of time intervals is known as a time series. These numbers are dynamic. As a result, it might alter over time. Cross-sectional data only evaluates behavior based on a single snapshot of the data and only works with a population sample.

5.9.6 Detecting missing values

The initial step conducted by the authors is finding any missing values.While collecting data, there might be incomplete records in the data set. These records need to have exclusive features for the algorithm. It results in an error-prone outcome from the novel algorithm.

The authors collected 8160 trending and non-trending videos for classification, with 4211 being None Trending Videos and 3949 being Trending Videos. All categorical variables were ”one-hot” encoded, and the data set was then scaled using a standard scaler.

5.9.7 Evaluating time series data

Evaluating forecasting

- Residual Analysis
- Principal of Parsimony
- Simplest model types
  - least-squares linear
  - least-square quadratic
  - 1st order autoregressive

- 2nd and 3rd order autoregressive
- least-squares exponential

5.10 Centrality measures in the Social network

The centrality between each node in the network graph created above was measured to look at the information dispersion in the network. For that diffusion, data was input into a data frame using a list comprehension.

Following figure 5.13 represents a clear analysis for likes vs retweets in a long term distribution.

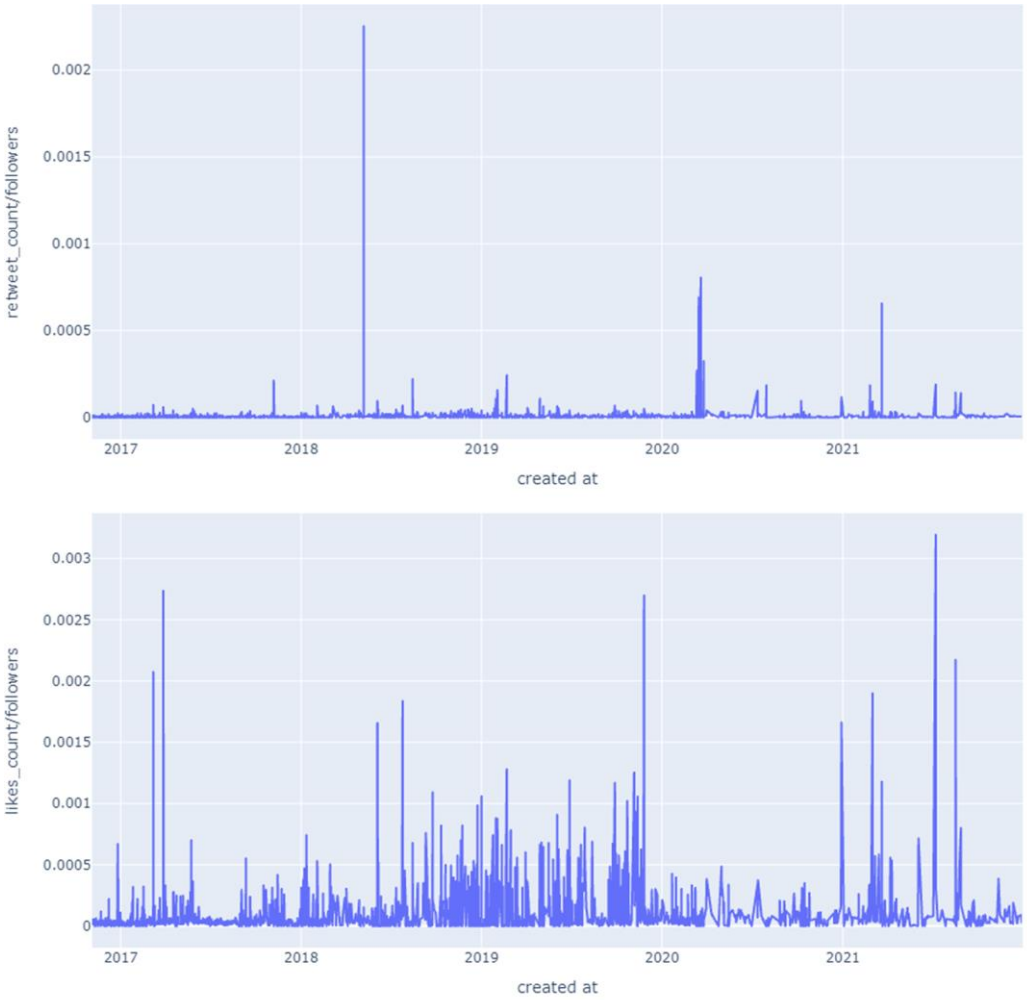


Figure 5.13: Retweets vs likes in long term diffusion

This study conducted a network analysis on Twitter to identify the influencers.

Following Figure 5.14 consist of the follower's network with Python, the ids were mapped in Microsoft Excel for network analysis with R. The network depicted below has 6625 edges and 7522 nodes. The network graph's longest distance (diameter) is 9. This network graph is imperfect because each vertex is not connected to every other vertex through an edge. The network diagram clearly shows two unconnected components, which are circled. This may be identified as a huge component by studying the degree distribution histogram. Aside from the network dimension, the properties of both components are almost identical, implying that component 02 has graph compactness.

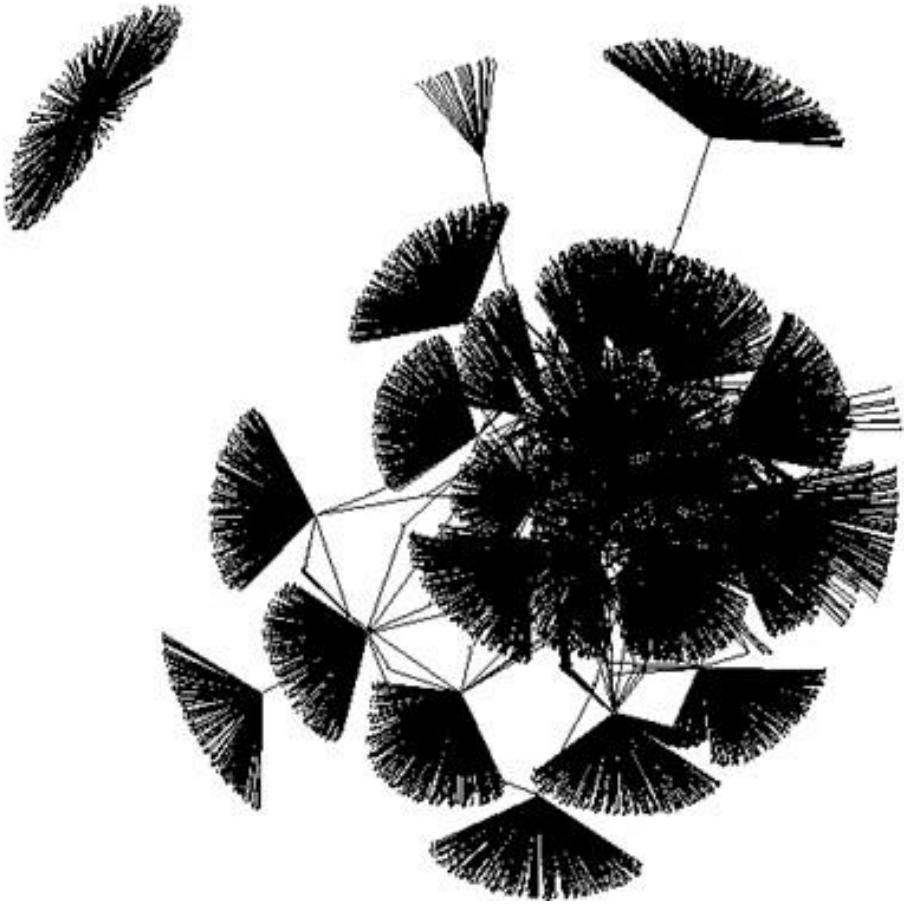


Figure 5.14: The network is illustrated using Twitter followers

The following Table 8 explains that the table compares some elements of two datasets as well as a general comparison. The first has 300 edges and 301 vertices, whereas the second has 7,222 edges and 6,324 vertices. The total number of edges and vertices found in the comparison is 7,522 and 6,625, respectively. The edge density

Attribute	Data Set 1	Data Set 2	Overall
Number of edges	300	7,222	7,522
Number of vertices	301	6,324	6,625
edgedensity	0.0	0.0	0.0
disconnected	TRUE	TRUE	FALSE
Average Degree	2.0	2.3	2.3
transitivity	0.0	0.0	0.0
Average PageRank	0.0	0.0	0.0
diameter	2.0	9.0	9.0
Average Eigenvector	0.1	0.0	0.0
Average Normalized Betweenness	0.0	0.0	0.0

Table 8: Network base statistical overview

for all three datasets is zero, indicating that there are no edges in the dataset. The unconnected property is set to true in datasets 1 and 2, whereas it is set to false in the aggregate dataset.

The diameter of dataset 1 is two, whereas the diameters of dataset 2 and the entire dataset are nine. This indicates that in dataset 2, the greatest distance between any two nodes is greater than the maximum distance in dataset 1. Transitivity is 0.0 for all datasets, indicating no clustering in the networks. The average degree of the three datasets is 2.0 or 2.3, indicating that each node is linked to two other nodes on average. For all datasets, the average eigenvector, normalized betweenness, and PageRank are all 0.0, indicating no critical network nodes.

Overall, the table provides a quick summary of the various properties of the two datasets and may be used to analyze and contrast the aspects of the networks in the datasets.

5.10.1 Edge density distribution

As per the graph 5.15 describes how the edge densities are distributed. The edge density distribution is one when the neighborhood order is one. The X axis represents the post Ids and the Y axis represents edge density.

Explanation of the graph

A network graph with nodes and edges is depicted in the figure. Each node represents a user account, and each edge represents a relationship between two accounts (e.g.,

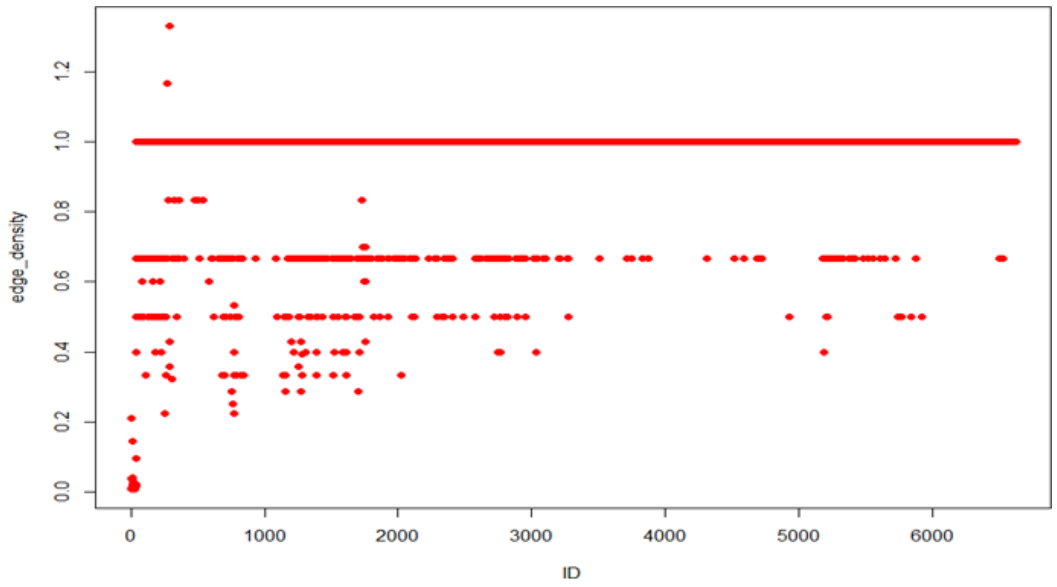


Figure 5.15: The distribution of edge density

following/followers or mentions). The network is depicted using a force-directed layout approach, which seeks to arrange nodes so that linked nodes are closer together than unconnected ones.

