Republic of Iraq

Ministry of Higher Education and Scientific Research

University of Babylon

College of Information Technology

Software Department

Video Popularity Prediction based on YouTube Metadata and Thumbnail Images

A Thesis

Submitted to the Council of the College of Information Technology for Postgraduate Studies of University of Babylon in Partial Fulfilment of the Requirements for the Degree of Master in Information Technology Software

by:

Heba Hussein Abd-Alabas Naif

Supervised by:

Lecturer Dr. Wadhah Razooqi Abood Hassan Baiee

2023 A.C. 1445 A.H.

أَلَمْ نَشْرَحْ لَكَ صَدْرَكَ ۞ وَوَضَعْنَا عَنكَ وِزْرَكَ ۞

الَّذِي أَنقَضَ ظَهْرَكَ ۞ وَرَفَعْنَا لَكَ ذِكْرَكَ ۞ فَإِنَّ مَعَ الْعُسْرِ يُسْرًا

۞ إِنَّ مَعَ الْعُسْرِ يُسْرًا ۞ فَإِذَا فَرَغْتَ فَانصَبْ ۞

وَإِلَىٰ رَبِّكَ فَارْغَب ۞

صدق الله العظيم

سورة الشرح

Dedications

To the Prophet of Mercy, Muhammad, may God bless him and grant him peace:

I dedicate this thesis with deep reverence and respect to the Prophet Muhammad, whose teachings of compassion, wisdom, and guidance continue to inspire countless lives around the world. His legacy serves as a beacon of light and an enduring source of spiritual strength for all of humanity.

And to the Imam, the Awaited Mahdi:

Whose profound spiritual presence offers hope and guidance to those who await his appearance.

To My Family:

You have consistently been the spark that reignited my light when it dimmed. Thank you for your unwavering love and support along this journey I have undertaken. I love you all, now and for all time.

Heba Hussien

Acknowledgements

I am deeply grateful to God for His blessings and for enabling my success in reaching this significant moment in my academic journey.

There are no proper words to convey my deep gratitude and respect for my supervisor, Dr. Wadhah Baiee His patience, guidance, valuable advice, and particularly his uplifting support and enthusiastic demeanor have been instrumental in helping me complete this thesis.

I extend my heartfelt appreciation to my family, especially my mother and aunt Maryam, for their unwavering support during my master's thesis.

I also extend my gratitude to my brothers and sisters for their endless support. It would not have been possible to write this thesis without their unwavering encouragement. Their guidance, love, and support have played an instrumental role in my development and success.

I sincerely thank my dear friend Shahad Hussien for her inspiring friendship, understanding, and unwavering belief in

my abilities. Her presence and support have proven invaluable throughout this challenging endeavor.

Furthermore, I express my heartfelt gratitude to all those who stood by me and provided moral support during this research journey. Your unwavering belief in me, kind words, and encouragement have been pivotal in keeping me motivated and focused on achieving my goals.

Declaration

I hereby declare that this Thesis, submitted to the University of Babylon in partial fulfillment of requirements for the degree of Master in Information Technology \Software, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose sources are appropriately cited in the references.

Signature:

Name: Heba Hussein Abd-Alabas

Date: / /2023

Supervisor Certification

I certify that this thesis was prepared under my supervision at the Department of Software / Collage of Information Technology / University of Babylon, by **Heba Hussien** as a partial fulfilment of the requirements for the degree of **Master in Information Technology**.

Signature:
Name: Lecturer. Dr. Wadhah Razooqi Baiee
Date: / / 2023

The Head of the Department Certification

In view of the available recommendation, we forward this Thesis for debate by the examining committee.

Signature:
Name: Prof. Ahmed Saleem Al-Saffar
Date: / / 2023

Abstract

The exponential growth of social media platforms and online video content has revolutionized the way we communicate, interact, and consume information. With millions of videos being shared daily on platforms like YouTube, predicting the popularity of these videos has become a crucial aspect for influencers, platform administrators, and marketers. Understanding the factors that contribute to a video's popularity can significantly impact content marketing and YouTube platform strategies.

This thesis delves into the challenge of accurately predicting video popularity on social media platforms, with a specific focus on YouTube platform. I propose a new combined approach that harnesses both metadata analysis (including selected impact features) and thumbnail image analysis to create a more effective prediction model. By extracting relevant features from these sources, I aim to enhance the precision of forecasting video popularity. To achieve this goal, employ a range of powerful classification algorithms, including Support Vector Machine, Gradient Boosting, Random Forest, Extreme Gradient Boosting, and K-nearest neighbour. Some preprocessing operations were conducted on the data to make it suitable for machine learning algorithms. Subsequently, features were extracted in three categories: textual features, visual features, and time-related features. These newly added features, combined with the existing ones, enriched the dataset with additional impactful attributes capable of enhancing the model's accuracy. By training the model using these algorithms and implementing feature extraction techniques, substantial accuracy was achieved, particularly with the Extreme Gradient Boosting and Random Forest algorithms. The

forecasting process was conducted in two stages: predicting both original and extracted features, and predicting only the extracted features. The model's performance was evaluated using various metrics, including accuracy, which resulted in rates of 96% and 97% for the Extreme Gradient Boosting and Random Forest algorithms, respectively. Combining metadata and thumbnail analysis provides valuable insights into the factors influencing video popularity. This enables influencers and marketers to better personalize their content, resonating with their target audience and increasing the likelihood of success on social media platforms.

Table of Contents

List Of Figures

List Of Tables

List Of Algorithms

List of Abbreviations

Abbreviation	Meaning
YOLO v5	You Only Look Once Version 5
XGB	Extreme Gradient Boosting
SMHP	Social Media Headline Prediction
RF	Random Forest
LTRCN	Long-Term Recurrent Convolutional Network
Popularity-SVR	Popularity- Support Vector Regression
MRBF	Multivariate Radial Basis Function
NLP	Natural Language Processing
AL	Artificial Intelligent
URLs	Uniform Resource Locators
CNN	Convolutional Neural Network
VGG	Visual Geometry Group
ResNet	Residual Network
IOU	Intersection over Union
NMS	Non-maximum suppression
TF-IDF	Term Frequency-Inverse Document Frequency
Word2Vec	Word-To-Vector
GloVe	Global Vectors for Word Representation
Doc2Vec	Document-To-Vector
SVM	Support Vector Machine
KNN	K-Nearest Neighbours
MSE	Mean Squared Error
MAE	Mean Absolute Error
WL	Weak Learner
GBM	Gradient Boosting Machines
API	Application Programming Interface
ID	Identification
WWW	World Wide Web
https	Hypertext Transfer Protocol Secure

AFINN	Affective Norms for English Words

Chapter One

General Introduction

1 General Introduction

1.1 Introduction

In our modern era, we've witnessed the emergence of online video-sharing platforms, which have revolutionized how we consume and engage with media content. Among these platforms, YouTube has emerged as the dominant player, with billions of users worldwide and an immense library of videos covering a vast range of topics. As YouTube continues to grow, influencers and marketers strive to understand the factors that contribute to the popularity of videos on the platform. Predicting the popularity of YouTube videos has become a subject of great interest, as it offers valuable insights for influencers, marketers, and platform administrators[1,2].

Understanding what makes a video popular on YouTube is not only intriguing from a social perspective but also carries significant practical implications. For influencers, accurately predicting video popularity can guide decisions regarding content creation, title optimization, thumbnail design, and promotional strategies[3,4]. Marketers can leverage predictive models to identify potential viral videos and allocate their advertising budgets effectively[5]. Moreover, YouTube itself can benefit from predictive analytics by enhancing user experience, optimizing recommendations, and attracting more creators to the platform.

In recent years, advancements in data science, machine learning, and natural language processing techniques have opened up exciting possibilities for predicting the popularity of YouTube videos. By analyzing various features of a video, such as view count, likes, comments, video duration, tags, and channel characteristics, predictive models can be trained to estimate the

likelihood of a video becoming popular. These models can uncover patterns and relationships that humans may overlook, enabling more accurate predictions of video performance.

Predicting the popularity of YouTube videos is a multifaceted task with several challenges[3]. The considerable volume of accessible data, the platform's dynamic nature, and the impact of external factors like trending topics and algorithmic changes all contribute to these obstacles[4]. Additionally, the subjective nature of popularity and the inherent unpredictability of viral phenomena add further complexity to the prediction process[5].

This thesis encompasses a thorough analysis of various features that contribute to a video's popularity. Specifically, we will investigate the impact of different video metadata, such as the ratio of likes to dislikes and the number of comments, as well as examine the role of video title, description, and thumbnails, among other factors. By carefully analyzing these features, we aim to uncover the underlying patterns and relationships that drive video popularity on YouTube. This analysis will provide valuable insights into the aspects that captivate viewers and influence their engagement with videos.

Moreover, we will evaluate the performance of different predictive models in the realm of video popularity prediction. Through rigorous testing and comparison, we will assess the effectiveness of these models in accurately estimating the likelihood of a video becoming popular. This evaluation process will enable us to identify the most reliable and accurate predictive models for video popularity.

By integrating our exploratory analysis of video characteristics with the evaluation of predictive models, we aim to make a valuable contribution to

the continually expanding domain of social media analytics and predictive modeling. Our findings will not only enhance our understanding of YouTube's popularity dynamics but also provide actionable guidance for Influencers, marketers, and YouTube administrators in optimizing their video strategies and maximizing audience reach. Ultimately, this research endeavor holds the potential to advance our knowledge of online content consumption patterns and shape the future of video analytics and prediction techniques.

1.2 Thesis Problem

Predict the popularity of YouTube videos, which is crucial for influencers, content creators, and marketing companies. This involves understanding the multifaceted factors that contribute to a video's success and determining how effectively machine learning algorithms and predictive models can analyze metadata and user engagement metrics for precise popularity predictions. Furthermore, we need to explore how leveraging this predictive capability can lead to enhanced marketing and content creation strategies for YouTube creators and businesses.

1.3 Thesis Question

- Can a predictive system be developed to accurately forecast the popularity of YouTube videos?
- What additional features can be incorporated to enhance the richness of the dataset for improved prediction accuracy?
- How can overall classification accuracy be enhanced in the context of video popularity prediction?

1.4 Aim of Thesis

- The primary aim of this thesis is to develop effective prediction models to empower influencers, marketers, and platform administrators in maximizing video popularity across various social platforms like YouTube.
- Extracting additional features that could influence the prediction process to improve model accuracy.
- Enhance model accuracy through the various stages to prepare the dataset for learning algorithms.

1.5 Thesis Contribution

- The thesis introduces the use of sentiment analysis to extract subjective information from textual features as its contribution.
- Additionally, the thesis proposes a pre-training model (YOLO v5) for extracting visual features from video thumbnails. The integration of these visual and textual features aims to improve prediction rates and provide a more comprehensive understanding of their combined impact on video popularity, thus enhancing the model's performance and accuracy in predicting video trends.

1.6 Related Works

This section reviews various previous research that has addressed the issue of predicting the popularity of specific content on multiple platforms. These studies have explored different features, factors, and algorithms to achieve their predictions.

Trzcinski et al [6]. proposed a model to predict the popularity of an online video before its content is published, using Support Vector Regression with Gaussian Radial Basis Functions. They showed that predicting popularity patterns with this approach provides more precise and stable results. Furthermore, they demonstrated that combining early distribution patterns with social and visual features improves the accuracy of popularity prediction. In terms of video popularity prediction, social features were found to be a much stronger signal than visual features. The best results were achieved by combining visual features, social features, and early view counts, allowing for the prediction of video popularity on Facebook with a Spearman correlation rank of up to 0.94, just 6 hours after publication.

Y. Li et al.[1] proposed the use of several machine learning algorithms to predict performance, and backward search is employed to select the most relevant features. As a result, extreme gradient boosting with three features (time gap, category, description) is chosen due to its optimal balance between cost and performance, resulting in an F-score of 0.73.

M. U. N. Nisa et al. [7]proposed a method that predicts the popularity of videos using the XGBoost model. The approach involves features selection, fusion, min-max normalization, and precision parameters such as gamma, eta, and learning rate. The XGBoost model achieved an accuracy of 86% and a precision of 64%.

R. Shreyas et al. The Random Forest regression model is used in this paper [8]to predict the popularity of articles using the Online News Popularity data set. The Random Forest model's performance is evaluated and compared to that of other models. Standardization, regularization, correlation, strong bias/high variance, and feature selection all have an effect on learning

models. The Random Forest technique predicts popular/unpopular articles with an accuracy of 88.8%, according to the results.

F. Huang et al.[9]Introduce a thriving application scenario called Social Media Headline Prediction (SMHP), which focuses on predicting the popularity of posts shared on social media. The research proposes a method that utilizes multi-aspect features combined with the random forest (RF) model for popularity predictions. It explains the process of feature extraction by combining metadata of the posts and users' features, as well as strategies for dealing with missing values. The result of this paper indicates that user-related features, such as the number of followers and following, along with the random forest regression model, are the most effective features and model for the current social media headline prediction task. These features and the chosen model have demonstrated strong predictive capabilities in accurately forecasting the popularity of posts on social media platforms.

T. Trzcinski et al.[10]The researchers propose a new method based on a Long-term Recurrent Convolutional Network (LRCN) to address the challenge of predicting the popularity of online videos shared on social media. This approach utilizes deep neural network architectures that consider the sequential nature of information in the videos. The popularity prediction problem is formulated as a classification task, with the goal of predicting popularity using only visual cues extracted from the videos. The results of this study demonstrate that their proposed LRCN-based approach outperforms traditional shallow methods, achieving over a 30% improvement in prediction performance. The experiments are conducted on a dataset comprising more than 37,000 videos published on Facebook.

N. Sangwan and V. Bhatnagar in The research [11] focuses on predicting the prevalence of video content uploaded online by users. It uses visual and time-based highlights of videos as features for popularity prediction. Among the regression models tested, MRBF performs best, and Popularity-SVR shows stable results at around 90% accuracy. To overcome challenges in predicting video popularity accurately, fuzzy logic is proposed. The research suggests exploring more visual, time-based, and semantic features to improve the accuracy of video fame forecasting.

1.7 Thesis Outline

After Chapter one, which presents a general introduction the rest of the thesis is structured as follows:

- Chapter Two (Theoretical Background): In this chapter, we offer a comprehensive overview of the fundamental concepts and theoretical underpinnings that form the basis of the research in this thesis. Specifically, we focus on predicting video popularity on social media platforms. The topics covered include social media and predicting content popularity, factors affecting video popularity, preprocessing techniques, feature extraction, machine learning algorithms for prediction, and performance evaluation metrics. Understanding these concepts will enable readers to recognize the importance of the proposed system and its potential impact on enhancing content strategies, increasing user engagement, and improving video visibility on social media.
- Chapter Three (The Proposed System): In this chapter, we present the practical aspects of the proposed system, focusing on the algorithms and techniques used in developing the advanced prediction model for video popularity on social networks. We outline the data collection process, the

features considered, and the machine learning methodology employed to build the predictive model.

- Chapter Four (Results and Discussions): his chapter showcases the results obtained from the implementation of the proposed system. The results will be presented using tables, graphs, and visualizations to offer a comprehensive overview of the predictive model's performance. The primary findings and insights derived from the study will be discussed in detail.
- Chapter Five (Conclusions and Future Works): The final chapter will present a comprehensive assessment of the fundamental concerns addressed in this thesis and the contributions made by the proposed system. We will summarize the main findings and discuss their implications for influencers, marketers, and platform administrators.