The colour of each node represents a community or cluster of accounts with similar connections or interactions. The size of each node corresponds to the number of followers for that account. The thickness of each edge shows the strength of the connection between the two accounts.

The image displays the structure and patterns of interactions in a social network. It may be used to identify account clusters or communities and large accounts with network influence. It can help identify network outliers or anomalies.

The following graph 5.16, represents Edge density distribution when the neighborhood rank is two. Here, the X-axis represents the IDs of the social media posts, and the Y-axis represents the edge density distribution.

Further analysis was carried out based on the clustering coefficient of the network.Here, the X-axis represents the IDs of the social media posts, and the Y-axis represents the edge density distribution.

Centrality measurements reflect a node's influence in the network. It is crucial to identify the most important nodes when establishing a social media strategy. As a result, to identify notable nodes, this study will use centrality measures. Because the eigenvector centrality value is one, the most influential node may be found. Six nodes in the network have eigenvector centrality values equal to or greater than 0.6.

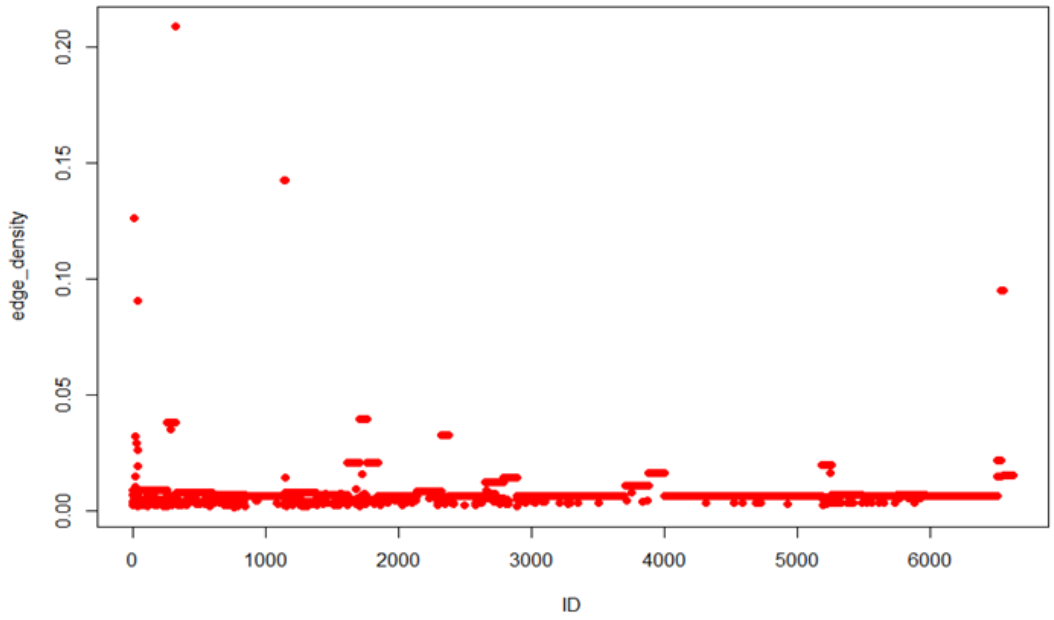


Figure 5.16: The distribution of edge density, where n=2

5.11 Sensitivity analysis

Sensitivity analysis is a method used to determine how different inputs to a model or system will affect the output. It is often used in decision-making and risk management to identify the most important factors that need to be considered when making decisions. The goal of sensitivity analysis is to understand which inputs have the greatest impact on the output, and how changes to those inputs will affect the outcome. Sensitivity analysis can be used in a variety of fields, including finance, engineering, and environmental science. It can be performed using a variety of techniques, such as sensitivity indexes, partial derivatives, and global sensitivity analysis. Sensitivity analysis examines the effects of various independent variable values on a particular dependent variable, given a specific set of underlying assumptions. In other words, sensitivity analysis looks at how different types of uncertainty in a mathematical model affect the model's overall level of uncertainty. This strategy is employed within defined bounds that are dependent on one or more input variables.

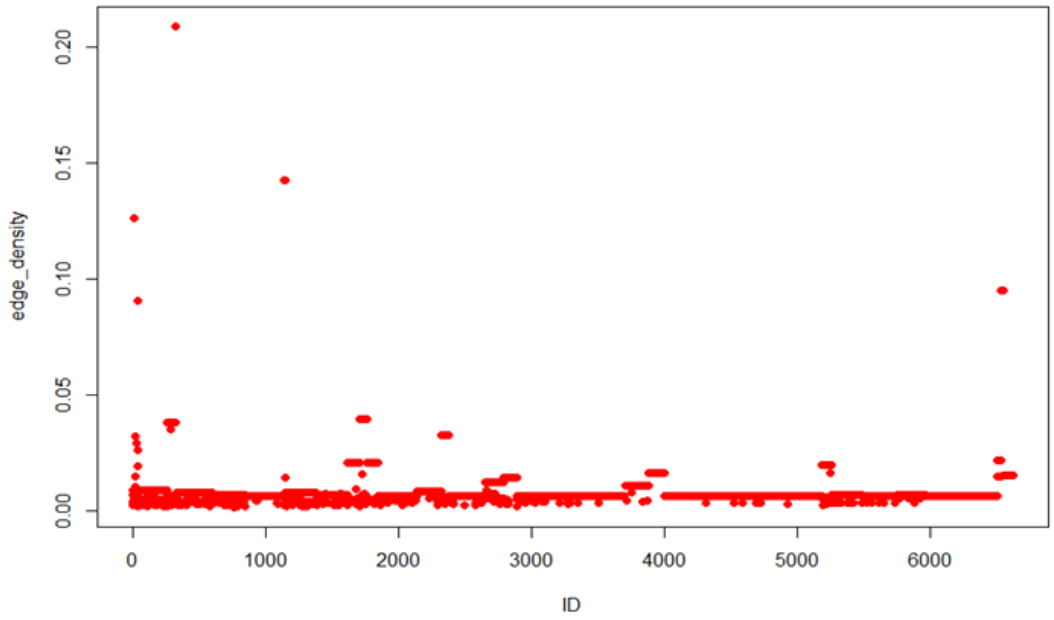


Figure 5.17: The distribution of edge density,where n=2

5.12 contingency table

A contingency table in statistics is a form of matrix-style table that displays variables' (multivariate) frequency distribution. It is also known as a "crosstab" or "cross-tabular" table. They are extensively utilised in business intelligence, engineering, scientific research, and research investigations.

5.13 Profiling the micro services

Profiling microservices entails examining each service's behavior and performance to spot possible problems and areas for improvement. The following measures should be taken while profiling microservices:

Determine which microservices require profiling: Determine the services that are essential to your application and need profiling first. Start with the services that are often utilized or have a history of causing performance problems.

Collect performance data: Obtain performance data for each service, including response time, throughput, error rates, and resource use (CPU, memory, and network usage). You might use technologies like APM (Application Performance Management) systems, log analyzers, and network monitoring tools to get this information.

As per the analysis, I have found Process time for profiling 5.859375 seconds. This is a bit slow profile for service.

Evaluate the data: Look for patterns and trends in the performance indicators by analyzing the data that has been gathered. Search for any irregularities that could point to performance problems, such as sharp increases in reaction times or mistake rates.

The next step is pinpointing the root cause after locating any potential performance issues. This might entail troubleshooting problems, looking at code, and reviewing logs.

Identify opportunities for improvement in the microservices based on the root cause investigation. Coding modifications, configuration adjustments, or resource scaling are necessary.

Monitor and iterate: Monitor the microservices' performance to ensure the improvements are working as intended. Repeat the procedure as necessary to find further opportunities for improvement.

These procedures will help you successfully profile microservices to ensure they operate at peak efficiency and produce the desired outcomes.

5.14 Error handling

5.14.1 Least square method

This method is used to minimize the error of the trend line by minimizing the sum of squared errors. Calculated by subtracting the observed and predicted values. Then it squared off and found the best possible solution.

5.14.2 Model Evaluation

Model validation and improvements Method –

Residual Analysis

Residuals occur due to variations that the model does not explain. There are two unexplained reasons for variation. Pure random noise Pure random noise cannot be eliminated. Hence this is unpredictable. Identification of random noise Upon satisfying the following factors, the model can describe there is random noise.

- Linearity: The residuals in the model are randomly distributed.
- Independence – All residuals are performing an independence

- Normality – All residuals are distributed as gaussian distribution
- Equal variance – variance is constant among residuals

5.14.3 Model errors

Reasons The model needed to include some variables while it was being developed. The model was developed with the wrong methodology. Due to random noise, if any residuals occur, the model can be somewhat valid. However, the model is not valuable if there are any model errors due to missing parameters.

5.14.4 Accuracy of algorithm

An algorithm's accuracy is frequently employed as a performance indicator, particularly for supervised learning tasks like classification and regression.The percentage of cases adequately categorized by the algorithm out of all instances in the dataset is known as accuracy in classification tasks. It is determined by dividing the total number of guesses by the number of correct predictions.

As each algorithm performs differently depending on the volume and complexity of the data, the computing power available, and the specific problem the algorithm is intended to solve, it is challenging to evaluate the effectiveness of AI algorithms.

Performance comparison of AI algorithms

To assess the effectiveness of AI algorithms, it is customary to use metrics like accuracy, precision, recall, F1-score, ROC AUC, time complexity, and memory complexity.

Popular evaluation metrics for unsupervised learning algorithms, such as clustering algorithms, include the Calinski-Harabasz index, Davies-Bouldin index, and silhouette score.

It is critical to take into account the trade-off between performance and other factors, like as complexity and interpretability, while choosing an AI algorithm for a certain task. Testing several algorithms and assessing their effectiveness on a validation set are equally crucial steps in order to select the best algorithm for the task.

Evaluation of the K means the algorithm's performance.

The K-Means algorithm's performance may be measured using a variety of methods, including: This is the "silhouette score": The silhouette score measures how much, in relation to other clusters, each instance resembles its own cluster. On a scale of -1 to 1, a high score indicates good instance separation, whereas a low number indicates a potential error in cluster assignment.

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Algorithm	0.9939	0.7377	0.6716	0.7031
Common Neighbors	0.8829	0.0103	0.1046	0.0188
Jaccard Coefficient	0.9757	0.0119	0.0155	0.0135
Adamic-Adar Index	0.9288	0.0105	0.0605	0.0179
Preferential Attachment	0.0107	0.0107	1	0.0212

Table 9: Performance Metrics for Different Algorithms

Rand Index Reconciled: A score of 1 represents a perfect match, while a score of -1 denotes a completely random match. The ”modified Rand index” is what is used for this. It evaluates the similarity between the actual labels and the anticipated labels.

”inertia” refers to the sum of the squared distances between each instance and its nearest centroid. A lower inertia value indicates better cluster separation. The analysis below is described in an example code.

```
# Fit the K-Means model to the data
kmeans = KMeans(n_clusters=n_clusters)
kmeans.fit(X)

# Predict the cluster labels for each data point
labels = kmeans.labels_

# Compute the adjusted Rand index (ARI)
ari = metrics.adjusted_rand_score(labels, true_labels)
print("Adjusted Rand index:", ari)

# Compute the silhouette score
silhouette_score = metrics.silhouette_score(X, labels, metric='euclidean')
print("Silhouette score:", silhouette_score)
```

Chapter 6

DISCUSSION

6.1 Overview

Social media has made a profound impact on society. While it has benefits, such as connecting people worldwide and promoting social causes, it has also brought new challenges, such as cyberbullying and spreading false information. As a society, we must be aware of these challenges and take appropriate steps to overcome them. These steps include educating individuals about the importance of their digital citizenship and the ways in which they can follow responsible social media useage. By doing so, we can ensure that social media continues to be a positive force in our society.

This research has explored several technical and research methods to enhance the understanding of the research topic: In addition to using the Information Diffusion Analysis Framework, analytical methods, including time series analysis, social network analysis, graph theory, and event-driven architecture, may be applied to learn more about how information spreads across multiplex social media platforms.

A time series analysis is a statistical method that examines data across time to spot patterns and trends. This method may be used to analyze social media data to locate shifts in user behavior, such as activity peaks or the popularity of particular content categories. Time series analysis, for instance, can be used to pinpoint the most talked-about subjects on Twitter at a particular moment, such as during a significant event like the Super Bowl.

In social network analysis, the structure and behavior of social networks are examined to find patterns of influence and interaction. A social network's prominent users, communities, and clusters may be found via social network analysis. Social network analysis, for instance, may be used to locate groups of Instagram users that often engage with one another, such as fashion influencers who frequently work together.

Graph theory, a mathematical body of knowledge, can be used to describe and examine complex networks, including social networks. Social networks may be modeled using graph theory, and interactions and influence patterns can be found. For instance, the relationships between Facebook users may be modeled using graph theory, and people who connect various communities can be found.

Developing software programs that react to events, such as user interactions or system events, is known as "event-driven architecture." Applications that can react to changes in social media activity, such as spikes in user activity or shifts in the popularity of particular types of content, can be created using event-driven architecture. Event-driven architecture, for instance, may be utilized to create a social media monitoring tool that notifies marketers if there is a sudden spike in activity about their brand on Twitter.

During this research, I have focused mainly on the following research Questions.

1. How to investigate a mechanism to identify information diffusion in multiplex social media platforms?

   During this research, I have realized that "graph networking" is an excellent method to identify information diffusion in multiplex social media platforms. This identification involves selecting a group of initial nodes or content items as the starting point for disseminating information. The seeds in this context may pertain to distinct users, prominent accounts, or prevalent content items. Throughout this research, I have used this mechanism to identify information diffusion.

   Moreover,the aim is to create a multiplex network that can accurately

capture the complex and diverse interactions within the social media platform. In order to represent a variety of interaction types or content characteristics, such as retweets, mentions, hashtags, or content similarity, it is recommended to incorporate multiple layers. Diffusion is the term used to describe observing the dissemination of information from its primary sources to the broader network. Observe the transmission of data from the origin to various recipients or content elements, scrutinizing the trends of engagement, circulation, or augmentation it generates.

2. What are the methods to implement a proper mechanism for information diffusion in multiplex social media platforms?

   During this study, I have realized implementation of tree-based graph-structured network is the best method to implement as the mechanism for information diffusion in multiplex social media platforms.

   Adopting a graph-based network architecture utilizing trees is a feasible strategy for integrating the mechanism of information propagation in complex social media environments. The subsequent explanation justifies the perceived benefits.

   The implementation of arboreal networks provides a hierarchical representation of the diffusion process. The implementation of a hierarchical arrangement in the form of a tree structure enables clear visualization and understanding of the distribution of data from a central node, commonly known as the root, to subsequent nodes, also referred to as branches, and subsequently to additional downstream elements, commonly known as leaves. This method of depiction enables the analysis and understanding of the channels through which information is disseminated.

   Their simplicity and efficiency distinguish the utilization of tree structures, as they are relatively straightforward to implement and analyze. Tree-based structures offer enhanced computational efficiency and reduced complexity compared to complex network arrangements such as fully connected graphs or multiplex networks. This particular

attribute makes them suitable for examining the distribution of information across various tiers of social media channels.

The information flow direction in a network based on a tree structure is unequivocal and explicit, guaranteeing unambiguous communication. The nodes of the tree demonstrate a clear parent-child relationship, indicating the transfer of information from a particular node to its subsequent counterpart. This enables a thorough understanding of how data is disseminated and communicated throughout the network.

The employment of tree-based networks enables the facile tracking of diffusion pathways. By scrutinizing the branches and foliage of the tree, researchers can track the exact pathways followed by the information as it spreads throughout the platform. The capacity to track the spread of information provides valuable insights into its dissemination's scope, speed, and direction.

The ability of tree structures to scale allows them to efficiently adapt to the requirements of large and complex social media networks with multiple connections. Tree-based networks can efficiently manage large volumes of data and complex interconnections of multiplex platforms by using appropriate algorithms and data structures.

Proficiency in data interpretation enables the acquisition of valuable insights and the formulation of effective strategies for managing or harnessing the dissemination of information across social media platforms. The hierarchical structure of tree-based networks is the reason behind their interpretability. Academic researchers can identify noteworthy nodes, particularly those with elevated levels of connectivity located closer to the root, and to understand their role in shaping the diffusion of information.

While tree-based graph-structured networks offer various advantages, it is essential to recognize that their suitability may depend on specific research objectives, the characteristics of multiple social media platforms, and the complexity of the information diffusion mechanisms under investigation. A thorough evaluation of variables is crucial

for researchers when selecting an appropriate methodology for their investigation.

3. What are the mechanisms for identifying the evolution of information diffusion in social media platforms?

   Multiple mechanisms exist to identify the evaluation of information diffusion in social media platforms. I have mainly used time-series analysis and graph theory to identify the evolution of information diffusion in social media platforms.

   The amalgamation of time-series analysis and graph theory provides a sturdy approach for identifying the advancement of information propagation in social media networks. The subsequent discourse expounds upon the dynamic relationship between these various methodologies. The study of time-series data involves analyzing patterns and trends over time. This type of analysis is commonly used in various fields, including economics, finance, and engineering, to forecast future outcomes and make informed decisions based on historical data. The time-series analysis process entails examining data points that have been gathered over consecutive time intervals to detect patterns, trends, and alterations that occur over time. Time-series analysis is a valuable tool for comprehending the temporal patterns and progression of information dissemination on social media platforms. Temporal patterns pertain to the repetitive sequences or cadences of occurrences or phenomena throughout a period. The implementation of time-series analysis facilitates the detection of patterns such as trends, periodicity, and fluctuations in the diffusion process. By analyzing time-series data, researchers can identify temporal trends in the distribution of information. This entails analyzing the pace at which information disseminates, pinpointing instances of heightened activity or involvement, and detecting fluctuations in the diffusion of information across temporal dimensions. Time-series analysis enables the detection of prolonged patterns in data distribution, commonly denoted as trends. Researchers can observe the pattern of

information distribution, detecting any variations in its course and ascertaining whether it is experiencing an upward or downward trend. Applying this analytical approach facilitates understanding the overarching trend of information dissemination throughout social media channels, encompassing both upward and downward inclinations. Graph theory provides a rigorous mathematical framework for examining the properties and features of networks. The field of study known as Graph Theory concerns graphs' mathematical properties and structures, which are collections of vertices and edges that can represent various types of relationships and connections between objects or entities. This area of mathematics has applications in various fields, including computer science, operations research, and social network analysis. Graph theory is applied in social media platforms to represent the underlying network of connections between users or content entities. It is also employed to scrutinize the dissemination of information within this network. Researchers can analyze the network architecture of social media platforms with the aid of graph theory. The procedure involves analyzing the interconnectedness between nodes, which can pertain to users or content entities. Furthermore, the process entails the recognition of nodes that exhibit substantial impact, such as those exhibiting a high degree of centrality or betweenness centrality. Moreover, it involves identifying the community structures present in the network. Comprehending the network topology is essential in understanding the diffusion of information and its susceptibility to influence.

The mechanisms of diffusion: Through the integration of time-series analysis and graph theory, scholars can track the dissemination routes of information throughout a given period. One can trace the diffusion process by identifying the precise channels or routes by which information spreads from its primary sources to subsequent recipients or content items. The present analysis enables the identification of pathways of influence and enhances the understanding of the underlying mechanisms that propel the dissemination of information.

The utilization of visual aids enables the illustration of temporal patterns of information dissemination and network configurations in a cohesive manner, thereby augmenting the effectiveness of communicating insights and discoveries. Various visualization techniques can visually represent time-series analysis and graph theory. These techniques include but are not limited to line plots, bar charts, network graphs, and heat maps.

The integration of time-series analysis and graph theory can provide researchers with a holistic comprehension of the temporal dynamics, progression, and fundamental network architectures that propel the dissemination of information on social media platforms. Implementing a comprehensive methodology enables the identification of regularities, tendencies, significant elements, and dissemination routes, thus providing valuable comprehension regarding the information dissemination mechanisms.

4. How to implement a mechanism to identify the distribution of identifying message contents across multiplex social media networks.

   Multiple mechanisms exist to identify the distribution of identifying message content across multiplex social media networks. However, I have used social networking concepts as the best method for identifying the distribution of identifying message contents across multiplex social media networks.

   Using social networking principles is the ideal strategy for ensuring the spread of identified message contents throughout several connected social media networks. The phrases that follow provide evidence for the viability of these concepts.

   The study of social networking concepts is on comprehending how user connections and interactions are organized in social media networks. By examining the network's structure, researchers may ascertain how information is sent and circulated throughout the system and the routes it follows.

   In social networking, nodes stand in for individuals or pieces of

material, while edges represent their connections or interactions. By examining the nodes and edges of such networks, researchers may ascertain how widely distributed identifiable message contents are in multiplex social media networks.

The concept of ”community detection” refers to finding communities or groups of individuals that engage in comparable discussions or have comparable interests. The study of these communities has the potential to provide significant findings on the distribution of unique message contents across distinct subgroups within the network.

Influence analysis—the research of influential users or content items within a network—is made possible by social networking strategies. By looking at metrics like node centrality, researchers may discover nodes that significantly influence the distribution of identifiable message contents, such as people with high degrees of connectivity or accounts with significant sway.

Information cascades may be better understood by referring to social networking concepts, which deal with user interactions that distribute message content. By monitoring the diffusion of information across cascades, researchers may spot patterns of information dispersion and monitor how message content moves throughout the network.

Monitoring the pathways that the information takes as it goes is necessary to comprehend the distribution of different message contents. Understanding social networking theories may help my research understand how people distribute, retweet, and discuss message content.

Network graphs or other proper visualization techniques may be used to illustrate social networking concepts visually. Using visual representations helps to clarify how identifying message contents are disseminated by emphasizing the relationships between nodes, communities, and information flow throughout the network.

By using social networking principles, academics may better understand the distribution of identifying message contents across multidimensional social media networks. Using this technique, we get

crucial new insights into the communication mechanisms underpinning complex networks, which makes it simpler to examine how networks are set up, how community links develop over time, how nodes impact things, and how information flows.

In conclusion, I have focused on all of the above research questions that I have identified as part of my research.

With respect to the objectives of the research and their achievements following can be summarized as the concluding remarks.

- To develop an algorithm to measure platform-independent information diffusion speed/information diffusivity. i.e., Investigate and implement a mechanism to perform information diffusion in multiplex social media platforms regardless of specific content

  This is one of the main objectives of my research work. Hence I have proposed a novel algorithm that is platform-independent, and that can apply to any number of platforms that can measure the information diffusion speed for a given content. Algorithm 1.1 onwards this algorithm is discussed in this thesis.

  A short overview of algorithm 1.1, including its objective and the problem it tries to answer in assessing the rate of information diffusion for a particular piece of content across several platforms.

  Details of the algorithm: Describe the details of Algorithm 1.1, including the stages it takes and the exact techniques, measures, or methodologies it employs to determine the rate of information dispersion. User interactions, pattern propagation, network analysis, or temporal dynamics could all be considered in this.