Chapter two

Theoretical Background

2 Theoretical Background

2.1 Overview

This chapter provides a comprehensive overview of the fundamental principles underlying social media platforms, datasets, data mining, and natural language processing (NLP). It also covers the techniques involved in preprocessing, text analysis, feature extraction, and prediction algorithms. The primary emphasis of this chapter revolves around the methodologies and strategies employed in this thesis, shedding light on their significance.

2.2 Social Media and Popularity Predication

Social media refers to online platforms that enable users to create, share, and exchange information within virtual communities. It has revolutionized global communication and interaction, offering features for sharing text, photos, videos, and links. Social media's origins can be traced back to the development of computer networks and the internet, with the launch of YouTube in 2005 marking a significant milestone[12]. Since then, influential platforms like Facebook, Twitter, YouTube, LinkedIn, Instagram, and Snapchat have emerged. Social media platforms have not only changed the way people communicate but have also impacted various aspects of society. They have enabled individuals, businesses, organizations, and even governments to engage with audiences, promote products and services, share news and information, and foster communities based on shared interests and values[2]. TikTok, for instance, specializes in video editing and sharing, allowing users to create and share short video clips that can be charming, funny, or even cringe-inducing[13].

The widespread use of smartphones and mobile apps has further fueled its growth[14]. While providing opportunities for self-expression and networking, social media also poses challenges concerning privacy, security, and excessive screen time[13]. Overall, it has transformed the way people connect, communicate, and shape various aspects of personal and professional life.

2.2.1 Predicting Content Popularity

Predicting popularity in social media is a common application of data mining and analytics. With the enormous amount of user-generated content and interactions on platforms such as Facebook, Twitter, Instagram, and YouTube, businesses and researchers are interested in understanding which posts, videos, or content will become popular and gain significant engagement[15]. Data mining techniques can be employed to analyse various factors that contribute to popularity in social media. These factors may include text features likes (word embeddings from video descriptions , titles) and visual and metadata features [15]. By examining historical data and patterns, data mining models can be built to predict the potential popularity of web content. Machine learning algorithms can be applied to social media data to uncover patterns and relationships that influence popularity. These algorithms can consider various features, such as textual features ,video descriptions and titles, the source of the content, category, number of views, timing of the post to make predictions about the potential popularity of a post or content[15]. Additionally, sentiment analysis can be used to understand the sentiment or emotional tone expressed in social media posts. Analysing sentiment can provide insights into the factors that contribute to content popularity. For example, positive sentiment in user comments or interactions may indicate a higher likelihood of popularity[16]. By predicting popularity in social media, businesses and influencers can optimize their strategies, identify trends, and improve their chances of reaching

a larger audience. However, it's important to note that predicting popularity in social media is a complex task and can be influenced by various factors, including user preferences, viral trends, and unpredictable events.

2.2.2 Factors Affecting Video Popularity

The factors that affect video popularity can be categorized into two main groups: external factors related to the platform and users, and internal factors related to the video itself [17][4] it is shown in figure 2.1. Here's a breakdown of each category:

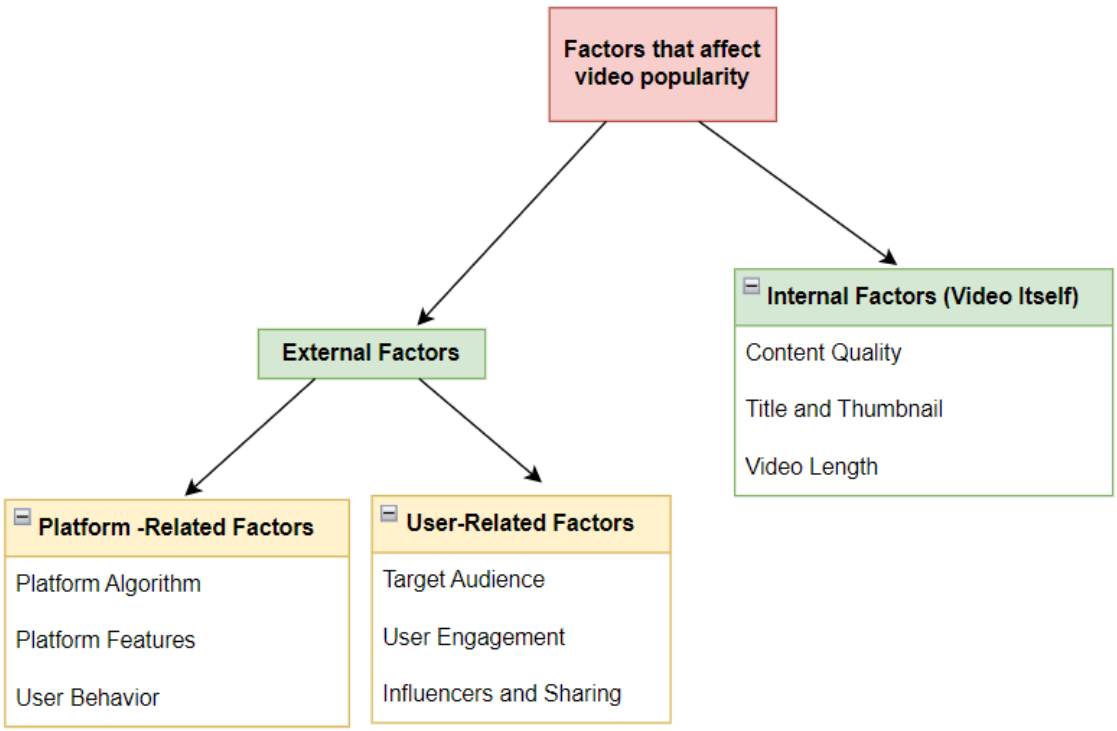


Figure 2.1: Some factor that effect video popularity.

2.2.2.1 External factors

They encompass elements outside of the video that can impact its popularity, including platform-related and user-related factors.

- Platform-Related Factors

  Platform Algorithm: Each platform has its own algorithm that determines which videos get recommended and shown to a broader audience. Understanding and optimizing for these algorithms can significantly impact video visibility[18].

  Platform Features: Utilizing platform-specific features, such as live streaming, stories, or interactive elements, can enhance engagement and popularity[19].

  User Behaviour: User behaviour, such as viewing habits, likes, comments, and shares, can influence the visibility and recommendation of a video.

- User-Related Factors[17]

  Target Audience: Understanding the preferences and interests of the target audience is crucial for creating content that resonates with them.

  User Engagement: Higher engagement metrics like likes, comments, and shares signal to the platform that the video is valuable, potentially leading to increased visibility[2,19].

  Influencers and Sharing: When influential users or influencers share or feature a video, it can lead to a significant boost in popularity[2,20, 5].

2.2.2.2 Internal Factors (Video Itself)

- Internal factors refer to the characteristics or attributes of the video itself that influence its popularity. These factors include:

Content Quality and Type: High-quality content that is informative, entertaining, or valuable to the audience is more likely to be shared and recommended[3,20].

Challenges and Trends: Participating in popular challenges or following internet trends can help videos gain visibility and attract a wider audience[20].

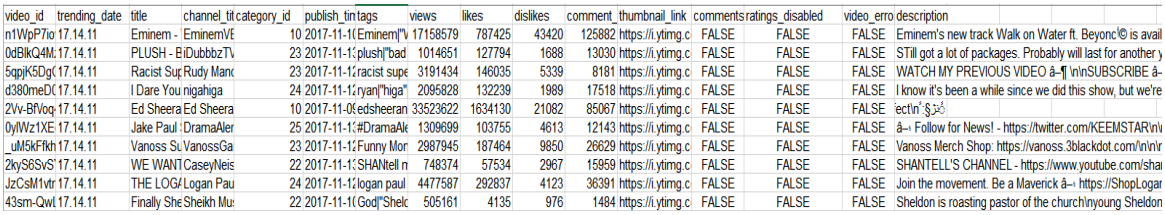
Title and Thumbnail: An attention-grabbing title and thumbnail can encourage users to click on the video, increasing its views[5].

Video Length: The length of the video is essential. It should be neither too long nor too short, as viewers might lose interest in lengthy videos, while short videos might not provide enough value[5].

It's important to note that the impact of each factor can vary based on the specific platform and content type. Creators should analyse their audience's preferences, experiment with different strategies, and stay adaptable to achieve long-term popularity and growth.

2.3 Dataset

In this thesis, you utilized the "Trending YouTube Video Statistics" dataset obtained from Kaggle[22]. This dataset consists of information on 24,427 unique trending YouTube videos. The dataset includes various details such as the video id, video title, channel title, publish time, trending date, tags, views, likes, dislikes, description, thumbnail link, and comment count, explain in table 2.1. By utilizing this dataset, we can explore and analyse the factors that contribute to the popularity and engagement of trending YouTube videos. This can involve applying various data mining and machine learning techniques to uncover patterns, relationships, and insights within the data. (See Figure 2.2 for a sample of the image dataset).



video_id	trending_date	title	channel_tit	category_id	publish_tin	tags	views	likes	dislikes	comment_	thumbnail_link	comments	ratings_disabled	video_erro	description
n1WpP7io	17.14.11	Eminem -	EminemVE	10	2017-11-1(	Eminem|"\	17158579	787425	43420	125882	https://i.ytimg.c	FALSE	FALSE	FALSE	Eminem's new track Walk on Water ft. Beyonc© is avail
0dBlkQ4M	17.14.11	PLUSH - Bi	DubbbzTV	23	2017-11-1(	plush|"bad	1014651	127794	1688	13030	https://i.ytimg.c	FALSE	FALSE	FALSE	STill got a lot of packages. Probably will last for another y
5qpjK5Dg	17.14.11	Racist Sup	Rudy Manc	23	2017-11-1(	racist supe	3191434	146035	5339	8181	https://i.ytimg.c	FALSE	FALSE	FALSE	WATCH MY PREVIOUS VIDEO â–¶ \n\nSUBSCRIBE â–
d380meD(	17.14.11	I Dare You	nigahiga	24	2017-11-1(	ryan|"higa"	2095828	132239	1989	17518	https://i.ytimg.c	FALSE	FALSE	FALSE	I know it's been a while since we did this show, but we're
2Vv-BfVoq	17.14.11	Ed Sheera	Ed Sheera	10	2017-11-0(	edsheeran	33523622	1634130	21082	85067	https://i.ytimg.c	FALSE	FALSE	FALSE	ect\n˜§ـ؟
0ylWz1XE	17.14.11	Jake Paul :	DramaAler	25	2017-11-1(	#DramaAle	1309699	103755	4613	12143	https://i.ytimg.c	FALSE	FALSE	FALSE	â–¬ Follow for News! - https://twitter.com/KEEMSTAR\n\n
_uM5kFfkh	17.14.11	Vanoss Su	VanossGa	23	2017-11-1(	Funny Mor	2987945	187464	9850	26629	https://i.ytimg.c	FALSE	FALSE	FALSE	Vanoss Merch Shop: https://vanoss.3blackdot.com/\n\n\n
2kyS6SvS	17.14.11	WE WANT	CaseyNeis	22	2017-11-1(	SHANtell n	748374	57534	2967	15959	https://i.ytimg.c	FALSE	FALSE	FALSE	SHANTELL'S CHANNEL - https://www.youtube.com/shan
JzCsM1vtr	17.14.11	THE LOGA	Logan Pau	24	2017-11-1(	logan paul	4477587	292837	4123	36391	https://i.ytimg.c	FALSE	FALSE	FALSE	Join the movement. Be a Maverick â–¬ https://ShopLogar
43sm-QwL	17.14.11	Finally She	Sheikh Mu	22	2017-11-1(	God|"Shelc	505161	4135	976	1484	https://i.ytimg.c	FALSE	FALSE	FALSE	Sheldon is roasting pastor of the church\nyoung Sheldon

Figure 2.2: sample of dataset.

Table 2.1 Illustration of the dataset features.

Feature	Description
Video id	refers to a unique identifier assigned to each individual video on the platform
Video Title	The title or name of the video.
Channel Title	The title or name of the YouTube channel that uploaded the video.
Publish Time	The date and time when the video were published on YouTube.
Trending date	refers to the date on which a particular video appeared on the list of trending videos
Tags	Keywords or tags associated with the video, which can provide additional information or context.
Views	The number of times the video has been viewed.
Likes	The number of likes received by the video.
Dis Likes	The number of dislikes received by the video.
Description	A text description or summary of the video provided by the uploader.
Comment Count	The number of comments posted on the video.
Category ID	A field indicating the category or genre to which the video belongs. This ID can vary depending on the region or classification system used.

2.3.1 Dataset Labelling

Data labelling is the process of assigning labels or tags to data points in a dataset. It is typically performed by humans or automated systems to categorize the data according to predefined criteria. The labelled data is then used to train machine learning models, enabling them to learn and make predictions on new, unlabelled data. The classification was based on the equation found in the papers[1].

$$score = (likes - 1.5 * dislikes) * comment\_count / views$$

- Class 0 denotes videos that are not popular, with less than 100,000 views, while the other classes refer to popular videos.
- Class 1 videos are those with a score of less than 0 and have a significant number of unfavourable views, indicating that the likes and dislikes for the videos are approximately equal.
- Class 2 videos cannot be classified as either receiving positive or negative feedback, such as news stories that close their comment section. However, if the score is higher than 300, and the video has a high ratio of comments to views or a large number of favourites, it falls under this class.
- Class 3 includes videos with overwhelmingly positive reviews. In either scenario, the video may have received a lot of positive feedback from viewers.

2.4 Feature Extraction Technical

Feature extraction also known as feature engineering refers to the process of creating new features or variables from existing data to enhance the performance of a machine learning model. It involves transforming raw data

into a format that captures relevant patterns or relationships, making it easier for the model to learn and make accurate predictions[23],[24].

There are several techniques and strategies for feature generation, depending on the nature of the data and the specific problem at hand. Feature extraction is particularly useful for unstructured or high-dimensional data types such as images, audio, text, or sensor readings. Here are some common techniques for feature extraction:

2.4.1 Visual feature extraction

Visual feature extraction is a crucial step in working with image data. Techniques for feature creation involve extracting meaningful visual attributes from images. In the context of image classification and object recognition[25] , raw data often comprises pixel values that may not be directly compatible with certain classification algorithms. By extracting higher-level features, such as edges and regions correlated with the presence of human faces, the data can be transformed into a more suitable representation for a wider range of classification techniques[23]. These techniques encompass various approaches, including the utilization of Convolutional Neural Networks (CNN) to extract deep features, or the application of pre-trained models like VGG, ResNet, YOLO[26],Alex Net[27].These methods enhance the ability to capture meaningful information from images, facilitating accurate classification and object recognition tasks.

2.4.1.1 Yolo For Object Detection

(YOLO) is a powerful real-time object detection method for deep learning and computer vision. YOLO is a technique for detecting and locating objects within image frames[28]. It treats object detection as a regression issue and calculates the class probabilities for each bounding box[29]. The base model detects

images at an astounding 45 frames per second, while the Fast YOLO model detects at 155 frames per second, figure 2.3 shown the yolo architecture. It is well-known for its precision and speed, and it is frequently employed in a variety of applications[28]. The main advantage of the YOLO model is its speed and accuracy. It can rapidly detect multiple objects in a single pass through the neural network, making it particularly well-suited for real-time object detection applications.