  Insist on the recommended algorithm's platform independence, enabling its usage with various social media platforms. Indicate the generalization and applicability of the strategy, highlighting its adaptability and effectiveness.

  Evaluation: Describe the procedure used to evaluate or verify the algorithm. Discuss the datasets used, the performance metrics employed, and the experimental setup to demonstrate the effectiveness

and reliability of the proposed strategy. This analysis should verify the correctness of the algorithm's assessment of the pace of information distribution across multiple platforms. Analyze the recommended algorithm compared to other methods or tactics used in the literature, where relevant. Indicate how the algorithm advances the topic, closes knowledge gaps, or provides enlightening data on the mechanisms of information transmission across numerous social media platforms. Indicate the advantages, creativity, and improvements the recommended algorithm provides over current practices. Discuss any limitations or challenges encountered while putting the algorithm into practice or assessing it.

- To design an information diffusion framework that calculates streaming information diffusivity. i.e., derive a mechanism to identify the information diffusion and its nature on multiplex social media platforms. The completed mechanism identified the trend regardless of the content and its nature.

  This objective focuses on the calculation of information diffusion speed. Algorithm 1.2 onwards (mainly all the algorithms that accept time series analysis) focuses on calculating streaming information diffusivity.

  Algorithm 1.2 in detail, including the steps it takes and the particular techniques, tests, or processes it employs to estimate the diffusivity of streaming data. It has been made clear that the algorithm's objective is to determine the rate of information diffusion while focusing on the diffusivity of streaming information. Beginning with version 1.2, algorithms, particularly those that use time series analysis, have this objective as their primary objective. It is essential to be clear about the purpose of measuring information diffusion speed and streaming information diffusivity. To do this, it could be required to look at time series data, identify diffusion patterns, estimate the rate of information spread, or monitor the rate of information spread over time. Describe the significance of these measurements for comprehending

the dynamics of information diffusion across diverse platforms.

- To implement a mechanism for identifying the distribution of information diffusion across multiplex social media platforms.

  Under this objective, I have focused on implementing the proposed algorithm. All of the coding implementations are given in the analysis section.

- To experiment and evaluate the analytical capabilities of the proposed solution.

The proposed algorithm is robust to handle data distribution on multiple aspects. An actual implementation is essential when it comes to testing. Multiple testing methods are used and experiment with the capabilities of the algorithm. The section ”results” discusses all of the above implementations.

Robustness in handling data distribution: The recommended method has been designed to manage data distribution in several ways. This shows that it can handle a wide range of data distribution properties or patterns, including skewed or uneven distributions, a wide range of data sources, and complex network topologies. The algorithm’s adaptability ensures it can handle various data distribution scenarios.

Actual Implementation: An actual implementation must be performed when testing the algorithm. The approach has been evaluated using real-world data or scenarios instead of theoretical analysis or simulations. The actual application allows for a more accurate and helpful evaluation of the algorithm’s performance and effectiveness.

Multiple Testing Techniques: Several testing methodologies have been utilized to evaluate the capabilities of the proposed algorithm. The algorithm’s effectiveness may be thoroughly assessed by using a range of testing methodologies to examine different aspects of its performance. These methods could come from various experiments, simulations, or benchmarking techniques. Utilizing the assessment metrics and criteria previously developed in this phase, you describe the experimental find-

ings, display the implementation findings, and evaluate the algorithm's performance.

In conclusion, effective analytical tools that may be employed to acquire a deeper understanding of the dissemination of information across multiplex social media platforms include time series analysis, social network analysis, graph theory, and event-driven architecture. Businesses and marketers can gain important insights into user behavior and preferences by analyzing patterns of interaction and influence and developing applications that react to changes in social media activity. These insights can inform their marketing strategies and help them reach their target audience more successfully.

The modern business environment is complex because of a wide range of aspects, user requirements, and technological enhancements, such as globalization, and several of these variables impact business decisions. Making business decisions nowadays is primarily driven by ”data,” i.e., many organizations use ”data-driven business.”

The most used video-sharing website is YouTube. YouTube has grown extremely popular because it allows users to post user-generated material that is free to view (i.e., without a license, subscription, or other fees). YouTube gives consumers various options to interact with videos, including reactions (likes, comments, and sharing within peer communities). It also gives video makers a variety of possibilities to distribute films among communities.

The gathering, examination, and interpretation of data produced by social media platforms are referred to as ”social media data processing.” As social media has grown, enormous volumes of data are produced every second, giving us important insights into user behavior, trends, and preferences. Data collection, cleansing, analysis, and visualization are just a few phases of processing social media data.

Data collection entails acquiring information from social media sites using various tools and techniques, including web scraping, application programming interfaces (APIs), and social listening tools. Data must be cleansed after collection to eliminate duplicate or unnecessary information.

To do this, the data must be sorted, filtered, and organized for simple analysis.

The following step is data analysis, which entails using various methods to comprehend and analyze the data. Some examples of data analysis techniques are descriptive statistics, sentiment analysis, network analysis, and machine learning algorithms. Descriptive statistics include measures like frequency, distribution, and central tendency, which give a broad picture of the data. The tone and emotions represented in social media messages may be understood via sentiment analysis. Network analysis may be used to locate essential communities and influencers within a social network. Algorithms for machine learning can be used to spot trends and patterns in data.

The last step is data visualization, which comes after data analysis. Data visualization aims to make the data simple to grasp and analyze. These can include maps, graphs, and charts that show the data's patterns and trends.

Processing social media data has emerged as a crucial tool for companies, marketers, and researchers trying to comprehend consumer preferences and behavior. Analyzing social media data, organizations may learn more about consumer attitudes, market trends, and hot topics. To preserve user privacy and avoid data exploitation, it is crucial to ensure that social media data is gathered and used responsibly, following privacy laws.

All the algorithms mentioned above analyze social media posts and determine spread patterns. The first method generates a diffusion tree for each post in the collection, with each tree representing the flow of information from that post to other posts in the collection. The system then blends trees as needed depending on post-homologous similarities.

The second technique is similar but contains integrated time series analysis, allowing for post-weighting depending on timestamps. This method computes the similarity of each pair of posts and adds directed edges between them if the similarity is more significant than a predefined threshold. The generated diffusion trees depict the flow of information over time, and the algorithm seeks to uncover diffusion trends.

Both methods help understand information flow in social networks and

might be used for a variety of purposes, including recognizing notable persons and detecting misinformation dissemination.

The first technique, ”Enhanced Diffusion Tree Algorithm with Social Networking Concepts,” is a strategy for finding diffusion trees in a social network. It takes as input a collection of social media postings, each with a timestamp, a unique identity, and other tags. The approach sorts the posts by date, creates a new diffusion tree for each post, and creates a new set of diffusion trees. Based on the tags for each post, the algorithm computes the similarity between each post and all following posts. A directed edge is added between the nodes representing the earlier and later postings if the similarity exceeds a predefined threshold. If the subsequent post is already present in a diffusion tree,

The second technique, ”Algorithm with Embedded Time Series Analysis,” is a time series analysis diffusion tree version. It takes as input the same collection of social media postings, a similarity criterion, and a weight function. The technique sorts the posts according to their timestamp, creates a new tree for each post, and starts with an empty set of trees. The approach computes the similarity between each post and all future posts based on their tags and weight, depending on the difference in their timestamps. A directed edge from the earlier post node to the last post node is added if the similarity exceeds the threshold. If the following post has The presented algorithm is a social network analysis tool that attempts to extract information from social media postings and identify trends in information propagation. The approach employs diffusion trees, directed trees that illustrate knowledge transfer in a social network.

To begin, the algorithm sorts the posts by date and generates an empty diffusion tree collection. It then loops through each post, creating a new diffusion tree with a single node for each post. The method then iterates through all posts with later timestamps, applying a weight function based on their timestamp relative to the initial post for each post.

If the similarity exceeds a predetermined threshold, the approach inserts a directed edge in the diffusion tree from the original post node to the last post node. The approach merges the new and existing diffusion trees if the

later post is already in a diffusion tree.

The approach generates a series of diffusion trees that describe information dispersion patterns in a social network.

The algorithm's performance depends on the size of the social media dataset and the complexity of the weight function used to calculate post similarity. It can predict the spread of rumors or disinformation, identify influential persons in a social network, and examine social network dynamics over time.

The method presented in this question is a valuable tool for studying information distribution patterns in social media, and it has various possible applications in social network analysis.

Chapter 7

CONCLUSION

In today's intricate business world, data-driven decisions heavily rely on data from various sources. YouTube, Facebook, and Twitter are leading social media platforms that allow users to post and engage with free, user-generated content. Social media data processing involves gathering, cleaning, analyzing, and visualizing vast data volumes to gain insights into user behavior and trends. Techniques such as statistics, sentiment analysis, network analysis, and machine learning aid this process. Responsible data collection and usage, following privacy laws, are emphasized.

This research delves into various technical methodologies, such as time series analysis, social network analysis, graph theory, and event-driven architecture, to understand how information propagates on diverse social media platforms. Time series analysis identifies trends in user interactions, aiding in pinpointing popular content during specific events. Social network analysis reveals influential users and group dynamics, which is essential for understanding interactions within the network. These analytical approaches shed light on information diffusion patterns, contributing to a more informed and responsible use of social media for a better society. The research explores graph theory to analyze social networks and their influence patterns. Event-driven architecture is highlighted as a method to create responsive applications, especially in social media monitoring. The study focuses on two main research questions:

1. investigating information diffusion mechanisms in multiplex social media latforms using a graph networking approach

2. implementing a tree-based graph-structured network as an efficient mechanism for information diffusion in complex social media environments.

Tree structures are emphasized for their efficiency, clear visualization, and scalability in understanding information dissemination pathways, making them suitable for analyzing data distribution across various social media channels. However, the methodology selection was tried to align with specific research objectives and platform characteristics.

The report discusses implementing mechanisms to identify the distribution of message content across multiple social media networks. It focused on using social networking concepts as a practical approach. Understanding user connections, network structure, and community detection within social media networks helps comprehend information dissemination patterns—influence analysis, information cascades, and monitoring information pathways aid in understanding message content distribution. Utilizing network graphs for visualization enhances clarity in disseminating and identifying message contents. The approach helps with communication mechanisms in complex networks, facilitating a better understanding of network setup, community links, node impact, and information flow.

The primary research objective involves developing a platform-independent algorithm to measure information diffusion speed, achieving this through a novel algorithm outlined in the paper.

The report discusses Algorithms 1.1,1.2, and 1.3, explaining their stages, methodologies, and platform independence, emphasizing their adaptability and effectiveness across various social media platforms. The algorithm focuses on determining the rate of information dispersion, considering user interactions, pattern propagation, network analysis, and temporal dynamics. Algorithms discussed focus on analyzing social media posts and understanding information spread patterns. The ”Enhanced Diffusion Tree Algorithm with Social Networking Concepts” and the ”Algorithm with Embedded Time Series Analysis” aim to identify diffusion trends in social networks, utilizing techniques like similarity computation and time series analysis. These algorithms offer insights into information

flow, aiding in recognizing influential individuals and detecting information diffusion.

The evaluation procedure for the algorithm is outlined, including the datasets used, performance metrics, and experimental setup to demonstrate its effectiveness in assessing the pace of information distribution across multiple platforms. Comparative analysis is conducted against existing methods in the literature to highlight advancements, data enlightenments, and improvements brought about by the recommended algorithm. The paper also addresses limitations and challenges encountered during the algorithm's implementation and evaluation.

Additionally, the objectives related to designing an information diffusion framework that calculates streaming information diffusivity and implementing a mechanism for identifying the distribution of information diffusion are discussed. The focus is on measuring information diffusion strength and streaming information diffusivity, clearly describing their significance in comprehending information diffusion dynamics across diverse platforms. The coding implementations related to these objectives are provided in the analysis section.

The proposed algorithms demonstrate robustness in handling diverse data distributions, including skewed or uneven patterns, various data sources, and complex network topologies. Its adaptability ensures effectiveness across a spectrum of data distribution scenarios.

Actual implementation is emphasized for accurate testing, focusing on real-world data or scenarios rather than theoretical analysis or simulations. This approach provides a more precise evaluation of the algorithm's performance and effectiveness.

Multiple testing techniques have been employed to evaluate the algorithm's capabilities comprehensively. These methodologies, drawn from experiments, simulations, or benchmarking, allow a thorough assessment of performance aspects using predefined assessment metrics.

In conclusion, analytical tools like time series analysis, social network analysis, graph theory, and event-driven architecture offer valuable insights into information dissemination across multiple social media plat-

forms. Businesses and marketers can utilize these insights to understand user behavior, preferences, interaction patterns, and influence dynamics, enabling informed marketing strategies to target their audience effectively.

7.1 Limitations and drawbacks

The main drawbacks of these algorithms can be outlined as follows:

All approaches are rather complex, with several nested loops and computing operations. As a result, large datasets may require more work to deploy and scale.

The algorithms make assumptions about the structure and features of social media data, such as the presence of tags and the importance of timestamp data.

A variety of user-defined parameters, such as the similarity threshold and weight function, are used by the algorithms. The selection of these elements can significantly impact the output of the algorithms, but there are sometimes clear guidelines for doing so.

Sensitivity to noise: social media data can be noisy, containing irrelevant or misleading information. Because the algorithms are subject to noise, the results may need to be corrected or made trustworthy.

Ethical concerns: Because the algorithms rely on user-generated data, there are concerns regarding privacy, consent, and potential bias. Considering these problems before adopting or deploying these algorithms properly is necessary.

7.2 Future work

Extension to more Social Media Platforms: extend the framework to include more social media platforms not initially considered. This can include altering the paradigm to account for various platforms' unique features and capabilities, enabling a more in-depth analysis of information dispersion across various social media ecosystems.

Integration of Advanced Machine Learning Techniques: Include ma-

chine learning techniques in the framework to enhance the analysis of information dispersion. This may include using algorithms for content classification, sentiment analysis, or user behavior prediction to provide more accurate and nuanced insights into how information is distributed and influences users' behaviors within multiplex social media networks.

Look at how the framework may be adjusted to consider influence and opinion dynamics. This might include examining the relationships between important users, thought leaders, and information transmission. It can also entail investigating how views evolve and how that influences how information moves through multiplex social media networks.

Information dispersion analysis should be done while considering privacy and ethical considerations. Gathering and analyzing user data entails addressing potential privacy risks, ensuring that data is anonymized and in compliance with privacy laws, and considering the ethical repercussions of manipulating information or unintended consequences of tracking and influencing information diffusion.

Last but not least, I desire to emphasize these algorithms' value to society as they arise from their capacity to assist us in better interpreting and analyzing social media data, which is a rich source of information about human behavior and social interactions. By detecting and evaluating patterns in social media postings and the networks of links between them, we may get insights into a wide variety of social phenomena, such as the diffusion of ideas and information, the development of social groups, and the emergence of cultural trends. Such insights can help policymakers, marketers, and academics make informed judgments regarding various subjects, including public health programs, marketing strategies, and social interventions. For example, social media data analysis has been used to track the spread of infectious diseases, monitor public opinion on political issues, identify groups at risk of radicalization, and track the spread of contagious diseases.

Overall, the algorithm's effectiveness stems from its ability to help us make sense of the big and intricate social media world and then utilize that knowledge to impact our actions and judgments in various disciplines.

Appendix A

Appendix

A.1 This code is a sample implementation of the base algorithm

```
def measure_information_diffusion(posts, threshold):
    # Sort the posts using timestamp
    posts = sorted(posts, key=lambda x: x.timestamp)

    # Initialize empty trees
    trees = set()

    # Iterate each post
    for i, p_i in enumerate(posts):
        # Initialize a new diffusion tree
        T_i = {p_i}
        for j in range(i + 1, len(posts)):
            p_j = posts[j]
            # Compute the similarity of p_i and p_j/ Given tags
            similarity = compute_similarity(p_i, p_j)
            # Check with the similarities and thresholds
            if similarity > threshold:
                T_i.add(p_j)
            # If p_j has added to a tree then merge T_i with that
            for T_j in trees:
```

```
                if p_j in T_j:
                    T_j.update(T_i)
                    break
            else:
                # add T_i to trees.
                # coz p_j hasn't been added tree, then it works fo
                trees.add(T_i)

    # Return trees. this will gives all the vales of the combinati
    return trees
```

A.2 This code is a sample implementation of the algorithm that working with a timeseries data

```
import pandas as pd
import numpy as np
import datetime as dt
from sklearn.metrics.pairwise import cosine_similarity

# Load data
df = pd.read_csv('postings.csv')
df['timestamp'] = pd.to_datetime(df['timestamp'])

# Sort DataFrame
df = df.sort_values(by='timestamp')

# Initialize empty dictionary
T = {}

# this function is calculating similarity between two sets of tags
def calculate_similarity(tag_set1, tag_set2):
    vec1 = np.array(tag_set1).reshape(1,-1)
    vec2 = np.array(tag_set2).reshape(1,-1)
```

```
    return cosine_similarity(vec1, vec2)[0][0]

# Iterate through each post in the DataFrame
for i, post_i in df.iterrows():
    # Initialize tree for post i
    T[i] = {i: []}
    # Iterate with all posts that are later than origial
    for j, post_j in df.iloc[i+1:].iterrows():
        # Similarity calculation
        similarity = calculate_similarity(set(post_i['tags'].split
        if similarity > 0.5:
            T[i][i].append(j)
        # Merge trees
        for key, value in T.items():
            if j in value:
                T[i] = {**T[i], **{key: value}}
    # Add a tree for post i. This will collect all trees
    if i == 0:
        diffusion_trees = T[i]
    else:
        diffusion_trees = {**diffusion_trees, **T[i]}

# trees
print(diffusion_trees)
```

Sample code

```
# extracting YouTube video delails
import os
import google.auth
import google.auth.transport.requests
from google.oauth2.credentials import Credentials
from googleapiclient.discovery import build
```

```
# Set up YouTube API credentials
credentials, project_id = google.auth.default(scopes=

["https://www.googleapis.com/auth/youtube.force-ssl"])
if not credentials:
    credentials_path = os.path.join(os.getcwd(), 'client_secrets.j
    flow =

    google.auth.flow.InstalledAppFlow.from_client_secrets_file(cre

    scopes=["https://www.googleapis.com/auth/youtube.force-ssl"])

    credentials = flow.run_console()

# Set up YouTube API client
youtube = build('youtube', 'v3', credentials=credentials)

# Set up search query parameters
search_query = 'python programming'
max_results = 10

# Execute search query
search_response = youtube.search().list(
    q=search'_query,
    type='video',
    part='id,snippet',
    maxResults=max_results
).execute()

# Extract video data
videos = []
for search_result in search_response.get('items', []):
    video = {}
```

```
    if search_result['id']['kind'] == 'youtube#video':
        video['title'] = search_result['snippet']['title']
        video['description'] = search_result['snippet']['descripti
        video['thumbnail'] = search_result['snippet']['thumbnails'
        video['url'] = f"https://www.youtube.com/watch?v={search_r
        videos.append(video)

# Print video data
for video in videos:
    print(f"Title: {video['title']}")
    print(f"Description: {video['description']}")
    print(f"Thumbnail: {video['thumbnail']}")
    print(f"URL: {video['url']}\n")
```

Sample Facebook API key - EAAGrMLDZA67gBAEStGyKMVhLIQZCDhdbVGSSHYM

A.2.1 Parameter definitions for the proposed algorithm

Input Parameters:

- P: A collection of social media posts, where each post is represented by a timestamp, a unique identity, and other tags. This is the input data for the algorithm.
- Predetermined Threshold: A value that defines the minimum similarity required for two posts to be connected in the diffusion tree.

Output Parameter:

- T: A collection of diffusion trees, where each diffusion tree represents the propagation of posts in the social media network.

Internal Variables/Parameters:

- T_i: A diffusion tree representing the propagation from a specific post p_i.
- T: The collection of diffusion trees.
- p_i: An individual post from the collection P.
- p_j: An individual post from the collection P that has a timestamp later than p_i.
- Similarity: A measure of similarity between the tags of two posts, used to determine if a connection should be made in the diffusion tree.
- T_k: A diffusion tree to which post p_j has already been added, if applicable.

A.2.2 Diffusion Tree Construction Algorithm Implementation Guide

A.3 Introduction

This implementation guide provides step-by-step instructions for implementing the Diffusion Tree Construction Algorithm in Python. The algorithm constructs diffusion trees from a collection of social media posts with timestamps, unique identities, and tags.

A.4 Prerequisites

Before implementing the algorithm, ensure you have the following prerequisites:

- Python: You should have Python installed on your system. You can download Python from the official website: https://www.python.org/downloads/

- Required Libraries: The algorithm uses standard Python libraries, so no additional libraries are needed.

A.5 Implementation Steps

To implement the Diffusion Tree Construction Algorithm, follow these steps:

A.5.1 Data Preparation

Prepare a collection of social media posts with timestamps, unique identities, and tags. You can load this data from a file or generate it programmatically.

A.5.2 Algorithm Implementation

The following Python code implements the Diffusion Tree Construction Algorithm. It assumes that you have loaded the data into a list called ‘posts‘.

```
def construct_diffusion_trees(posts, threshold):
    # Sort the posts by timestamp
    posts.sort(key=lambda post: post['timestamp'])

    diffusion_trees = []

    for post_i in posts:
        tree_i = DiffusionTree(post_i)

        for post_j in posts:
            if post_j['timestamp'] > post_i['timestamp']:
                similarity = compute_similarity(post_i, p
                if similarity > threshold:
                    tree_i.add_edge(post_i, post_j)
```

```
                    for tree_k in diffusiontrees:
                        if post_j in tree_k:
                            tree_i.merge(tree_k)
                            diffusion_trees.remove(tree_k)

            diffusiontrees.append(tree_i)

    return diffusiontrees

class DiffusionTree:
    def __init__(self, root):
        self.nodes = [root]
        self.edges = []

    def add_edge(self, source, target):
        self.edges.append((source, target))

    def merge(self, other_tree):
        self.nodes.extend(other_tree.nodes)
        self.edges.extend(other_tree.edges)
```

The ”construct diffusion trees” function takes a list of posts and a similarity threshold as input and returns a list of diffusion trees.

A.5.3 Usage

You can use the following code to construct diffusion trees and analyze them:

```
# Example usage of the algorithm
threshold= 0.7
diffusion_trees = construct_diffusion_trees(posts, thresho

# Analyze the diffusion trees, e.g., calculate tree sizes,
```

A.6 Conclusion

This implementation guide provides an overview of a step-by-step approach to implementing the Diffusion Tree Construction Algorithm in Python. By following these steps, you can construct diffusion trees and analyze the propagation of posts in a social media network.

B.1 List of publications

- Long-Term Trend Analysis for Social Media Content Published During COVID-19 Pandemic
- A Novel Annotation Scheme to Generate Hate Speech Corpus through Crowdsourcing and Active Learning
- Classification of Trending Videos in YouTube
- Collective Sentimental Trend Identification for Social Media Content using Time Series Forecasting
- A User Experience Measuring Technique to Moderate Social Media Contents through Crowdsourcing
- A Peer Recommendation Model to Avoid Hate Speech Engagements in Multiplex Social Networks
- Hate speech corpus generation using crowdsourcing
- Time Series Based Trend Analysis for Hate Speech in Sri Lankan Social Media Platforms During COVID 19 Pandemic
- A comparative study of the characteristics of hate speech propagators and their behaviours over Twitter social media platform

Bibliography

[1] S. Dixon, " Statista social media statistics and facts, the global statistics," https://www.statista.com/topics/1164/social-networks/ #topicOverview, accessed: 2023-01-30.