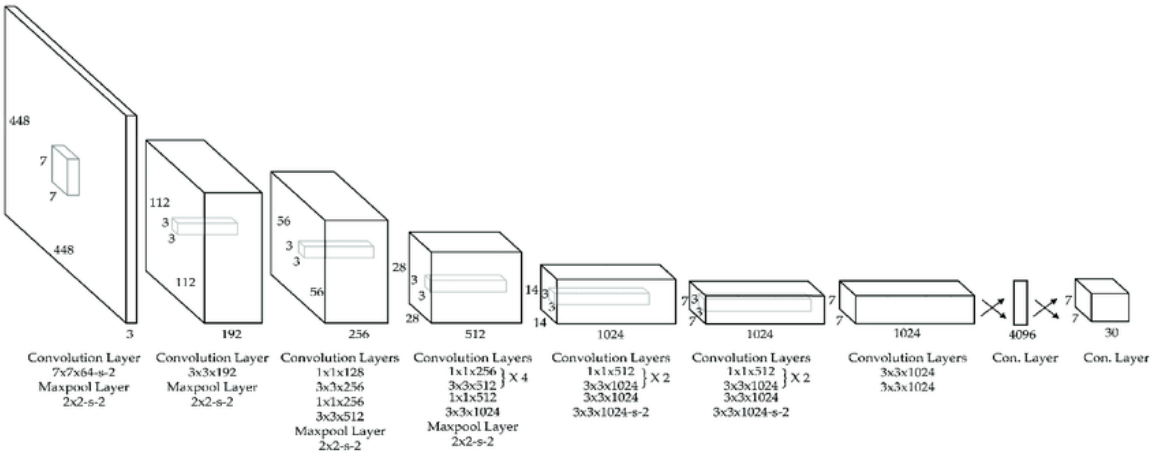


Figure 2.3: The YOLO architecture.

YOLO works dividing the input image into a grid with fixed size (S*S) and where N = 7 shown on the figure 2.4 for each grid cell, it is predicting bounding boxes (rectangles) that might contain objects. These bounding boxes are described by their coordinates (center, width, and height) relative to the grid cell. YOLO determines the attributes of these bounding boxes, where Y is the final vector representation for each bounding box[30], main steps of YOLO model in(Algorithm 2.1)[31].

$Y = [P_r, b_x, b_y, b_h, b_w, c_i]$

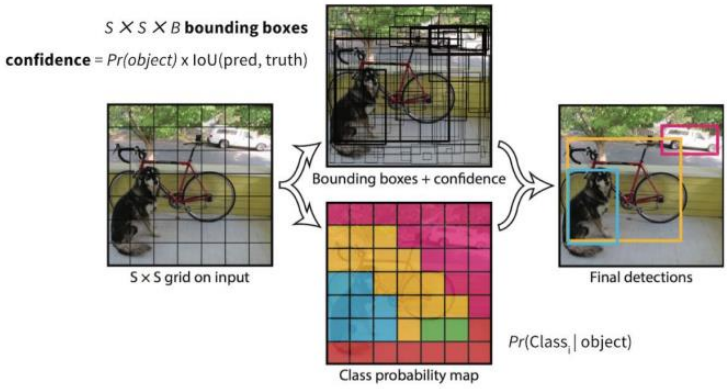


Figure 2.4: Yolo for object detection.

This is especially important during the training phase of the model.

P_r: corresponds to the probability score of the grid containing an object.

b_x,b_y : are the x and y coordinates of the center of the bounding box with respect to the enveloping grid cell.

b_h,b_w : correspond to the height and the width of the bounding box with respect to the enveloping grid cell.

c_i: correspond to the classes (number of classes).

Algorithm 2.1 Pre-trained YOLO For object detection
Input: An image.
Output: Bounding boxes that represent detected objects along with their class probabilities.
Begin: -YOLO divides an input image into an S × S grid cell. -For each grid cell: Calculate confidence scores for each bounding box. These scores represent the model's confidence that the box contains an object and how accurate it thinks the predicted box is.

These scores represent the model's confidence that the box contains an object and how accurate it thinks the predicted box is.
-During training, YOLO assign one bounding box predictor to be "responsive" for predicting an object based on which prediction has the highest Intersection over Union (IOU) with the ground truth
-Non-maximum suppression: NMS is a post-processing step that is used to improve the accuracy and efficiency of object detection.
NMS is used to identify and remove redundant or incorrect bounding boxes and to output a single bounding box for each object in the image.
End

2.5 Data Preparation Methods

Preprocessing in data mining refers to the various techniques and procedures used to clean, transform, and prepare raw data before it can be analysed[23]. The quality of the output from data mining algorithms largely depends on the quality of the input data. Therefore, preprocessing is essential to ensure that the data is in a suitable format for the analysis. Overall, the goal of preprocessing in data mining is to ensure that the data is ready for analysis and that the analysis results are accurate, reliable, sand meaningful[32][33].

In this thesis, several steps were undertaken to prepare the research data for the prediction model. Several important steps included:

2.5.1 Removal of Row Duplications

This step involves identifying and removing any duplicate rows in the dataset. Duplicates may occur due to various reasons such as data entry

errors or system issues. Removing duplicates ensures that each data point is unique and avoids any bias or inconsistencies in the analysis[23].

2.5.2 Handling Missing Values

Dealing with missing data is a crucial part of data preprocessing. Choosing the right strategy for handling missing data depends on factors like data extent, nature, analysis goals, and the techniques used. It is important to think each strategy's pros and cons to select the best fit for your dataset and analysis[34]. There are several strategies available for handling missing values[23]:

Removing Data Objects: If a significant portion of the data objects have missing values, one option is to eliminate those objects entirely.

Removing Attributes: Similarly, the choice is to remove attributes with a high number of missing values.

Estimating Missing Values: In certain cases, missing values can be estimated or imputed using various techniques. Common methods include mean imputation, regression imputation, or using specialized machine learning algorithms for imputation[35][36].

Ignoring Missing Values: it may be possible to ignore missing values during the analysis[35]. Some algorithms can handle missing data by default, or they can be modified to accommodate missing values without imputing them.

2.5.3 Processing and Extracting Text Features

Cleaning textual features involves preprocessing steps to transform raw text data into a cleaner and more suitable format for analysis or modeling. Here are some common techniques used for cleaning textual features[37] 36]:

Lowercasing Text: Converting all text to lowercase helps to standardize the text and minimize case sensitivity difficulties[39].

Removing of Punctuation: Getting rid of punctuation markings like commas, periods, and exclamation points might assist minimize noise in the text and improve future analysis.

Handling Special Characters and Numbers: special characters, URLs, or numbers that may not have significant value can be eliminated or substituted with appropriate placeholders[39].

In natural language processing (NLP), text feature creation involves extracting meaningful information from text data using various techniques such as bag-of-words, term frequency-inverse document frequency (TF-IDF), word embeddings (e.g., Word2Vec or GloVe), sentiment analysis, or document embeddings (e.g., Doc2Vec). These techniques capture the semantic and contextual information present in the text[40].

Natural Language Processing (NLP) and sentiment analysis are closely related fields that both deal with the processing and analysis of human language. NLP is a multidisciplinary field that combines linguistics, computer science, and artificial intelligence to enable computers to understand, interpret, and generate human language[41].Its primary goal is to bridge the gap between human language and machine language[42].

NLP techniques is a fairly generic term that covers a very wide range of applications[43], Figure 2.5 show are the most popular applications, allowing computers to perform tasks such as language translation, sentiment analysis, text summarization, question answering, part-of-speech

tagging, named entity recognition, syntactic parsing, and machine translation[16].

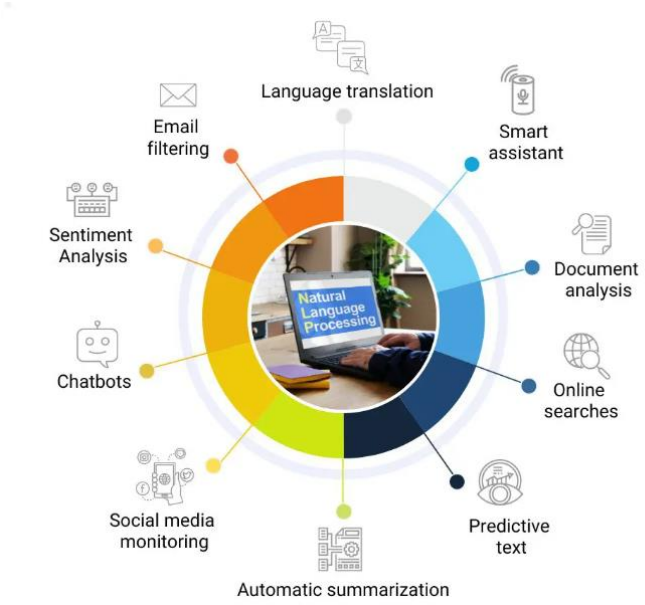


Figure 2.5: Application of NLP.

To accomplish these tasks, NLP algorithms rely on a variety of techniques, including statistical modeling, machine learning, deep learning, and rule-based approaches. Among these applications, sentiment analysis, also known as opinion mining, has gained significant attention in recent years due to the increasing use of social media platforms. Sentiment analysis employs machine learning, data mining, natural language processing, and computational linguistics techniques to identify, extract, and analyze opinions and sentiments present in textual data[16]. Sentiment analysis is a Lexicon-based approach. its aim of to determine whether a given piece of text expresses a positive, negative, or neutral sentiment, figure 2.6 show example Process of lexicon-based sentiment analysis[44].

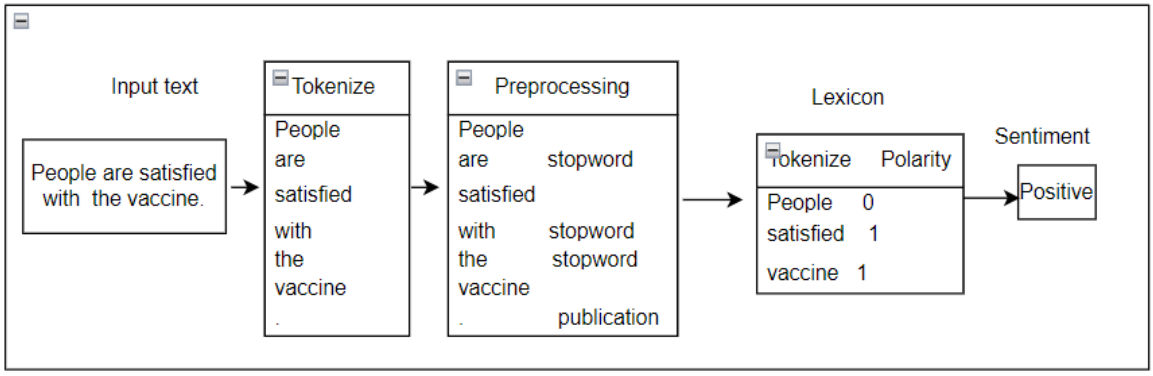


Figure 2.6: Example Process of lexicon-based sentiment analysis.

Instead of training a model on a labelled dataset, this approach relies on predefined sentiment lexicons or dictionaries containing words and their associated sentiment scores (positive, negative, or neutral), Algorithm 2.2 represent main steps in Lexicon-based approach.

Algorithm 2.2 Lexicon-based sentiment analysis
Input: text
Output: positive, negative, natural
Begin: - Input text is usually tokenized into individual words or phrases (tokens). -For each token, it searches for the corresponding word in the sentiment dictionary. - The Sentiment Dictionary associates each word or phrase with a range of polarity scores (from -1 to 1), with 0 indicating neutral sentiment. -Aggregating the individual polarity scores of all tokens in the text to calculate an overall polarity score for the entire text. This aggregation may involve simple averaging or other techniques depending on the specific implementation.

-The final polarity score is often normalized to ensure that it falls within the range of -1 to 1, even if the vocabulary contains scores on a different scale.

- Classify the overall sentiment score into predefined categories such as positive, negative, or neutral.

-Use predefined thresholds or rules for classification, e.g., score surpasses a threshold for positive classification, falls below for negative classification.

End

2.5.4 Encoding Features

It is an important aspect of data preprocessing, especially when working with machine learning algorithms that require numerical inputs[23]. There are several methods to handle non numerical features: Label Encoding, One-hot encoding, Target Encoding, Binary Encoding.

2.6 Data Mining

Data mining mean extracting useful information from massive amounts of data[23] .It is a multidisciplinary field that combines elements of statistics, artificial intelligence (AI), and database research. It has become increasingly important due to the exponential growth in the size of datasets. By collecting and analysing large amounts of data, data mining aims to discover patterns and insights that may not be immediately apparent. Data mining has two main goals, shows in figure 2.7: prediction and description[45].

Prediction: Data mining aims to develop predictive models that can forecast future outcomes or behaviours based on historical data[23]. This is valuable for

various applications such as sales forecasting, customer churn prediction, demand forecasting, and stock market prediction[46].

Description: Data mining also focuses on providing descriptive insights by summarizing and understanding the characteristics of the data. This involves identifying key attributes, summarizing statistical measures, visualizing data distributions, and generating reports or dashboards[23].

Both prediction and description are crucial in extracting meaningful insights and actionable knowledge from data, enabling informed decision-making and gaining a competitive advantage in different domains[47][8].

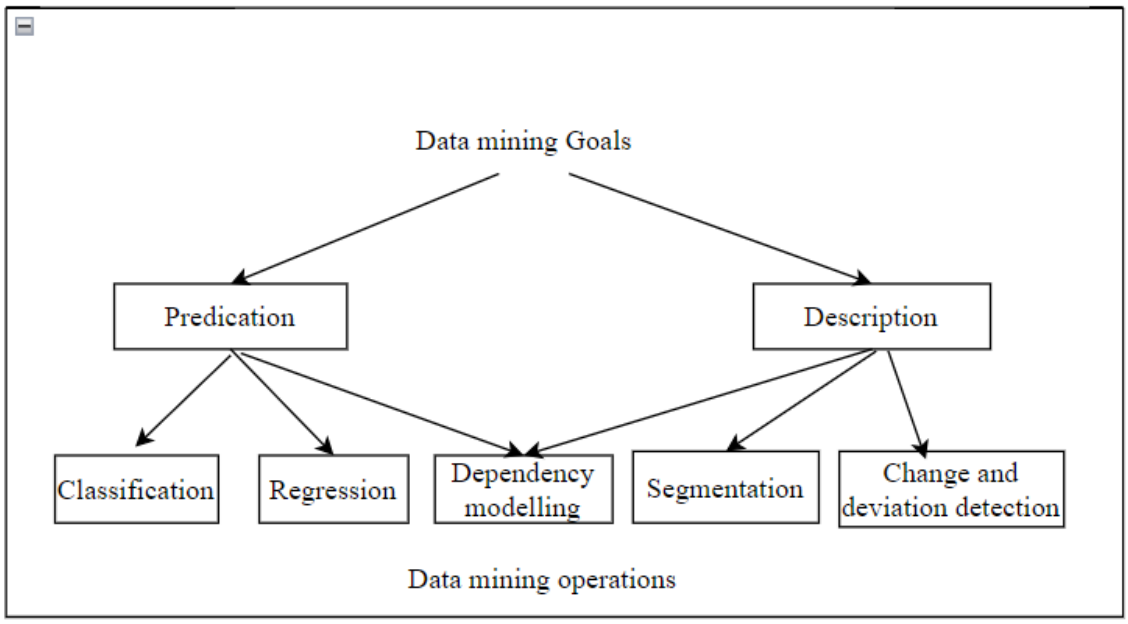


Figure 2.7: Data mining goals.

2.6.1 Machine Learning Techniques

Machine learning refers to a variety of techniques and approaches that allow computers and systems to learn from data and make predictions[41]. Here are a few examples of common machine learning techniques:

1-Supervised Learning: The set of models trained using a labeled dataset.

2-Unsupervised Learning: The set of models trained using an unlabeled dataset

Classification techniques are methods used in machine learning and statistics to classify data into predefined classes or categories. These techniques are supervised learning algorithms that learn from labelled training data to make predictions or assign class labels to new, unseen data[23]. Here are some used classification techniques in this thesis:

2.6.1.1 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm that is versatile and can be applied to regression and classification tasks[48]. The basic idea of SVM (Support Vector Machines) is to find an optimal hyperplane that separates data points of different classes in a high-dimensional space, shows in figure 2.8. SVM can handle linear and non-linear problems and is commonly used for classification tasks. Here's types of SVMs and their corresponding equations[23][49]:

In linear SVM, the goal is to find a linear decision boundary. The equation of the hyperplane can be represented as[50]:

$$w^t \cdot x + b = 0 \qquad (1)$$

w^t: represents the weight vector perpendicular to the hyperplane

x: represents the input data vector

b: is the bias term.

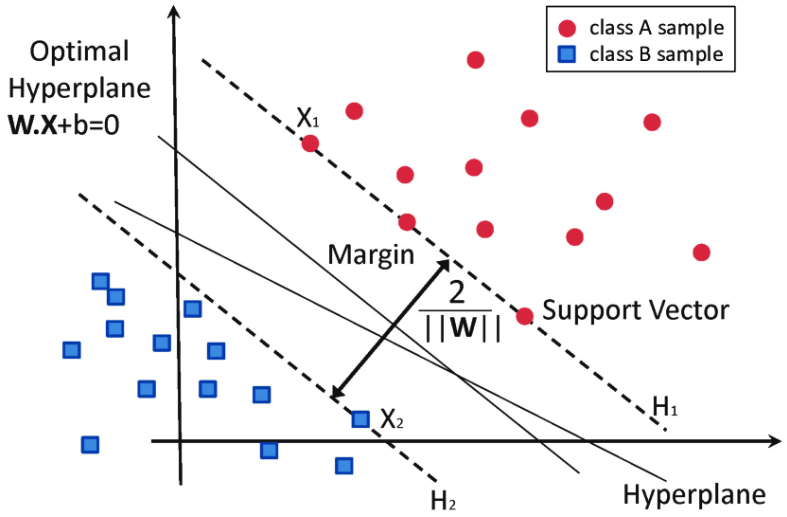


Figure 2.8: Classification of data by support vector machine (SVM).

Non-linear SVM, also known as kernel SVM, is used to handle non-linearly separable data by mapping it to a higher-dimensional feature space using a kernel function. The decision boundary equation becomes:

$$\sum(\alpha_i \cdot y_i \cdot k(x_i, x)) + b = 0 \qquad (2)$$

where α_i and y_i are the Lagrange multipliers and class labels of the support vectors, x_i represents the support vectors, $k(x_i, x)$is the kernel function that computes the similarity between the support vectors and the input data point x. Commonly used kernel functions include: Gaussian (Radial Basis Function), Polynomial, Sigmoid and Linear Kernel. The choice of kernel function depends on the problem and the underlying data. These equations represent the decision boundaries for different types of SVMs. The goal of SVM is to find the optimal values for the parameters (weights, bias, Lagrange multipliers) that maximize the margin between the classes or minimize classification errors, depending on the problem setup.

2.6.1.2 K-Nearest Neighbor Classifier

The K-Nearest Neighbor (KNN) Classifier is a widely employed technique in machine learning for both classification and regression tasks. It is a non-parametric and instance-based algorithm, meaning it does not assume any particular data distribution and relies on the actual training instances for predictions. The fundamental concept of KNN involves classifying a new data point based on the majority class among its K nearest neighbors [28]. To determine the neighbors, KNN calculates the distance between the new data point and each training data point[23]. Common distance metrics used include Euclidean distance and Manhattan distance. Euclidean distance, for example, calculates the distance between two points (x1, y1) and (x2, y2) in a 2D space as follows:

$$\text{Distance} = sqrt((x_2 - x_1)^2 + (y_2 - y_1)^2) \quad (3)$$

Once the distances are calculated, the subsequent stage involves determining the K nearest neighbors of the new data point. The value of K, representing the number of neighbors to consider, is specified by the user. The KNN algorithm then employs a voting mechanism to assign a class label to the new data point. In classification scenarios, the predicted class is determined by selecting the majority class among the K neighbors. When K equals 1, the class label of the nearest neighbor is directly assigned to the new data point. For regression tasks, the predicted value is obtained as the average or weighted average of the values associated with the K nearest neighbors. In the k-nearest neighbors (KNN) algorithm, the choice of the right value for the parameter k is crucial. The value of k determines the number of nearest neighbors that will be considered when making predictions for a new or test instance. Choosing the right value for k in K-Nearest Neighbors (KNN) is important. If k is too small,

the classifier may overfit the training data and be sensitive to noise. If k is too large, the classifier may misclassify test instances by including distant neighbors. It's crucial to find the optimal k value to balance between overfitting and misclassification, ensuring accurate predictions (see Figure 2.9).

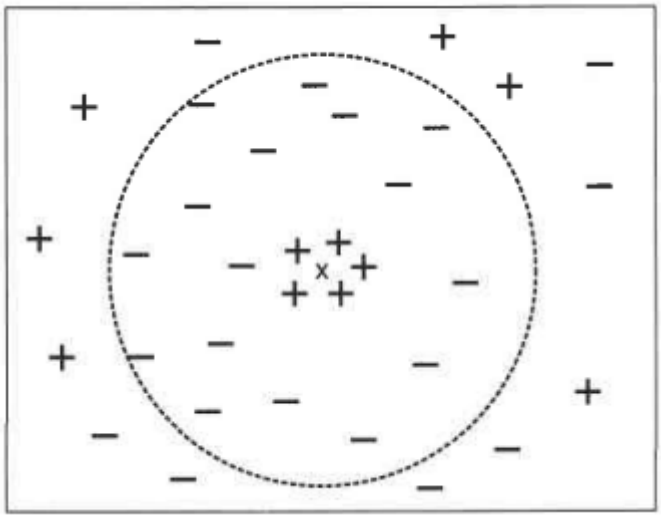


Figure 2.9: KNN classification with large k.

2.6.1.3 Random Forest Classifier

Random Forest is an ensemble method that combines multiple decision trees to make predictions. It creates a forest of decision trees rather than relying on a single decision tree to improve prediction accuracy. Each decision tree in the Random Forest is trained independently on a random subset of the training data. This random subset is typically selected through bootstrapping, where data points are sampled with replacement[23]. This process introduces diversity among the individual trees. During prediction, the algorithm aggregates the predictions of all the individual decision trees to make the final prediction. The way this aggregation is done depends on the type of task (classification or regression). In classification tasks, the predictions of individual decision trees are combined using majority voting. The class that

receives the most votes across the decision trees becomes the final prediction. In regression tasks, the predictions of individual decision trees are often averaged or aggregated to obtain the final prediction[51]. The main objective of using a Random Forest is to reduce variance and bias, thus improving the generalization ability of the model. Unlike a single decision tree, which can have low bias but high variance (prone to overfitting), the Random Forest's ensemble approach helps to mitigate overfitting by reducing variance, leading to better overall performance on unseen data.

Overall, Random Forest is a powerful and widely used machine learning algorithm known for its ability to handle complex datasets, reduce overfitting, and provide reliable predictions. The main idea of the random forest algorithm is presented in Figure (2.10).

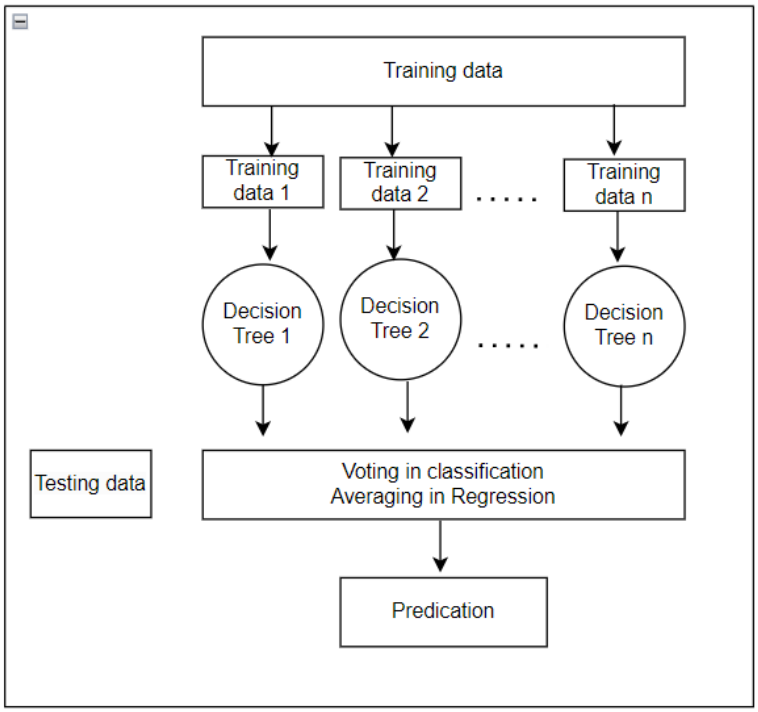


Figure 2.10: The idea of random forest.

The impurity measures for regression and classification tasks in the Random Forest algorithm are different. In regression tasks, the Random Forest algorithm typically uses the mean squared error (MSE) or the mean absolute error (MAE) to measure impurity in equations (4,5) In classification tasks, the Random Forest algorithm uses Gini impurity or entropy to measure impurity(6,7),whereat the number of features in each tree is identified according to the equations (8) [23], Random Forest work steps[59] is presented in Algorithm 2.3:

For regression:

$$variance\ , MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \mu| \quad (4)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \mu)^2 \quad (5)$$

Where y_i: is the actual label (target) for the 'i' instance.

N: is the number of instances in the current node.

μ= is the mean of the labels (targets) in the current node, given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

For classification:

$$Gini = 1 - \Sigma_i f_i^2 \quad (6)$$

$$entropy = -\sum_i f_i log_2 \cdot f_i \quad (7)$$

Where:

- f_i: is the frequency (proportion) of instances with class 'i' at the current node.
- Σ_i: is the sum over all classes.

$$F = log_2 f + 1 \quad (8)$$

- f is the number of input features.

Algorithm 2.3 Random Forest
Input: - Training data - Number of trees in the forest (num_trees) - Stopping criteria for tree growth (e.g., max depth, min-samples-leaf)
Output: Trained Random Forest model
Begin: -For each tree in the forest (from 1 to number of trees): Create a bootstrap sample (a random subset with replacement) from the training data. Randomly select a subset of features for this tree. Build a decision tree using the bootstrap sample and selected features: If stopping criteria are met or only one class remains, make this node a leaf and assign the majority class. Otherwise, choose the best feature and split the data into child nodes. Repeat the splitting process for child nodes until stopping criteria are met. -Store the decision tree in the forest. -To make a prediction for new data (test data): -For each tree in the forest: Traverse the tree based on the input features (test data) to reach a leaf node. Record the class (for classification) -Aggregate the predictions use majority voting among tree predictions.

Return the final prediction.

End

2.6.1.4 Gradient Boosting Classifier

Gradient Boosting, like Random Forest, is a machine learning method that uses a mixture of weak learner (decision trees) to produce a strong prediction model[52]. A weak learner (WL) is a learning algorithm capable of producing classifiers with a strictly (but only slightly) higher likelihood of error. A powerful prediction model, on the other hand, can produce classifiers with arbitrarily low error probability given adequate training data. It constructs trees in a sequential manner, with each new tree rectifying the mistakes of the previous ones. It operates by fitting new decision trees iteratively to preceding trees' residual errors (differences between true values and predictions) [53][54]. This approach allows the ensemble to focus on patterns that the previous trees missed, ultimately enhancing the overall model. Complex issues like as data analysis, text processing, and image identification can be performed more accurately and efficiently utilizing gradient boosting. It's very good at addressing supervised learning problems like classification and regression[55].

2.6.1.5 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is a highly popular and powerful implementation of gradient boosting machines (GBM) used for supervised learning tasks. Gradient boosting is an ensemble learning technique that combines the predictions of multiple weak learners (usually decision trees) to create a strong predictive model. XGBoost is known for its exceptional performance in various machine learning competitions and real-world

applications. It has become a preferred choice for data scientists and machine learning practitioners [56].

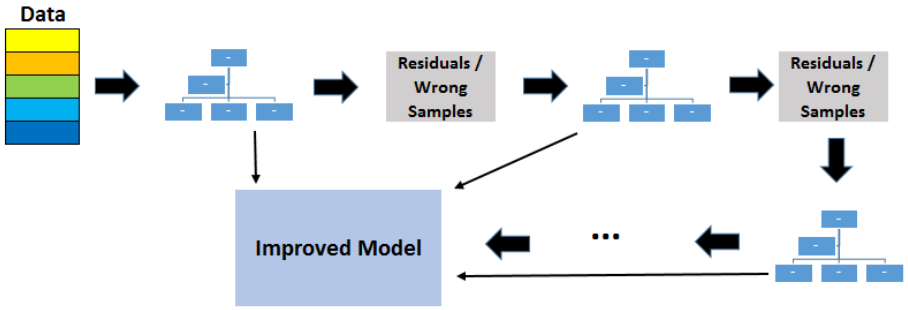


Figure 2.11: Boosting method.

XGBoost combination of high performance, regularization, feature importance analysis, and scalability has made it a popular choice for various machine learning tasks, including classification, regression, ranking, and recommendation systems, among others. Its widespread adoption and continued development in the data science community contribute to its reputation as one of the best performing algorithms for supervised learning. Figure (2.11) shows the boosting method.

The mathematical equations for the XGBoost algorithm involve the objective function, regularization terms, and the optimization process during training[57][50]. To represent the XGBoost algorithm mathematically, we'll use the following notations:

X represents the input features (matrix) of the dataset with n examples and m features.

y_i represents the target output (vector) of the dataset with n examples.

represents the k_{th} tree in the ensemble.

K represents the number of trees in the XGBoost ensemble.

F_k represents the k_{th} tree in the ensemble.

γ is the tree complexity term.

Ω is the regularization term.

L is the loss function.