[2] ——, "Digital 2023: Sri lanka — datareportal – global digital in- sights," https://datareportal.com/reports/digital-2023-sri-lanka, ac- cessed: 2023-01-30.

[3] D. C. Raza, Shaina, "Fake news detection based on news content and social contexts: a transformer-based approach," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, May 2022. [Online]. Available: https://doi.org/10.1007/s41060-021-00302-z

[4] Z. S. Dong, L. Meng, L. Christenson, and L. Fulton, "Social media information sharing for natural disaster response," *Natural Hazards*, vol. 107, no. 3, pp. 2077–2104, Jul 2021. [Online].
Available: https://doi.org/10.1007/s11069-021-04528-9

[5] T. K. K. D. S. K. S. T. K. R. Sharma, A., "Disaster analysis through tweets," in *Third Congress on Intelligent Systems*. Singapore: Springer Nature Singapore, 2023, pp. 543–554.

[6] D. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[7] A. Nash, "Chapter 1 - affect, people, and digital social networks," in *Emotions, Technology, and Social Media*, ser. Emotions and

Technology, S. Y. Tettegah, Ed. San Diego: Academic Press, 2016, pp. 3–23. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128018576000014

[8] E. Lai, Linda S. L.and Turban, "Groups formation and operations in the web 2.0 environment and social networks," *Group Decision and Negotiation*, vol. 17, no. 5, pp. 387–402, Sep 2008. [Online]. Available: https://doi.org/10.1007/s10726-008-9113-2

[9] H. Purtik and D. Arenas, "Embedding social innovation: Shaping societal norms and behaviors throughout the innovation process," *Business & Society*, vol. 58, no. 5, pp. 963–1002, 2019. [Online]. Available: https://doi.org/10.1177/0007650317726523

[10] N. K. Hayles, *How we think: Digital media and contemporary technogenesis*. University of Chicago Press, 2012.

[11] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, "Social features of online networks: The strength of in- termediary ties in online social media," *PloS one*, vol. 7, no. 1, p. e29358, 2012.

[12] D. Camacho, Á. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria, "The four dimensions of social network analysis: An overview of research methods, applications, and soft- ware tools," *Information Fusion*, vol. 63, pp. 88–120, 2020.

[13] A. Majeed and I. Rauf, "Graph theory: A comprehensive survey about graph theory applications in computer science and social net- works," *Inventions*, vol. 5, no. 1, p. 10, 2020.

[14] R. F. Bales, "Interaction process analysis: A method for the study of small groups," *Cambridge, Mass.: Addison-Wesley*, 1950.

[15] ——, "A set of categories for the analysis of small group interaction," *American Sociological Review*, vol. 15, no. 2, pp. 257–263, 1950. [Online]. Available: http://www.jstor.org/stable/2086790

[16] H. M. M. Caldera, G. S. N. Meedin, and I. Perera, “Time series based trend analysis for hate speech in twitter during covid 19 pan- demic,” in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2020, pp. 1–2.

[17] K. Z. Khanam, G. Srivastava, and V. Mago, “The homophily prin- ciple in social network analysis,” *arXiv preprint arXiv:2008.10383*, 2020.

[18] G. Vasanthakumar, “Cascading behavior in networks,” *Social Net- work Analysis: Theory and Applications*, pp. 51–61, 2022.

[19] M. Cai, H. Luo, and Y. Cui, “A study on the topic-sentiment evo- lution and diffusion in time series of public opinion derived from emergencies,” *Complexity*, vol. 2021, pp. 1–23, 2021.

[20] G. F. Hollewell and N. Longpré, “Radicalization in the social media era: Understanding the relationship between self-radicalization and the internet,” *International journal of offender therapy and compar- ative criminology*, vol. 66, no. 8, pp. 896–913, 2022.

[21] R. R. Mourão and D. K. Brown, “Black lives matter coverage: How protest news frames and attitudinal change affect social media en- gagement,” *Digital Journalism*, vol. 10, no. 4, pp. 626–646, 2022.

[22] G. G. W and K. R. M, “Cyberbullying via social media and well- being,” *Current Opinion in Psychology*, p. 101314, 2022.

[23] M. Saldaña and H. T. Vu, “You are fake news! factors impacting journalists’ debunking behaviors on social media,” *Digital Journal- ism*, vol. 10, no. 5, pp. 823–842, 2022.

[24] Y. Chen and L. Wang, “Misleading political advertising fuels incivility online: A social network analysis of 2020 u.s. presidential election campaign video comments on youtube,” *Comput. Hum. Behav.*, vol. 131, no. C, jun 2022. [Online]. Available: https://doi.org/10.1016/j.chb.2022.107202

[25] S. Sinha, S. Bhattacharya, and S. Roy, “Impact of second-order network motif on online social networks,” *The Journal of Super- computing*, vol. 78, no. 4, pp. 5450–5478, 2022.

[26] W. Lee, “Machine learning and security: The good, the bad, and the ugly,” in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–2. [Online]. Available: https://doi.org/10.1145/3372297.3424552

[27] A. Ceron, *Elgar Encyclopedia of Technology and Politics*. Cheltenham, UK: Edward Elgar Publishing, 2022. [Online]. Avail- able: https://www.elgaronline.com/view/book/9781800374263/ 9781800374263.xml

[28] D. L. Hoffman, C. P. Moreau, S. Stremersch, and M. Wedel, “The rise of new technologies in marketing: A framework and outlook,” *Journal of Marketing*, vol. 86, no. 1, pp. 1–6, 2022. [Online].
Available: https://doi.org/10.1177/00222429211061636

[29] J. a. L. H. Frade, J. H. C. d. Oliveira, and J. d. M. E. Giraldi, “Advertising in streaming video: An integrative literature review and research agenda,” *Telecommun. Policy*, vol. 45, no. 9, oct 2021. [Online]. Available: https://doi.org/10.1016/j.telpol.2021.102186

[30] S. K. V, N. KP, and G. B. Kamath, “Social media advertisements and their influence on consumer purchase intention,” *Cogent Business & Management*, vol. 8, no. 1, p. 2000697, 2021. [Online].
Available: https://doi.org/10.1080/23311975.2021.2000697

[31] M. T. Febriyantoro, “Exploring youtube marketing communication: Brand awareness, brand image and purchase intention in the millennial generation,” *Cogent Business & Management*, vol. 7, no. 1, p. 1787733, 2020. [Online]. Available: https://doi.org/10.1080/23311975.2020.1787733

[32] A. O. Insorio and D. M. Macandog, "Video lessons via youtube channel as mathematics interventions in modular distance learning," *Contemporary Mathematics and Science Education*, vol. 3, no. 1, p. ep22001, 2022.

[33] S. Asur and B. A. Huberman, "Predicting the future with social me- dia," *Proceedings of the 2010 IEEE/WIC/ACM International Confer- ence on Web Intelligence and Intelligent Agent Technology-Volume 1*, pp. 492–499, 2010.

[34] O. Boichak, "511Digital War: Mediatized Conflicts in Sociological Perspective," in *The Oxford Handbook of Digital Media Sociology*. Oxford University Press, 09 2022. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780197510636.013.31

[35] H. Pang, "Connecting mobile social media with psychosocial well- being: Understanding relationship between wechat involvement, network characteristics, online capital and life satisfaction," *Social Networks*, vol. 68, pp. 256–263, 2022.

[36] YouTube, "Statistics," https://www.youtube.com/about/press/, n.d., accessed January 3, 2023.

[37] D. Conway, "The data science venn diagram." https: //drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, accessed: 2023-01-30.

[38] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2014.

[39] R. E. Kent and C. Neuss, "Creating a web analysis and visualization environment," *Comput. Netw. ISDN Syst.*, vol. 28, no. 1–2, p. 109–117, dec 1995. [Online]. Available: https://doi.org/10.1016/0169-7552(95)00095-X

[40] R. T. Fielding and R. N. Taylor, *Architectural Styles and the Design of Network-based Software Architectures*. Irvine, CA: University of California, Irvine, 2000.

[41] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min- redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[42] J. Anitha, I.-H. Ting, S. A. Agnes, S. I. A. Pandian, and R. Belfin, "Chapter 3 - social media data analytics using feature engineering," in *Systems Simulation and Modeling for Cloud Computing and Big Data Applications*, ser. Advances in ubiquitous sensing applications for healthcare, J. D. Peter and S. L. Fernandes, Eds. Academic Press, 2020, pp. 29–59. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128197790000034

[43] G. Dong and H. Liu, *Feature Engineering for Machine Learning and Data Analytics*, 1st ed. USA: CRC Press, Inc., 2018.

[44] M. Newman, *Networks: An Introduction*, 1st ed. Oxford, UK: Oxford University Press, 2010.

[45] A. Hodler and M. Needham, "Introducing graph data science," https://neo4j.com/blog/introducing-graph-data-science/, July 2019, accessed on March 13, 2023.

[46] K. Ognyanova, "Social network analysis," in *Elgar Encyclopedia of Technology and Politics*. Edward Elgar Publishing, pp. 126–130.

[47] J. Zhang and P. S. Yu, *Information Diffusion*. Cham: Springer International Publishing, 2019, pp. 315–349. [Online]. Available: https://doi.org/10.1007/978-3-030-12528-8 9

[48] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987. [Online]. Available: https://doi.org/10.1086/228631

[49] C. Laghridat and M. Essalih, "A set of measures of centrality by level for social network analysis," *Procedia Computer Science*, vol. 219, pp. 751–758, 2023, cENTERIS – International

Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022. [Online]. Available: https:

//www.sciencedirect.com/science/article/pii/S1877050923003575

[50] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID: 334423

[51] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 137–146. [Online].

Available: https://doi.org/10.1145/956750.956769

[52] W. Kermack and A. McKendrick, "Contributions to the mathematical theory of epidemics—i," *Bulletin of Mathematical Biology*, vol. 53, no. 1, pp. 33–55, 1991. [Online]. Available: https:

//www.sciencedirect.com/science/article/pii/S0092824005800400

[53] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, pp. 391–407, 1990. [Online]. Available: https://api.semanticscholar.org/CorpusID:3252915

[54] D. D. Lee, M. Hill, and H. S. Seung, "Algorithms for non- negative matrix factorization," in *Neural Information Processing Systems*, 2000. [Online]. Available: https://api.semanticscholar.org/ CorpusID:2095855

[55] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation,"
J. Mach. Learn. Res., vol. 3, no. null, p. 993–1022, mar 2003.

[56] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 137–146. [Online]. Available: https://doi.org/10.1145/956750.956769

[57] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978. [Online]. Available: https://doi.org/10.1086/226707

[58] G. Box, *Box and Jenkins: Time Series Analysis, Forecasting and Control*. London: Palgrave Macmillan UK, 2013, pp. 161–215. [Online]. Available: https://doi.org/10.1057/9781137291264_6

[59] P. N. Nohuddin, F. Coenen, R. Christley, C. Setzkorn, Y. Patel, and S. Williams, "Finding "interesting" trends in social networks us- ing frequent pattern mining and self organizing maps," *Knowledge- Based Systems*, vol. 29, pp. 104–113, 2012.

[60] D. Y. T. A. A. Avetisyana, M. D. Drobyshevskiya and T. Ghukasyane, "Methods for information diffusion analysis." *Programming and Computer Software*, vol. 45, no. 1, pp. 1608–3261, 2019.