The objective function of XGBoost can be written as:

$$\bar{O}_i = \phi(x_j) = \sum_{k=1}^{k} f_k(x_i), f_k \in y \quad (12)$$

$$L(\phi) = \Sigma_i L(\check{y}_i, y_i) + \sum_k \Omega(f_k) + \gamma \text{ (K)} \quad (13)$$

2.7 Evaluation of the Prediction Models

Evaluating the performance of classification prediction models is essential to assess their effectiveness and make informed decisions. Here are some commonly used evaluation metrics for classification models[23]:

Tp: True positive Tn: True negative

Fn: False negative Fp: False positive

Accuracy: The ratio of the correctly classified instances to the total number of instances [40].

$$Accurecy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (14)$$

Precision: Precision measures the proportion of true positive predictions (correctly predicted positive instances) out of all positive predictions (true positives + false positives)[40].

$$Precision = \frac{Tp}{Tp + Fp} \quad (15)$$

Recall (Sensitivity or True Positive Rate): Recall measures the proportion of true positive predictions out of all actual positive instances[40].

$$Recall, r = \frac{Tp + Fn}{Tp} \quad (16)$$

F1-Score: The F1-score is the harmonic mean of precision and recall and provides a balanced measure that considers both precision and recall. [40].

$$F1\ score = 2 * \frac{precision + recall}{precision + recall} \quad (17)$$

Chapter Three

The Proposed System

3 The Proposed System

3.1 Overview

This chapter describes the main stages of the proposed system methodology. It begins by presenting the architecture of the proposed system. Next, it covers feature extraction and data preparation, including an explanation of preprocessing techniques. Lastly, the chapter demonstrates the utilization of classifiers and their evaluation to achieve the key aim of this thesis.

3.2 The Proposed System Architecture

The proposed system architecture consists of several main stages, each comprising sub-stages designed to achieve the study objectives. These stages include visual feature extraction, dataset preparation, machine learning models, and their evaluation, as illustrated in Figure 3.1.

To predict popular videos, we utilize a dataset obtained from Kaggle. The first stage of the proposed system focuses on visual feature extraction using pre-trained deep learning techniques such as YOLO v5, specifically designed for object extraction to enrich our dataset.

The second stage involves data preparation, which includes labelling the dataset, processing missing values, and analysing text features using sentiment analysis and interval feature extraction as sub-stages, data preparation encompasses several steps aimed at effectively preparing the input for the system.

In the final stages, a variety of machine learning model are implemented, including K-Nearest Neighbor, Random Forest, Support Vector Machine, Gradient Boosting, and Extreme Gradient Boosting. Subsequently, it

evaluated the outcomes of the proposed model using various methods to estimate predictive errors.

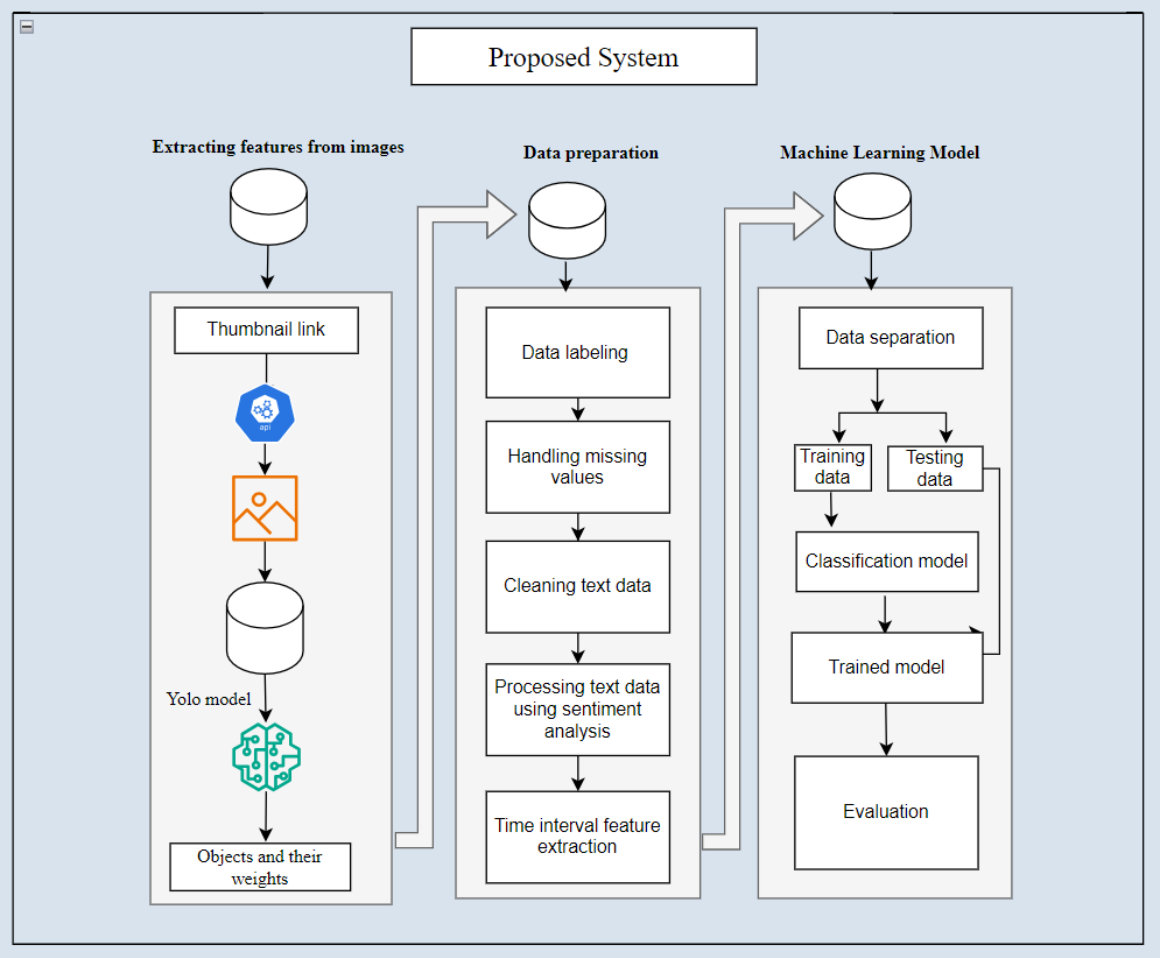


Figure 3.1: The proposed system architecture.

3.2.1 Extraction Features from Images

In this stage, the process begins of fetching thumbnails images for a set of video IDs from YouTube. The process begins with us extracting the video IDs, and API requests are sent to the YouTube API for each video ID. The purpose of these API requests is to retrieve video details, including thumbnail URLs and download the thumbnail to a file. If there are no video details in the API response, its attempting to extract the thumbnail URL from the HTML of the video watch page and download the image Default

thumbnail Algorithm 3.1 outlines the main steps to fetch images, (See Figure 3.2 for a sample of the image dataset).

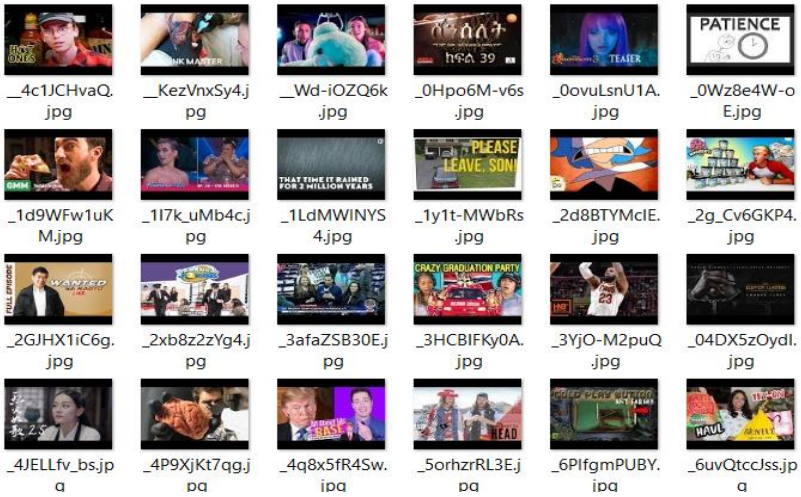


Figure 3.2: Sample of images dataset.

Algorithm 3.1 Retrieve Video Thumbnails
Input: thumbnail links
Output: images
Begin: -Extracting Video IDs from a thumbnail links. -Loop Through Video IDs: -To Retrieve Video Thumbnails: For each video ID, it makes API requests to the YouTube Data API to retrieve video Thumbnail. If video Thumbnail are found in the API response, it downloads the thumbnails to file. If no video Thumbnails are found, it attempts to extract the default thumbnail URL by parsing the video's watch page HTML using Beautiful Soup and downloads the default thumbnail. -Save all the images in a folder.

End

The image processing begins by loading the pre-trained deep learning model (YOLO V5) and its weight file, followed by setting initial values for specific parameters and resizing the image as a preprocessing step. Performing object detection on the preprocessed image involves passing the image through the YOLO model to obtain predictions, Algorithm 3.2 main steps to extract features from images.

Algorithm 3.2 Extract Features from Images
Input: images
Output: set of features (Three largest objects and each object weight) in csv file.
Begin: -Loading a pre-trained YOLOv5 model from specified weights and setting NMS parameters. -Iterating over images in a folder. -Processing each image by resizing it. -Using YOLOv5 to detect objects in the pre-processed image, obtaining attributes like bounding boxes and categories. -Filtering the detected objects and selecting the three largest based on object size, assigning weights based on object size. -Storing selected objects' categories, weights, and image IDs in CSV file for further analysis or use. **End**

Subsequently, non-maximum suppression is applied to remove redundant bounding box predictions. Following this, we process the output predictions

by extracting class labels. We then filter the predictions based on a confidence threshold or other criteria.

In the next step, we select the three largest objects based on their size. Weights are assigned to these objects according to their size, and these values are stored in a CSV file for further analysis or use (see algorithm 3.2). Missing values in the file are handled as described in the algorithm (3.3). When coming across instances in our data that have missing information, we don't want to just leave those missing unfilled. Instead, we go through our entire dataset to find other instances that like the ones with missing data.

Algorithm 3.3 Handling missing values
Input: data with null values
Output: data without missing value
Begin: -Check the selected columns to identify missing values (NaN or null values). -For rows with missing values, search for other rows with similar features in the dataset. - Fill in the missing values in the rows with data from matching rows (similar features) in randomly manner. -Save the dataset with missing values handled to a new file. **End**

This is pretty important because it lets us use information from those similar cases. It's kind of like asking your friends for help when you're missing some details; they might have the answers.

when we pick those similar instances to help us out, we don't always go for the same ones. We mix it up and choose them randomly. Why? There's a

good reason for it. It ensures that our new information comes from all over the dataset, not just one specific place. This keeps our data diverse and prevents us from relying too much on just one source.

it helps reduce any biases that might sneak in if we always picked the same source for filling in the gaps. It makes our data tougher and more reliable for whatever we want to do with it.

In the final step, these features are combined with the original data using the video ID as additional features, figure 3.3 explain Steps Visual Feature Extraction.

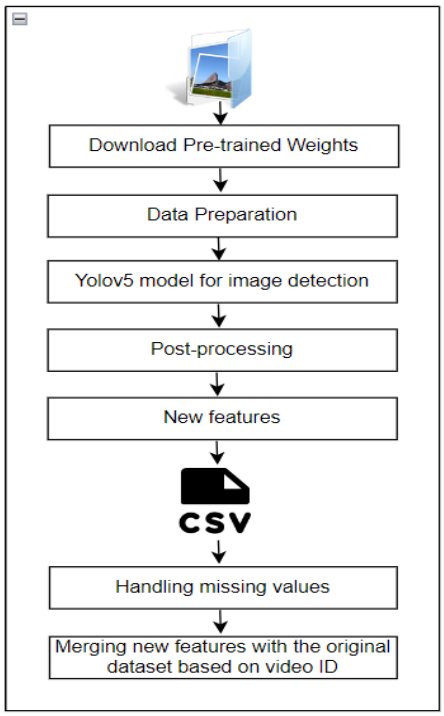


Figure 3.3: Steps Visual Feature Extraction and Merging with Original Dataset.

3.2.2 Data Preparation

3.2.2.1 Dataset Labelling

Data labelling is the process of assigning labels to instances in a dataset. This step was explained in the second chapter, section 2.3, for the purpose of

classifying the dataset I am currently working with. Algorithm 3.4 provides a comprehensive explanation of this labelling process.

Algorithm 3.4 Labelling Video
Input: video-ID, number of like, number of dislikes, number of comments, number of views.
Output: label for each video
Begin Calculated score according this equation: $$score = (likes - 1.5 * dislikes) * comment\_count / views$$ For each video: If views<100,000 Label=0 If views>=100,000 and score<0 Label=1 If views>=100,000 and 0<score<300 Label=2 If views>=100,000 and score>=300 Label =3 Return label **End**

Furthermore, eliminating duplicate entries is an important step in ensuring that each video is only represented once in the collection dataset. This method aids in the avoidance of redundancy and potential bias in subsequent analysis. The dataset becomes more streamlined and accurate by deleting

duplicate records based on unique video IDs, improving its appropriateness for future analysis and modelling.

3.2.2.2 Handling Missing Value

There are very few missing values in one of the features used, specifically the video description. The method employed to handle these missing values is imputation. The imputation methods used include replacing missing values with a constant value, in this case, 'N/A.' This allows the dataset to indicate that description information for these videos is not available or has not been provided, and the specific steps are outlined in Algorithm 3.5.

Overall, these imputation methods help maintain the integrity of the dataset by properly handling missing information and enabling subsequent analyses and modelling to be conducted accurately.

Algorithm 3.5 The Missing value in Description feature
Input: Feature with null values
Output: Feature without null values
Begin: - For each value in specific feature: // Iterate over the feature - IF value is null: // Check if the value is missing - value = filling specific value based on feature // Impute the missing value with specific value (for example not available, unknown) **End**

3.2.2.3 Cleaning Text data

Process text data, specifically video descriptions, video titles, and channel titles by removing any unnecessary characters. Cleaned and pre-processed

text data will yield better sentiment analysis results. This study employs a variety of preprocessing techniques, and the specific steps for cleaning textual features can be found in Algorithm 3.6. The following steps are implemented:

Convert to lowercase: This step converts all characters in the text to lowercase.

Remove URLs from the text: This step uses regular expressions to remove any URLs that start with 'http', 'www', or 'https' from the text.

Remove punctuation marks: This step removes all punctuation marks (e.g., : '!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~', etc.) from the text.

Remove Symbols: This step uses regular expressions to remove any characters that are not letters (a-z, A-Z) or whitespace from the text, for example(🖵 ۞ 🏛 ৷ ¥ € £ ˝ ◌́ ◌̣ ֍ ֎ ▲ ► ▼ ◄ ☺ ☻ ☼ ♀ ♂ ♠ ♣ ♥ ♫ ♯ ⅟ ↗,etc).

Remove words containing digits and individual digits: This step uses regular expressions to remove words that contain digits or individual digits from the text.

Split the text using the provided delimiters: This step splits the text into a list of words using specified delimiters: ',', ':', '.', '>', '<', and whitespace. It breaks the text into individual words, which will be further processed.

Remove empty elements from the list: This step removes any empty elements from the list of words.

Merage the cleaned words back into a single string: This step joins the cleaned words from the list back into a single string using a space (' ') as the separator.

Algorithm 3.6 Text Cleaning Process
Input: Input text
Output: A cleaned version of the input text.
Begin: -Convert text to lowercase -Remove URLs from the text -Remove punctuation marks -Remove Symbols -Remove words containing digits and individual digits -Split the text using the provided delimiters -Remove empty elements from the list -Join the cleaned words back into a single string -Return the cleaned text **End**

3.2.2.4 Features Encoding

After cleaning text data and handling missing values, the next step transforming non numerical features into numerical features for ease of use in data analysis and machine learning algorithms. In my dataset, the 'comment disable' feature can be transformed into a binary representation where True' indicates that comments are disabled, and it is replaced with '1,' while 'False' indicates that comments are enabled and is replaced with '0.'

The steps to perform this coding are outlined in Algorithm 3.7, which provides a systematic approach to convert the non-numerical feature into the desired numerical representation.

Algorithm 3.7 Convert Non numerical Feature into numerical Feature
Input: Non numerical feature (True, False)
Output: Numerical feature (1,0)
Begin: -Identify the non-numerical feature in the dataset that needs to be converted into a numerical feature. -For each value in non-numerical: // Iterate over the instances - If value is True, replace it with 1. -If value is False, replace it with 0. **End**

3.2.2.5 Processing text data using sentiment analysis

To start off, I begin by choosing the text features I want to work with, like the video description, video title, and channel title. These texts then go through a bit of a clean-up and preparation process. This cleaning step is super important because it helps me get better results and uncover more meaningful insights when I'm doing sentiment analysis.

Once the text is cleaned, I break it down into individual words. I use a sentiment dictionary, like the AFINN Dictionary, which rates words in English as positive, negative, or neutral. These ratings range from super negative to very positive, with zero meaning totally neutral.

After giving sentiment scores to each word in the text based on the AFINN sentiment list, I calculate an overall sentiment score for the whole text. I can do this by adding up all the individual scores or by figuring out a weighted average.

Finally, I put the overall sentiment score into different sentiment categories, like positive, negative, or neutral. This classification depends on specific thresholds or rules we've set. For example, if the score goes above a certain threshold, we call it positive; if it drops below, it's considered negative.

3.2.2.6 Time interval features extraction

In the context of video analysis, the feature " time interval " captures the duration it took for a video to attract significant attention or engagement from users. The time interval is calculated by subtracting the publication date from the trending date, seen algorithm 3.8. This calculation yields a time duration, typically measured in hours, days, or weeks, which represents how long it took for the video to gather significant attention or become trending.

This calculated time interval is then incorporated as feature in the dataset, enabling its use alongside other features to analyse patterns, correlations, or predictive relationships. By utilizing the " time interval " analysts can gain insights into the time dynamics and popularity growth of videos. It becomes a valuable tool for understanding the factors that contribute to a video's success and for predicting future trends or user engagement patterns.

Algorithm 3.8 Time Interval Feature Extraction
Input: published date, trending date.
Output: time interval value in days
Begin:

-Convert publish date to suitable format.

- Convert trending date to suitable format.

-For each instance:

Subtract the trending date value from the corresponding publish date value.

Store the result in the time interval feature

- Return time interval features

End

3.2.3 Machine Learning Models

Machine learning models were used to predict the popularity of YouTube videos by classifying them into four categories. To do this, I first extracted a set of features from the thumbnail video images using the pre-trained Yolo model.

Additionally, textual features were extracted and processed, along with the time interval between the video's publication and its popularity, and the statistical features that are originally present in the data set were also used, such as the number of likes, comments, dislikes, video disabling, and video category. After processing the data, an appropriate classifier was applied, and predictions were made in multiple stages, depending on the features used.

The total dataset was divided into a training set (70%) and a testing set (30%). Five classifiers were applied to find the best, these are:

- Support Vector Machine
- K-Nearest Neighbor Classifier
- Random Forest Classifier

- Gradient Boosting Classifier
- Extreme Gradient Boosting Classifier

After applying the five classifiers, the Random Forest and XGBoost algorithms demonstrated superior performance, achieving the highest accuracy and outperforming other classifiers.

3.2.4 Evaluating the Performance of the Proposed Model

Different separate measures were employed to assess the effectiveness of a specific categorization system. Accuracy, f1-score, precision, and recall are all included. These measurements are calculated using a confusion matrix, which is a matrix that represents the number of cases that are correctly or incorrectly predicted by a classification model. In this thesis, the predicted classes are four, so the form of the confusion matrix would be as follows (see Figure 3.4)

		Predicted Class			
		0	**1**	**2**	**3**
Actual Class	**0**	**TP**	**FP**	**FP**	**FP**
	1	**FN**	**TP**	**FP**	**FP**
	2	**FN**	**FN**	**TP**	**FP**
	3	**FN**	**FN**	**FN**	**TP**

Figure 3.4: The Confusion matrix of a four class.

Chapter Four

Result and Discussion

4 Results and Discussions

4.1 Introduction

This chapter presents the outcomes of the various stages of the proposed system, as discussed in Chapter three. It provides a comprehensive overview of the results obtained from each stage, including feature extraction from images, text cleaning and processing, and the performance of the machine learning models employed.

4.2 Research Requirements

Hardware: Processor Intel ci7, RAM 16GB.

Operating System: Windows10 (64) bit.

The system was implemented using Python 3.11 and the PyCharm.

4.3 Proposed System Results

4.3.1 Extract visual features from images

After using pre-trained YOLOv5 to process images and extract objects, see figure 4.1.

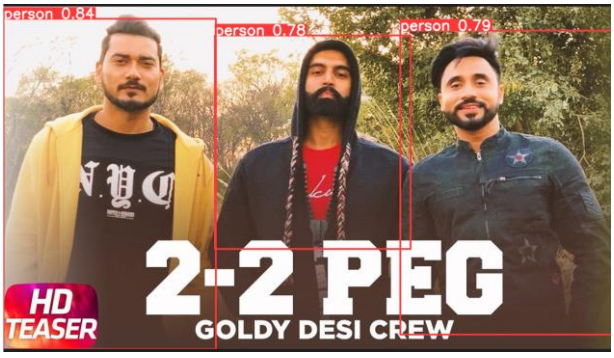


Figure 4.1: The result of using pre-trained yolo v5 model.

The result of this step is a group of objects and related information. The three largest objects are selected based on their size in the image (boundary box), and a weight is assigned to each object depending on its size in the image. Each object is represented by a number belonging to 80 classes with a weight for each object, as shown in the Table 4.2.

Table 4.1 A: Feature set extracted from images (before processing step).

video_id	ObjectCategory 1	ObjectCategory 2	ObjectCategory 3	ObjectWeight 1	ObjectWeight 2	ObjectWeight_3
0dBIKQ4Mz1M	0	0	nan	0.2400	0.3793	nan
d380meDoWoM	0	0	0	0.517	0.296	0.186
2Vv-BfVoq4g	16	74	nan	0.945	0.054	nan
0ylWz1XEeyc	0	0	0	0.5519	0.2490	0.1989
6ylqz1qWeyu	16	32	4	0.329	0.094	0.502

There are missing values in the extracted features, likely due to the absence of three objects in some images, as previously discussed in Section 3.2.1 Which were processed, (Table 4.1: A illustrates the data before processing step).

When there are missing values in object category 2 or object category 3, the dataset will be searched for videos that are similar to object category 1. From the found videos, one will be randomly selected to fill the missing values in either object category 2 or object category 3, along with their respective weights. and the (Table 4.1 B: shows the results after processing the empty values).

Table 4.1 B: Feature set extracted from images (after processing step).

video_id	ObjectCategory 1	ObjectCategory 2	ObjectCategory 3	ObjectWeight 1	ObjectWeight 2	ObjectWeight_3
0dBIKQ4Mz1M	0	0	0	0.2400	0.3793	0.1989
d380meDoWoM	0	0	0	0.517	0.296	0.186
2Vv-BfVoq4g	16	74	4	0.945	0.054	0.502
0ylWz1XEeyc	0	0	0	0.5519	0.2490	0.1989
6ylqz1qWeyu	16	32	0	0.329	0.094	0.502

Subsequently, the extracted features are integrated with the original features using the Image ID, which corresponds to the Video ID in the original

dataset. This integration is a crucial step as it allows for the combination of both sets of features for further analysis and modeling.

It's important to note that each value in the first three columns represent a category(classes), with the number 0 denoting a person and the number 27 representing a tie (To find out more, see Table 4.2), the other three columns are the weights for each object, its importance according to its size in the image, as indicated in Table 4.1 .

Table 4.2 shows the number and what it represents (category).

Number	Class	Number	Class
0	person	41	cup
1	bicycle	42	fork
2	car	43	knife
3	motorbike	44	spoon
4	aeroplane	45	bowl
5	bus	46	banana
6	train	47	apple
7	truck	48	sandwich
8	boat	49	orange
9	traffic light	50	broccoli
10	fire hydrant	51	carrot
11	stop sign	52	hot dog
12	parking meter	53	pizza
13	bench	54	donut
14	bird	55	cake
15	cat	56	chair
16	dog	57	sofa
17	horse	58	pottedplant
18	sheep	59	bed
19	cow	60	diningtable
20	elephant	61	toilet
21	bear	62	tvmonitor
22	zebra	63	laptop
23	giraffe	64	mouse
24	backpack	65	remote
25	umbrella	66	keyboard
26	handbag	67	cell phone
27	tie	68	microwave
28	suitcase	69	oven
29	frisbee	70	toaster
30	skis	71	sink
31	snowboard	72	refrigerator
32	sports ball	73	book
33	kite	74	clock
34	baseball bat	75	vase
35	baseball glove	76	scissors
36	skateboard	77	teddy bear
37	surfboard	78	hair drier
38	tennis racket	79	toothbrush
39	bottle	80	Footer
40	wine glass		

4.3.2 Data Preparation Results

4.3.2.1 Dataset labelling

The result of labeling the data into four classes, as mentioned in the proposed system, explain in table 4.3.

Table 4.3 The result of labelling the data into four classes.

video_id	trending_date	category_id	.....	likes	Label
n1WpP7iowLc	17.14.11	23	.....	127794	1
0dBIkQ4Mz1M	17.14.11	23	.....	146035	1
5qpjK5DgCt4	17.14.11	24	.....	132239	1
d380meD0W0M	17.14.11	10	.....	1634130	1

At this stage of pre-processing, duplicate records in the video dataset have been successfully removed based on duplicate video IDs. The dataset now contains only unique records (see figure 4.4), ensuring that each video is represented by a single entry without any redundant observations.

Table 4.4 A: The Results Before of Removing duplicate records.

video_id	trending_date	category_id	.....	likes
n1WpP7iowLc	17.14.11	23	.....	127794
0dBIkQ4Mz1M	17.14.11	23	.....	146035
5qpjK5DgCt4	17.14.11	24	.....	132239
0dBIkQ4Mz1M	17.14.11	23	.....	146035
d380meD0W0M	17.14.11	10	.....	1634130
n1WpP7iowLc	17.14.11	23	.....	127794

Table 4.4 B: The Results After of Removing duplicate records.

video_id	trending_date	category_id	.....	likes
n1WpP7iowLc	17.14.11	23	.....	127794
0dBIkQ4Mz1M	17.14.11	23	.....	146035
5qpjK5DgCt4	17.14.11	24	.....	132239
d380meD0W0M	17.14.11	10	.....	1634130

4.3.2.2 Missing value in Description

For some videos, the description field contains missing values, which are processed by assigning the value 'Not available' This allows the dataset to indicate that the description information for those videos is not available or not provided. The table (4.5) shows the results of this step.

Table 4.5 A: Description Features Before Handling the Missing Values.

video_id	description
n1WpP7iowL	Eminem's new track Walk on Water ft. Beyoncé i...
0dBIkQ4Mz1	nan
5qpjK5DgCt4	I know it's been a while since we did this sho...
8HNuRNi8t7(	nan

Table 4.5 B: Description Features After Handling the Missing Values.

video_id	description
n1WpP7iowL	Eminem's new track Walk on Water ft. Beyoncé i...
0dBIkQ4Mz1	Not available
5qpjK5DgCt4	I know it's been a while since we did this sho...
8HNuRNi8t7(	Not available

4.3.2.3 Cleaning text features

The results of the pre-processing stage are shown, where textual features such as video descriptions, video titles, and video channel titles are cleaned by removing irrelevant information. Table 4.6 shows the results of the pre-processing steps for the textual features within its dataset.

Table 4.6: The Results of Each Step on the Sample Video Description.

Original text	Ending Explained for the latest from master Guillermo Del Toro, the moving romantic monster movie2 THE SHAPE OF WATER starring Sally Hawkins and Doug Jones. Plus, analyzing the films bigger meaning and themes. Subscribe! ►► http://bit.ly/2jrstgM Support FoundFlix on Patreon! ►► http://www.patreon.com/foundflix === Connect with us on Social Media! === FACEBOOK ►► www.facebook.com/foundflix TWITTER ►► www.twitter.com/foundflix INSTAGRAM ►► www.instagram.com/foundflix
Step 1: After Convert to lowercase	ending explained for the latest from master guillermo del toro, the moving romantic monster movie2 the shape of water starring sally hawkins and doug jones. plus, analyzing the films bigger meaning and themes. subscribe! ►► http://bit.ly/2jrstgm support foundflix on patreon! ►► http://www.patreon.com/foundflix

	=== connect with us on social media! === facebook ►► www.facebook.com/foundflix twitter ►► www.twitter.com/foundflix instagram ►► www.instagram.com/foundflix
Step 2: After Remove URLs from the text	ending explained for the latest from master guillermo del toro, the moving romantic monster movie2 the shape of water starring sally hawkins and doug jones. plus, analyzing the films bigger meaning and themes. subscribe! ►► support foundflix on patreon! ►► === connect with us on social media! === facebook ►► twitter ►► instagram ►►
Step 3: After Remove punctuation marks	ending explained for the latest from master guillermo del toro the moving romantic monster movie2 the shape of water starring sally hawkins and doug jones plus analyzing the films bigger meaning and themes subscribe ►► support foundflix on patreon ►► connect with us on social media facebook ►► twitter ►► instagram ►►
Step 4: After Remove special	ending explained for the latest from master guillermo del toro the moving romantic monster movie2 the

characters, numbers, and extra spaces	shape of water starring sally hawkins and doug jones plus analyzing the films bigger meaning and themes subscribe support foundflix on patreon connect with us on social media facebook twitter instagram
Step 5: After remove words containing digits and individual digits	ending explained for the latest from master guillermo del toro the moving romantic monster movie the shape of water starring sally hawkins and doug jones plus analyzing the films bigger meaning and themes subscribe support foundflix on patreon connect with us on social media facebook twitter instagram
Step 6: After split the text using the provided delimiters	['', 'ending', 'explained', 'for', 'the', 'latest', 'from', 'master', 'guillermo', 'del', 'toro', '', 'the', 'moving', 'romantic', 'monster', 'movie', 'the', 'shape', 'of', 'water', 'starring', 'sally', 'hawkins', 'and', 'doug', 'jones', '', 'plus', '', 'analyzing', 'the', 'films', 'bigger', 'meaning', 'and', 'themes', '', '', 'subscribe', '', '', '', 'support', 'foundflix', 'on', 'patreon', '', '', '', '', '', '', '', '', 'connect', 'with', 'us', 'on', 'social', 'media', '', '', '', '', '', 'facebook', '', '', 'twitter',

	'', '', 'instagram', '', '', '']
Step 7: After remove empty element in list	['ending', 'explained', 'for', 'the', 'latest', 'from', 'master', 'guillermo', 'del', 'toro', 'the', 'moving', 'romantic', 'monster', 'movie', 'the', 'shape', 'of', 'water', 'starring', 'sally', 'hawkins', 'and', 'doug', 'jones', 'plus', 'analyzing', 'the', 'films', 'bigger', 'meaning', 'and', 'themes', 'subscribe', 'support', 'foundflix', 'on', 'patreon', 'connect', 'with', 'us', 'on', 'social', 'media', 'facebook', 'twitter', 'instagram']
Step 8: After Join the cleaned words back into a single string	ending explained for the latest from master guillermo del toro the moving romantic monster movie the shape of water starring sally hawkins and doug jones plus analyzing the films bigger meaning and themes subscribe support foundflix on patreon connect with us on social media facebook twitter instagram

4.3.2.4 Encoding features

In this step, the comment disable feature is converted from True/False representation to numerical values (0 or 1). This transformation prepares the data for machine learning algorithms that require numerical inputs, facilitating analysis, training, and predictions. Table 4.7 illustrates the outcome of this conversion process.