[61] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013. [Online]. Available: https://doi.org/10.2753/MIS0742-1222290408

[62] R. Li and A. Suh, "Factors influencing information credibility on social media platforms: Evidence from facebook pages," *Procedia Computer Science*, vol. 72, pp. 314–328, 2015, the Third Information Systems International Conference 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915036078

[63] J. G. Blumler and E. Katz, “The uses of mass communications: Cur- rent perspectives on gratifications research,” *Sage Annual Reviews of Communication Research: Advancing Communication Science: Merging Mass and Interpersonal Processes*, vol. 2, pp. 144–171, 1974.

[64] P. N. Petratos, “Misinformation, disinformation, and fake news: Cyber risks to business,” *Business Horizons*, vol. 64, no. 6, pp. 763–774, 2021, cIBER SPECIAL ISSUE:
CYBERSECURITY IN CRISIS. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000768132100135X

[65] M. Wang and K. Li, “Predicting information diffusion cascades using graph attention networks,” in *Neural Information Processing*,
H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham: Springer International Publishing, 2020, pp. 104–112.

[66] Z. Zhang and Z. Wang, “The data-driven null models for information dissemination tree in social networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 484, pp. 394–411, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437117304995

[67] S. A. Z. H. Goel A., Munagala K., “A note on modeling retweet cascades on twitter,” in *Algorithms and Models for the Web Graph*. Cham: Springer International Publishing, 2015, pp. 119–131.

[68] S. W. McCormack R., “An application of epidemiological modeling to information diffusion,” in *Advances in Social Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 382–389.

[69] F. Zhou, X. Xu, G. Trajcevski, and K. Zhang, “A survey of information cascade analysis: Models, predictions, and recent advances,” *ACM Comput. Surv.*, vol. 54, no. 2, mar 2021. [Online]. Available: https://doi.org/10.1145/3433000

[70] N. L. T. Charalambos Christoforou, Kalliopi Malerou and A. Vakali, "Difcurv: A unified framework for diffusion curve fitting and prediction in online social networks," *Array*, vol. 12, p. 100100, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005621000448

[71] Z. Y. Ren X., "Predicting information diffusion in social networks with users' social roles and topic interests," in *Information Retrieval Technology*. Cham: Springer International Publishing, 2016, pp. 349–355.

[72] A. Saxena, S. R. Iyengar, and Y. Gupta, "Understanding spreading patterns on social networks based on network topology," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ser. ASONAM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1616–1617. [Online]. Available: https://doi.org/10.1145/2808797.2809360

[73] R. R. Singh, *Centrality Measures: A Tool to Identify Key Actors in Social Networks*. Singapore: Springer Singapore, 2022, pp. 1–27. [Online]. Available: https://doi.org/10.1007/978-981-16-3398-0 1

[74] Y. Wang and G. Chirikjian, "A diffusion-based algorithm for workspace generation of highly articulated manipulators," in *Pro- ceedings 2002 IEEE International Conference on Robotics and Au- tomation (Cat. No.02CH37292)*, vol. 2, 2002, pp. 1525–1530 vol.2.

[75] K. S. H. J. Z. W. S. S. Majbouri Yazdi K., Yazdi A.M., "Integrating ant colony algorithm and node centrality to improve prediction of information diffusion in social networks," in *Security, Privacy, and Anonymity in Computation, Communication, and Storage*. Cham: Springer International Publishing, 2018, pp. 381–391.

[76] E. Yoo, B. Gu, and E. Rabinovich, "Diffusion on social media platforms: A point process model for interaction among

similar content," *Journal of Management Information Systems*, vol. 36, no. 4, pp. 1105–1141, 2019. [Online]. Available: https://doi.org/10.1080/07421222.2019.1661096

[77] Q. Bao, W. K. Cheung, Y. Zhang, and J. Liu, "A component-based diffusion model with structural diversity for social networks," *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 1078–1089, 2017.

[78] R. Sharma, T. Arya, S. Arora, A. Arya, and P. Agarwal, "A naive deep nets based approach for authenticating viral textual content on social media," in *Intelligent Systems and Applications*, K. Arai,
S. Kapoor, and R. Bhatia, Eds. Cham: Springer International Publishing, 2019, pp. 679–689.

[79] L. Liu, B. Qu, B. Chen, A. Hanjalic, and H. Wang, "Modelling of information diffusion on social networks with applications to wechat," *Physica A: Statistical Mechanics and its Applications*, vol. 496, pp. 318–329, 2018. [Online]. Available: https:
//www.sciencedirect.com/science/article/pii/S0378437117312785

[80] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Discov. Data*, vol. 5, no. 4, feb 2012. [Online]. Available: https://doi.org/10.1145/2086737.2086741

[81] M. EL-MOUSSAOUI, T. AGOUTI, A. TIKNIOUINE, and M. E. ADNANI, "A comprehensive literature review on community detection: Approaches and applications," *Procedia Computer Science*, vol. 151, pp. 295–302, 2019, the 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops. [Online]. Available: https:
//www.sciencedirect.com/science/article/pii/S1877050919305046

[82] A. Ahmad, T. Ahmad, and A. Bhatt, "Hwsmcb: A community- based hybrid approach for identifying influential nodes in

the social network," *Physica A: Statistical Mechanics and its Applications*, vol. 545, p. 123590, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437119319983

[83] C. M. G. F. Berlingerio, M., "Mining the temporal dimension of the information propagation," in *Advances in Intelligent Data Analysis VIII. IDA 2009. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2009, pp. 1163–1168.

[84] A. Antelmi, G. Cordasco, C. Spagnuolo, and P. Szufel, "Information diffusion in complex networks: A model based on hypergraphs and its analysis," in *Algorithms and Models for the Web Graph*, B. Kamiński, P. Pra lat, and P. Szufel, Eds. Cham: Springer International Publishing, 2020, pp. 36–51.

[85] L. Jain, R. Katarya, and S. Sachdeva, "Opinion leaders for information diffusion using graph neural network in online social networks," *ACM Trans. Web*, vol. 17, no. 2, apr 2023. [Online]. Available: https://doi.org/10.1145/3580516

[86] ——, "Opinion leaders for information diffusion using graph neural network in online social networks," *ACM Trans. Web*, vol. 17, no. 2, apr 2023. [Online]. Available: https://doi.org/10.1145/3580516

[87] K. Lytvyniuk, R. Sharma, and A. Jurek-Loughrey, "Predicting infor- mation diffusion in online social platforms: A twitter case study," in *Complex Networks and Their Applications VII*, L. M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, and L. M. Rocha, Eds. Cham: Springer International Publishing, 2019, pp. 405–417.

[88] S. P. Borgatti, "Identifying sets of key players in a social network," *Comput. Math. Organ. Theory*, vol. 12, no. 1, p. 21–34, apr 2006. [Online]. Available: https://doi.org/10.1007/s10588-006-7084-x

[89] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*. Springer, 2017.

[90] G. E. Box, G. M. Jenkins, G. C. Reinsel, and L. Ljung, *Time Series Analysis: Forecasting and Control*. Wiley, 2015.

[91] Z. Liu, Z. Zhu, J. Gao, and C. Xu, "Forecast methods for time series data: A survey," *IEEE Access*, vol. 9, pp. 91 896–91 912, 2021.

[92] C. Chatfield, *The Analysis of Time Series: An Introduction*. CRC Press, 2016.

[93] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer, 2016.

[94] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[95] E. S. Gardner, *Exponential smoothing: the state of the art*. Springer, 1985.

[96] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993– 1022, 2003.

[97] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," *Stanford knowledge systems laboratory technical report*, vol. No. KSL-01-05, pp. 1–24, 2001.

[98] X.-X. Z. X. L. C.-X. Z. Zi-Ke Zhang, Chuang Liu and Y.-C. Zhang, "Dynamics of information diffusion and its applications on complex networks," *Physics Reports*, vol. 651,
pp. 1–34, 2016, dynamics of information diffusion and its applications on complex networks. [Online]. Available: https:
//www.sciencedirect.com/science/article/pii/S0370157316301600

[99] C. J. G. Juan Pablo Alperin and S. Haustein, "Identifying diffusion patterns of research articles on twitter: A case study of online engagement with open access articles," *Public Understanding of Science*, vol. 28, no. 1, pp. 2–18, 2019, pMID: 29607775. [Online]. Available: https://doi.org/10.1177/0963662518761733

[100] H. T. Tu, T. T. Phan, and K. P. Nguyen, "Modeling information diffusion in social networks with ordinary linear differential equations," *Information Sciences*, vol. 593, pp. 614–636, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/ pii/S0020025522001025

[101] A. Attard and N. S. Coulson, "A thematic analysis of patient communication in parkinson's disease online support group discussion forums," *Computers in Human Behavior*, vol. 28, no. 2, pp. 500–506, 2012. [Online]. Available: https:
//www.sciencedirect.com/science/article/pii/S0747563211002391

[102] S. R, J. Pujari, V. S. Bhat, and A. Dixit, "Timeline analysis of twitter user," *Procedia Computer Science*, vol. 132, pp. 157–166, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: https:
//www.sciencedirect.com/science/article/pii/S187705091830913X

[103] C. Y. Joa and G. W. Yun, "Who sets social media sentiment?: Sentiment contagion in the 2016 u.s. presidential election media tweet network," *Journalism Practice*, vol. 16, no. 7, pp. 1449–1472, 2022. [Online]. Available: https:
//doi.org/10.1080/17512786.2020.1856708

[104] J. J. Dabrowski, J. P. de Villiers, and C. Beyers, "Na¨ıve bayes switching linear dynamical system: A model for dynamic system modelling, classification, and information fusion," *Information Fusion*, vol. 42, pp. 75–101, 2018. [Online]. Available: https:
//www.sciencedirect.com/science/article/pii/S1566253517300210

[105] L. Zhang and B. Liu, *Sentiment Analysis and Opinion Mining*. Boston, MA: Springer US, 2016, pp. 1–10. [Online]. Available: https://doi.org/10.1007/978-1-4899-7502-7907-1

[106] V. Krishnamurthy and W. Hoiles, "Chapter 21 - dynamics of information diffusion and social sensing," in *Cooperative and Graph*

Signal Processing, P. M. Djurić and C. Richard, Eds. Academic Press, 2018, pp. 525–600. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128136775000213

[107] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[108] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/ pii/S0957417411011511

[109] J. R. Quinlan, "Induction of decision trees," in *Machine Learning*, vol. 1, no. 1. Springer, 1986, pp. 81–106.

[110] P.-L. Tu and J.-Y. Chung, "A new decision-tree classification algo- rithm for machine learning," in *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92*, 1992, pp. 370–377.

[111] A. Ratnaparkhi and M. P. Marcus, "Maximum entropy models for natural language ambiguity resolution," 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:2600845

[112] Z. Liu, J. Yu, L. Gu, and X. Han, "Dynamic information diffu- sion model based on weighted information entropy," in *Computer Supported Cooperative Work and Social Computing*, Y. Sun, T. Lu,
B. Cao, H. Fan, D. Liu, B. Du, and L. Gao, Eds. Springer Nature Singapore, 2022, pp. 512–524.

[113] S. Gao, H. Pang, P. Gallinari, J. Guo, and N. Kato, "A novel embed- ding method for information diffusion prediction in social network big data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2097–2105, 2017.

[114] F. Wang, H. Wang, and K. Xu, "Diffusive logistic model towards predicting information diffusion in online social networks," in *2012*

32nd International Conference on Distributed Computing Systems Workshops, 2012, pp. 133–139.