Table 4.7 A: The Results Before of Handling non numerical Features.

Video_id	Comment_disable
n1WpP7iowLc	FALSE
0dBIkQ4Mz1M	FALSE
5qpjK5DgCt4	TRUE
8HNuRNi8t70	FALSE

Table 4.7 B: The Results After of Handling non numerical Features

Video_id	Comment_disable
n1WpP7iowLc	0
0dBIkQ4Mz1M	0
5qpjK5DgCt4	1
8HNuRNi8t70	0

4.3.2.5 Processing text data using sentiment analysis

The table 4.8 shows the proportion of sentiment conveyed in video descriptions, titles tags, and channel titles.

Video_id	Video Description	Video Title	Channal Title	Tags
0dBIkQ4Mz1M	0.250000	-0.700000	0.0	0.150000
d380meD0W0M	0.459091	0.000000	0.0	-0.350000
2Vv-BfVoq4g	0.200000	1.000.000	0.0	0.136364
0ylWz1XEeyc	-0.131818	0.000000	0.0	0.000000
_uM5kFfkhB8	0.250000	0.136364	0.0	0.056667

Table 4.8: The Textual Features Extraction.

Allowing the extraction and analysis of sentiment from textual data provides valuable insights into the emotions conveyed in the videos' descriptions, titles, tags and channel titles. The objective was to categorize the overall sentiment as positive, negative, or neutral, providing a deeper understanding of how viewers may perceive the videos based on their titles and

descriptions. Understanding the sentiment behind the textual content helps reveal the emotional impact and tone of the videos. This information is valuable for influencers, marketers, and researchers, as it can be used to optimize video titles, descriptions, tags, and channel titles to enhance engagement and audience response.

To illustrate the distribution of positive, negative, and neutral sentiments across the textual features, Figures (4.10), (4.11), and (4.12) are presented. Additionally, Figures (4.13), (4.14), and (4.15) show the ratio of positive, negative, and neutral (zero) sentiments in the text features.

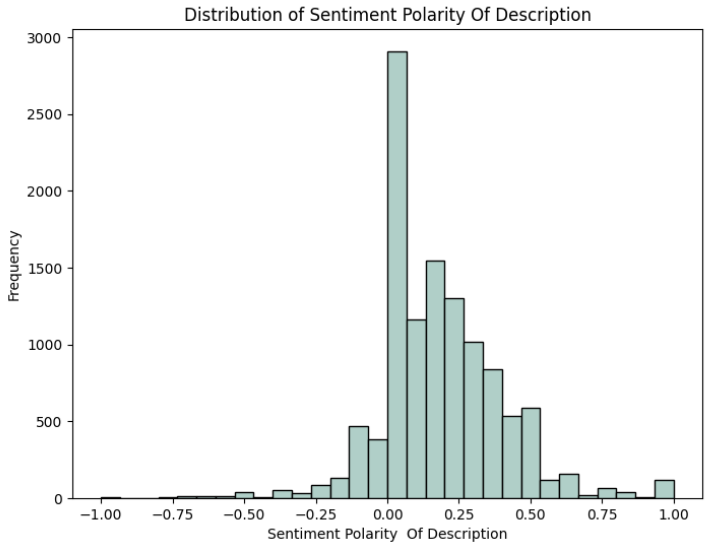


Figure 4.10: Distribution Sentiments Across the Description Features.

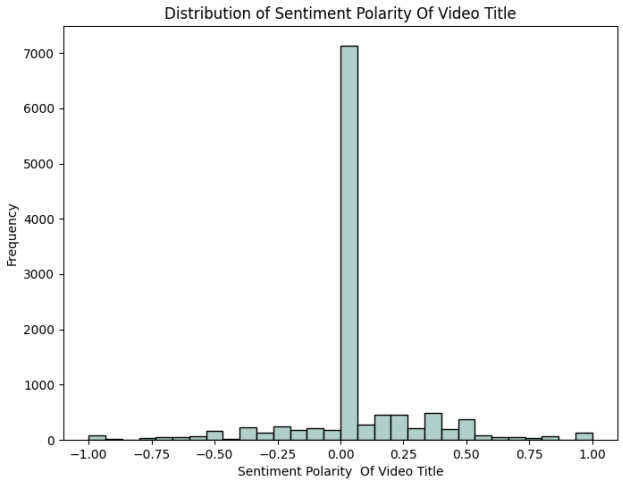


Figure 4.11: Distribution Sentiments Across the Video Title Features.

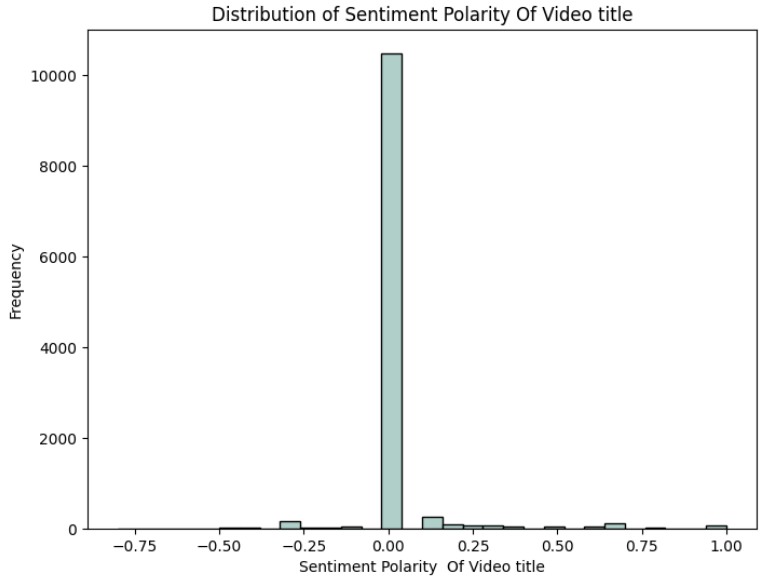


Figure 4.12: Distribution Sentiments Across the Description Features.

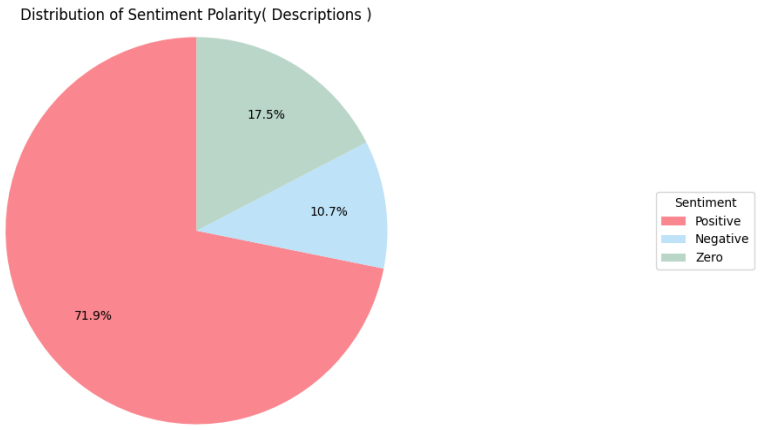


Figure 4.13: The Ratio of Positive (+), Negative (-), and Neutral (Zero) Sentiments in the Video Title Feature Description

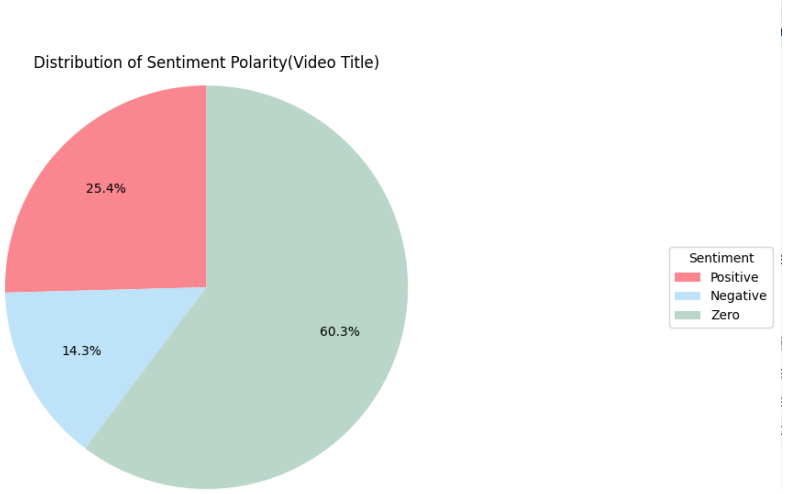


Figure 4.14: The Ratio of Positive (+), Negative (-), And Neutral (Zero)Sentiments in The Video Title Feature.

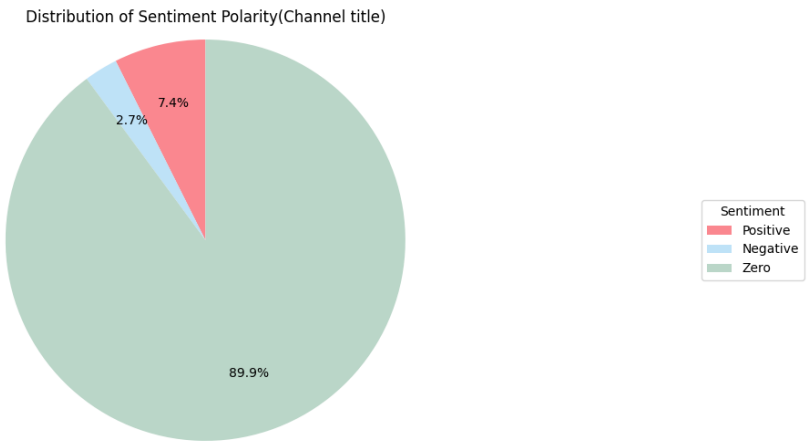


Figure 4.15: The Ratio of Positive (+), Negative (-), And Neutral (Zero)Sentiments in the Channel Title Feature.

4.3.2.6 Time interval features extraction

Time interval feature provides valuable insights into how quickly a video gained popularity after being published. A smaller Time interval indicates that a video became popular shortly after its release, while a larger Time interval suggests that the video took more time to gain traction and trend among viewers. Table (4.9) displays the calculated Time interval values for each video in the dataset.

Table 4.9: Time interval Feature for Each Video in the Dataset.

video_id	publish_time	trending_date	Time interval(in day)
0dBIkQ4Mz1M	13/11/2017 17:00	14/11/2017	0.291667
d380meD0W0M	12/11/2017 18:01	14/11/2017	1.248.831
2Vv-BfVoq4g	09/11/2017 11:04	14/11/2017	4.538.727
0ylWz1XEeyc	13/11/2017 07:37	14/11/2017	0.682049
_uM5kFfkhB8	12/11/2017 23:52	14/11/2017	1.005.405

4.3.3 Results of the Classification Methods

In my study, I conducted a detailed comparison of video classification techniques employing various machine learning algorithms and combinations of features. To evaluate the models, we employed a dataset comprising 24,427 categorized videos, which we divided into 70% for training and 30% for testing, as illustrated in Figure 4.17.

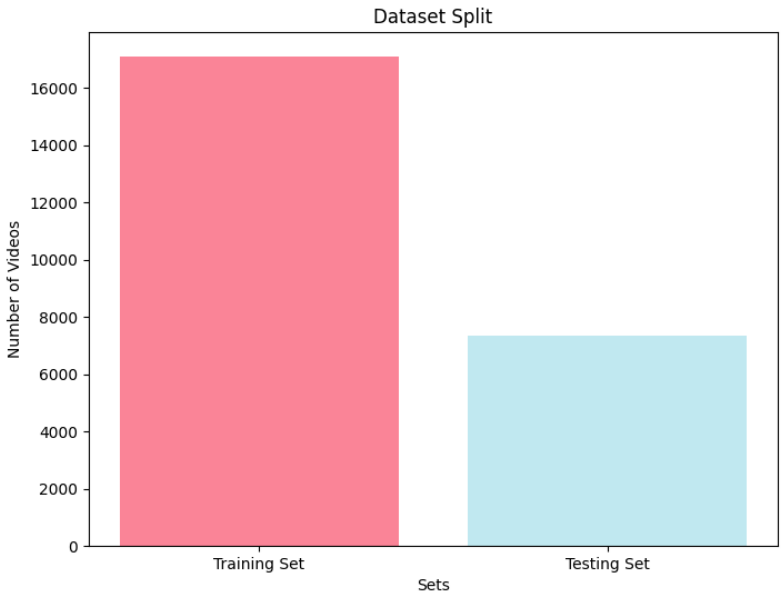


Figure 4.17: Separation of Data into Training and Testing Data.

Types of features used:

- Metadata (Statistic): such Like, Dislike, Comment, Comment Disable, Category ID which exist in dataset.
- Extracted Features:

  Extracted Text Features using Sentiment Analysis: Video Description, tags, Video Title, Video Channel Title.

  Time Interval: The time between when a video was published and when it became popular.

Visual or Extracted Features from Images: Object Category, Object Weights

Now, let me provide you with a brief overview of the key findings from my evaluations:

4.3.3.1 Evaluation Using All Features

Following the preprocessing phases, the stages of feature extraction, classification, and performance evaluation are implemented. The prediction accuracy is obtained in the first phase based on all extracted and existing features in the dataset.

Table 4.10 summarizes the findings. Other algorithms are outperformed by Random Forest and XGBoost. This implies that these algorithms are well-suited to the dataset and attributes employed. Among the algorithms, the Support Vector Classifier has the lowest accuracy. With the stated features, it might not be the best pick for this dataset. In terms of performance, K-Nearest Neighbor and Gradient Boosting sit in between RF/XGB and SVC. They outperform SVC but fall short of RF and XGB, seen figure 4.18.

- Random Forest accuracy reaching to 96.84.
- Extreme Gradient Boosting accuracy reaching to 96.49%.
- K-Nearest Neighbor accuracy to reaching 90.26%.
- Gradient Boosting accuracy to reaching 90.01%.
- Support Vector Machine accuracy reaching to 84.86%.