[115] H. Wang, C. Yang, and C. Shi, "Neural information diffusion prediction with topic-aware attention network," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1899–1908. [Online]. Available: https://doi.org/10.1145/3459637.3482374

[116] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet: A popular information diffusion mechanism–a survey paper," *Online Social Networks and Media*, vol. 6, pp. 26–40, 2018.

[117] E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou, "Temporal dynamics of information diffusion in twitter: Modeling and exper- imentation," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 256–264, 2018.

[118] H.-C. Chang, "A new perspective on twitter hashtag use: Diffu- sion of innovation theory," *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–4, 2010.

[119] K. Dey, S. Kaushik, and L. V. Subramaniam, "Literature survey on interplay of topics, information diffusion and connections on social networks," *arXiv preprint arXiv:1706.00921*, 2017.

[120] T. Kameda and R. Hastie, "Herd behavior," *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, pp. 1–14, 2015.

[121] J. Mattke, C. Maier, L. Reis, and T. Weitzel, "Herd behavior in social media: The role of facebook likes, strength of ties, and expertise," *Information & Management*, vol. 57, no. 8, p. 103370, 2020.

[122] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM Sigmod Record*, vol. 42, no. 2, pp. 17–28, 2013.

[123] I. Pitas, *Graph-Based Social Media Analysis*, ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2016. [Online]. Available: https://books.google.lk/books? id=BvYYCwAAQBAJ

[124] L. Wu, P. Cui, J. Pei, and L. Zhao, *Graph Neural Networks: Foundations, Frontiers, and Applications*. Springer Singapore, 2022. [Online]. Available: https://books.google.lk/books?id=
XplXEAAAQBAJ

[125] Y. Lu, L. Yu, T. Zhang, C. Zang, P. Cui, C. Song, and W. Zhu, "Col- lective human behavior in cascading system: discovery, modeling and applications," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 297–306.

[126] B. Wang, L. Ma, and Q. He, "Idpso for influence maximization under independent cascade model," in *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 2022,
pp. 1–6.

[127] ——, "Idpso for influence maximization under independent cas- cade model," in *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 2022, pp. 1–6.

[128] W. Yang, L. Brenner, and A. Giua, "Computation of activation probabilities in the independent cascade model," in *2018 5th Inter- national Conference on Control, Decision and Information Tech- nologies (CoDIT)*, 2018, pp. 791–797.

[129] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *2010 IEEE international conference on data mining*. IEEE, 2010, pp. 88–97.

[130] G. Fibich, "Bass-sir model for diffusion of new products in social networks," *Physical Review E*, vol. 94, no. 3, p. 032305, 2016.

[131] F. Riquelme and J.-A. Vera, “A parameterizable influence spread- based centrality measure for influential users detection in social networks,” *Knowledge-Based Systems*, vol. 257, p. 109922, 2022.

[132] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[133] B. Golub and M. O. Jackson, “Na¨ıve learning in social networks and the wisdom of crowds,” *American Economic Journal: Microeco- nomics*, vol. 2, pp. 112–149, 2010.

[134] A. Goyal, W. Lu, and L. V. Lakshmanan, “Simpath: An efficient algorithm for influence maximization under the linear threshold model,” in *2011 IEEE 11th international conference on data mining*. IEEE, 2011, pp. 211–220.

[135] C. Li, J. Luo, J. Z. Huang, and J. Fan, “Multi-layer network for influence propagation over microblog,” in *Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2012, Kuala Lumpur, Malaysia, May 29, 2012. Proceedings*. Springer, 2012, pp. 60–72.

[136] M. H. Alam, W.-J. Ryu, and S. Lee, “Hashtag-based topic evolution in social media,” *World Wide Web*, vol. 20, pp. 1527–1549, 2017.

[137] S. Louvigné, M. Uto, Y. Kato, and T. Ishii, “Social constructivist approach of motivation: social media messages recommendation system,” *Behaviormetrika*, vol. 45, pp. 133–155, 2018.

[138] J. Leskovec, L. A. Adamic, and B. A. Huberman, “The dynamics of viral marketing,” *ACM Trans. Web*, vol. 1, no. 1, p. 5–es, may 2007. [Online]. Available: https://doi.org/10.1145/1232722.1232727

[139] L. Q. C. E.-H. Chang B., Xu T., “Study on information diffusion analysis in social networks and its applications,” *International Jour- nal of Automation and Computing*, vol. 15, no. 4, pp. 377–401, 2018.

[140] B. C. E. A. A. F. A.-P. S. Agrawal D., Bamieh B., "Data-driven mod- eling and analysis of online social networks," in *Web-Age Informa- tion Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 3–17.

[141] R. Toivonen, L. Kovanen, M. Kivelä, J.-P. Onnela, J. Saramäki, and K. Kaski, "A comparative study of social network models: Network evolution models and nodal attribute models," *Social Networks*, vol. 31, no. 4, pp. 240–254, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378873309000331

[142] D. Yogish, T. N. Manjunath, and R. S. Hegadi, "Review on natural language processing trends and techniques using nltk," in *Recent Trends in Image Processing and Pattern Recognition*, K. C. Santosh and R. S. Hegadi, Eds. Singapore: Springer Singapore, 2019, pp. 589–606.

[143] S. Finlay, *Text Mining and Social Network Analysis*. London: Palgrave Macmillan UK, 2014, pp. 179–193. [Online]. Available: https://doi.org/10.1057/9781137379283 9

[144] X. Li, C. Wu, and F. Mai, "The effect of online reviews on product sales: A joint sentiment-topic analysis," *Information Management*, vol. 56, no. 2, pp. 172–184, 2019, social Commerce and Social Media: Behaviors in the New Service Economy. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378720617304597

[145] J. Kim, S. Hur, E. Lee, S. Lee, and J. Kim, "Nlp-fast: A fast, scalable, and flexible system to accelerate large-scale heterogeneous nlp models," in *2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2021, pp. 75– 89.

[146] V. Kocaman and D. Talby, "Spark nlp: Natural language understanding at scale," *Software Impacts*, vol. 8, p. 100058,

2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665963821000063

[147] M. Khan and A. Malviya, “Big data approach for sentiment analy- sis of twitter data using hadoop framework and deep learning,” in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–5.

[148] M. Fowler, “What do you mean by ”event-driven”?” https://martinfowler.com/articles/201701-event-driven.html, 2021.

[149] N. Railić and M. Savić, “Architecting continuous integration and continuous deployment for microservice architecture,” in *2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2021, pp. 1–5.

[150] G. Hohpe and B. Woolf, *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. USA: Addison- Wesley Longman Publishing Co., Inc., 2003.

[151] ——, *Enterprise integration patterns: Designing, building, and de- ploying messaging solutions*. Addison-Wesley Professional, 2020.

[152] J. Woo and H. Chen, “An event-driven sir model for topic diffusion in web forums,” in *2012 IEEE International Conference on Intelligence and Security Informatics*, 2012, pp. 108–113.

[153] V. Desai, “Building an event-driven solution for social media data ingestion: A high-level architecture,” *Medium*, May 2021. [Online]. Available: https://medium.com/@vedantdesai942000/building-an-event-driven-solution-for-social-media-data-ingestion-a-high-level

[154] M. A. Shiekh, K. Sharma, and A. H. Ganai, “Information diffusion: Survey to models and approaches, a way to capture online social networks,” in *Intelligent Data Communication Technologies and Internet of Things*, D. J. Hemanth, S. Shakya, and Z. Baig, Eds. Cham: Springer International Publishing, 2020, pp. 25–32.

[155] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th interna- tional joint conference on artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[156] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[157] J. Davis and M. Goadrich, "The relationship between precision- recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[158] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2020.

[159] . W. T. Reis L., Maier C., "Mixed-methods in information systems re- search: Status quo, core concepts, and future research implications." Communications of the Association for Information Systems, 2022,
pp. 51–52.

[160] J. W. Creswell and V. L. Plano Clark, *Designing and conducting mixed methods research*. Sage publications, 2017.

[161] S. A. Kumar P., "Information diffusion modeling and analysis for socially interacting networks," *Social Network Analysis and Mining*, vol. 11, no. 1, 2021.

[162] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation – a review on the statistical data driven approaches," *European Journal of Operational Research*, vol. 213, no. 1, pp. 1–14, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221710007903

[163] D. L. Hansen, B. Shneiderman, and M. A. Smith, "Analyzing social media networks with nodexl: Insights from a connected world," *Morgan Kaufmann*, 2010.

[164] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *International AAAI Conference on Weblogs and Social Media*, 2009.

[165] A. Gruzd, B. Wellman, and Y. Takhteyev, "Imagining twitter as an imagined community," *American Behavioral Scientist*, vol. 55, no. 10, pp. 1294–1318, 2011.

[166] S. M. G. M. Kumar, S., "Modeling information diffusion in on- line social networks using a modified forest-fire model," *Journal of Intelligent Information Systems*, vol. 56, no. 2, pp. 355–377, 2021.

[167] F. Li and N. Lin, "Social network analysis of information diffusion on sina weibo micro-blog system," in *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2015, pp. 233–236.

[168] L. Birt, S. Scott, D. Cavers, C. Campbell, and F. Walter, "Member checking: A tool to enhance trustworthiness or merely a nod to validation?" *Qualitative Health Research*, vol. 26, no. 13, pp. 1802– 1811, 2016.

[169] K. Krippendorff, *Content analysis: An introduction to its methodol- ogy*. Sage publications, 2013.

[170] K. A. Neuendorf, *The content analysis guidebook*. Sage publica- tions, 2016.

[171] D. Susser, "Ethical considerations for digitally targeted public health interventions," *American Journal of Public Health*, vol. 110, no. S3, pp. S290–S291, 2020, pMID: 33001734. [Online].
Available: https://doi.org/10.2105/AJPH.2020.305758

[172] K. Rodham and J. Gavin, "The ethics of using the internet to collect qualitative research data," *Research Ethics*, vol. 2, no. 3, pp. 92–97, 2006. [Online]. Available: https://doi.org/10.1177/174701610600200303

[173] H. T. Shen, *Principal Component Analysis*. Boston, MA: Springer US, 2009, pp. 2136–2136. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9 540

[174] S. Choi, *Independent Component Analysis*. Boston, MA: Springer US, 2009, pp. 735–741. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5 305

[175] *What is Independent Component Analysis?* John Wiley Sons, Ltd, 2001, ch. 7, pp. 145–164. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221317.ch7

[176] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Commun.*, vol. 30, no. 2, p. 169–190, jan 2017. [Online]. Available: https://doi.org/10.3233/AIC-170729

[177] B. Ghojogh, M. Crowley, F. Karray, and A. Ghodsi, *Locally Linear Embedding*. Cham: Springer International Publishing, 2023, pp. 207–247. [Online]. Available: https://doi.org/10.1007/978-3-031-10602-6 8

[178] Z. Xie, W. Zhang, H. Ding, and L. Ma, "Msfcnet: Multi-scale feature-crossing attention network for multi-field sparse data," in *Advances in Knowledge Discovery and Data Mining*, H. W. Lauw, R. C.-W. Wong, A. Ntoulas, E.-P. Lim, S.-K. Ng, and S. J. Pan, Eds. Cham: Springer International Publishing, 2020, pp. 142–154.

[179] B. A. Eclarin, A. C. Fajardo, and R. P. Medina, "Enhanced hash algorithm using a two-dimensional vector to improve data search performance," in *Intelligent and Interactive Computing*, V. Piuri, V. E. Balas, S. Borah, and S. S. Syed Ahmad, Eds. Singapore: Springer Singapore, 2019, pp. 59–69.

[180] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial Intelligence Review*, vol. 56,

no. 9, pp. 10 345–10 425, Sep 2023. [Online]. Available: https://doi.org/10.1007/s10462-023-10419-1