Table 4.10: Accuracy Comparison of Classifiers Using all features.

Algorithms	Accuracy	Precision	Recall	f-score
KNN	91.26	91.61	91.02	90.92
RF	96.94	97.00	96.86	96.89
SVC	87.55	87.77	87.37	87.40
GB	93.07	92.96	92.99	92.94
XGB	96.33	96.38	96.24	96.27

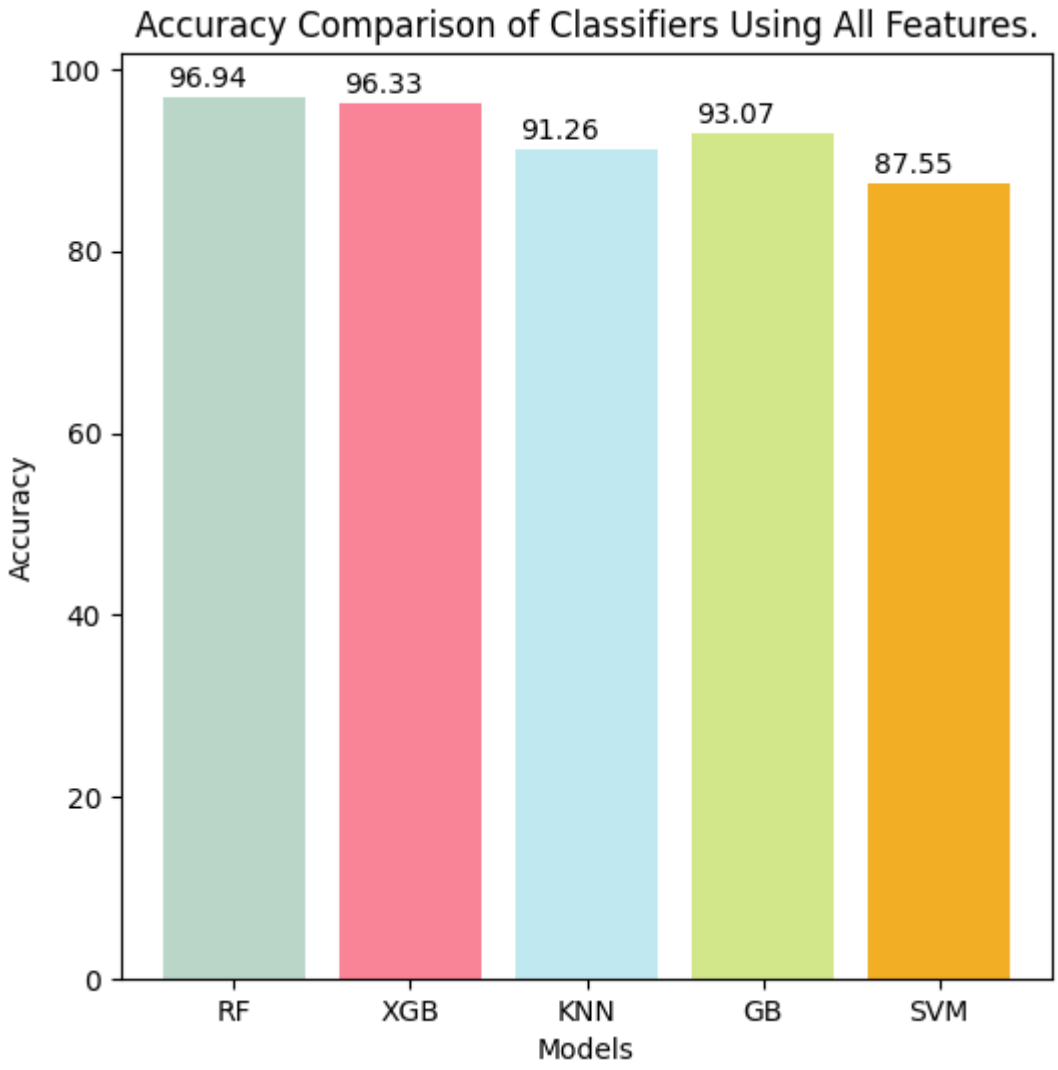


Figure 4.18 showing the results of the First evaluation.

4.3.3.2 Evaluation Using Extracted Features Only

It is also worth noting that when using only the extracted and processed features (text, time interval, and visual features), RF and XGB outperformed the previous research on the same dataset. This indicates the effectiveness of these algorithms for video classification tasks, as shown in Table 4.11. This demonstrates the efficacy of the features obtained from both images and text, which are superior to the features used in previous research. This underscores the effectiveness of this methods.

- RF achieved an accuracy of 93%.
- XGB achieved an accuracy of 84.48%.

Table 4.11: Shows That Random Forest (RF) And XGBoost (XGB) Outperformed the Previous Research on the Same Dataset Using Extracted Features Only.

References	Algorithms	F1 score using Extracted features
[1]	RF	70%
	XGB	74%
This study	**Algorithms**	**F1 score**
	RF	93.00 %
	XGB	84.48 %

Random Forest consistently achieved well across all evaluations, indicating its reliability and making it an appropriate option across different feature sets. In contrast, Extreme Gradient Boosting performed well in the first two evaluations but fell short in the third evaluation when only extracted features were used. Carefully chosen features extracted from video thumbnails,

descriptions, tags, and titles can benefit marketers and influencers by providing insights into the nature of the video, ultimately helping them make their videos more popular.

Chapter Five

Conclusions and Future Works

5 Conclusions and Future Works

5.1 Conclusions

- Combining various features to build our prediction model is a highly effective approach.
- Extracting additional features, including those related to video thumbnails, has proven to be an effective strategy.
- Employing sentiment analysis has helped me better understand emotions and feelings in textual features, enhancing our insights.
- Through the use of robust classification algorithms like Random Forest and Extreme Gradient Boosting, our study achieved remarkable accuracy levels of nearly 96% and 97%, respectively.
- SVM accuracy are lower compared to Random Forest and XGBoost, indicating that these models may not be the best choices for this specific dataset and feature set.
- The real value of our analysis is in gaining insights into the factors that influence video popularity.

In the final, Understanding the impact of titles, tags, descriptions, and thumbnails on viewer engagement offers influencers and marketers the opportunity to customize their video content, making it more appealing to their target audience. This, in turn, increases the likelihood of video success on social media platforms.

5.2 Future Work

Future work can build upon the findings and methodologies presented in this thesis to further enhance the prediction of video popularity on social media platforms. Some potential areas for future research include:

- Refining and optimizing the prediction model by exploring addition combinations of features using API.
- Extending the research to include other social media platforms besides YouTube.
- Comparing and contrasting the factors influencing video popularity on different platforms, allowing influencers and marketers to tailor their strategies for each platform.

By addressing these areas of future work, researchers can further advance the field of video popularity prediction and contribute to more effective content marketing and user engagement strategies on social media platforms.

References

6 References

[1] Y. Li, K. Eng, and L. Zhang, “YouTube Videos Prediction: Will this video be popular ?,” 2019, [Online]. Available: http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647615.pdf (accessed 9 May 2020).

[2] M. L. Khan, “Social media engagement: What motivates user participation and consumption on YouTube?,” *Comput. Human Behav.*, vol. 66, pp. 236–247, Jan. 2017, doi: 10.1016/J.CHB.2016.09.024.

[3] F. Figueiredo, F. Benevenuto, and J. M. Almeida, “The tube over time: Characterizing popularity growth of YouTube videos,” *Proc. 4th ACM Int. Conf. Web Search Data Mining, WSDM 2011*, pp. 745–754, 2011, doi: 10.1145/1935826.1935925.

[4] Q. I. A. O. Sibo, P. A. N. G. Shanchen, W. A. N. G. Min, Z. H. A. I. Xue, and D. A. I. Feng, “Online Video Popularity Regression Prediction Model with Multichannel Dynamic Scheduling Based on User Behavior,” *Chinese J. Electron.*, vol. 30, no. 5, pp. 876–884, Sep. 2021, doi: 10.1049/cje.2021.06.010.

[5] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, “Viral video style: A closer look at viral videos on YouTube,” *ICMR 2014 - Proc. ACM Int. Conf. Multimed. Retr. 2014*, pp. 193–200, 2014, doi: 10.1145/2578726.2578754.

[6] T. Trzcinski and P. Rokita, “Predicting popularity of online videos using Support Vector Regression,” Oct. 2015, doi: 10.1109/TMM.2017.2695439.

[7] M. U. N. Nisa, D. Mahmood, G. Ahmed, S. Khan, M. A. Mohammed, and R. Damaševičius, “Optimizing prediction of youtube video popularity using xgboost,” *Electron.*, vol. 10, no. 23, 2021, doi: 10.3390/electronics10232962.

[8] R. Shreyas, D. M. Akshata, B. S. Mahanand, B. Shagun, and C. M. Abhishek, “Predicting popularity of online articles using Random Forest regression,” *Proc. - 2016 2nd Int. Conf. Cogn. Comput. Inf. Process. CCIP 2016*, 2016, doi: 10.1109/CCIP.2016.7802890.

[9] F. Huang, J. Chen, Z. Lin, P. Kang, and Z. Yang, “Random forest exploiting post-related and user-related features for social media popularity prediction,” *MM 2018 - Proc. 2018 ACM Multimed. Conf.*, pp. 2013–2017, 2018, doi: 10.1145/3240508.3266439.

[10] T. Trzcinski, P. Andruszkiewicz, T. Bochenski, and P. Rokita, “Recurrent Neural Networks for Online Video Popularity Prediction,” Jul. 2017, doi: 10.1007/978-3-319-60438-1_15.

[11] N. Sangwan and V. Bhatnagar, “Video popularity prediction based on fuzzy inference system,” *J. Stat. Manag. Syst.*, vol. 23, no. 7, pp. 1173–1185, 2020, doi: 10.1080/09720510.2020.1799577.

[12] M. L. Khan, “Social media engagement: What motivates user participation and consumption on YouTube?,” *Comput. Human Behav.*, vol. 66, pp. 236–247, 2017, doi: 10.1016/j.chb.2016.09.024.

[13] A. Neyaz, A. Kumar, S. Krishnan, J. Placker, and Q. Liu, “Security, Privacy and Steganographic Analysis of FaceApp and TikTok,” *Int. J. Comput. Sci. Secur.*, vol. 14, no. 2, pp. 38–59, 2020, [Online]. Available: https://www.researchgate.net/publication/341782197

[14] Y. Fan, B. Yang, D. Hu, X. Yuan, and X. Xu, “Social- And Content-Aware Prediction for Video Content Delivery,” *IEEE Access*, vol. 8, pp. 29219–29227, 2020, doi: 10.1109/ACCESS.2020.2972920.

[15] S. L. de Sá, A. A. d. A. Rocha, and A. Paes, “Predicting popularity of video streaming services with representation learning: A survey and a real-world case study,” *Sensors*, vol. 21, no. 21, 2021, doi: 10.3390/s21217328.

[16] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, “A survey of sentiment analysis in social media,” *Knowl. Inf. Syst.*, vol. 60, no. 2, pp. 617–663, 2019, doi: 10.1007/s10115-018-1236-4.

[17] M. Ham and S. W. Lee, “Factors affecting the popularity of video content on live-streaming services: Focusing on V live, the South Korean live-streaming service,” *Sustain.*, vol. 12, no. 5, pp. 1–17, 2020, doi: 10.3390/su12051784.

[18] J. Davidson, B. Liebald, J. Liu, P. Nandy, and T. Van Vleet, “The YouTube video recommendation system,” *RecSys’10 - Proc. 4th ACM Conf. Recomm. Syst.*, no. August 2014, pp. 293–296, 2010, doi: 10.1145/1864708.1864770.

[19] “Key Features of Popular Social Media Platforms.” https://techeconomy.ng/key-features-of-popular-social-media-platforms/ (accessed Aug. 01, 2023).

[20] L. H. X. Ng, J. Y. H. Tan, D. J. H. Tan, and R. K. W. Lee, “Will you dance to the challenge?: Predicting user participation of TikTok challenges,” in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,*

ASONAM 2021, Nov. 2021, pp. 356–360. doi: 10.1145/3487351.3488276.

[21] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “Instagram post popularity trend analysis and prediction using hashtag, image assessment, and user history features,” *Int. Arab J. Inf. Technol.*, vol. 18, no. 1, pp. 85–94, 2021, doi: 10.34028/iajit/18/1/10.

[22] Mitchell J., “Trending YouTube Video Statistics | Kaggle,” *Kaggle*, 2019. https://www.kaggle.com/datasnaek/youtube-new (accessed Mar. 11, 2023).

[23] J. M. Luna, *Introduction to Data Mining*. 2021. doi: 10.1007/978-981-16-3964-7_1.

[24] S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” *Proc. 2014 Sci. Inf. Conf. SAI 2014*, no. July, pp. 372–378, 2014, doi: 10.1109/SAI.2014.6918213.

[25] O. I. Obaid, M. A. Mohammed, A. O. Salman, S. A. Mostafa, and A. A. Elngar, “Comparing the Performance of Pre-trained Deep Learning Models in Object Detection and Recognition,” *J. Inf. Technol. Manag.*, vol. 14, no. 4, pp. 40–56, 2022, doi: 10.22059/JITM.2022.88134.

[26] M. Jogin, “Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning,” *2018 3rd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol.*, no. November, pp. 2319–2323, 2020, doi: 10.1109/RTEICT42901.2018.9012507.

[27] M. Zabir, N. Fazira, Z. Ibrahim, and N. Sabri, “Evaluation of pretrained Convolutional Neural Network models for object recognition,” *Int. J.*

Eng. Technol., vol. 7, no. 3, pp. 95–98, 2018, doi: 10.14419/ijet.v7i3.15.17509.

[28] "A SURVEY ON OBJECT DETECTION TECHNIQUES USING TENSORFLOW , KERAS AND YOLO," no. 4, pp. 495–500, 2022.

[29] B. S. Rekha, A. Marium, G. N. Srinivasan, and S. A. Shetty, "Literature Survey on Object Detection using YOLO," *Int. Res. J. Eng. Technol.*, vol. 07, no. 06, pp. 3082–3088, 2020, [Online]. Available: https://www.irjet.net/archives/V7/i6/IRJET-V7I6576.pdf

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[31] A. P. Jana, A. Biswas, and Mohana, "YOLO based detection and classification of objects in video records," *2018 3rd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. RTEICT 2018 - Proc.*, no. April, pp. 2448–2452, 2018, doi: 10.1109/RTEICT42901.2018.9012375.

[32] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[33] M. H. Zhu, "Research on data preprocessing in exam analysis system," *Lect. Notes Electr. Eng.*, vol. 100 LNEE, no. VOL. 4, pp. 333–338, 2011, doi: 10.1007/978-3-642-21762-3_43.

[34] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data*

Anal., vol. 1, no. 1, pp. 1–22, 2016, doi: 10.1186/s41044-016-0014-0.

[35] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.

[36] E. Cho, T. W. Chang, and G. Hwang, "Data Preprocessing Combination to Improve the Performance of Quality Classification in the Manufacturing Process," *Electron.*, vol. 11, no. 3, pp. 1–15, 2022, doi: 10.3390/electronics11030477.

[37] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, 2009, doi: 10.4304/jetwi.1.1.60-76.

[38] A. Negi, "A Brief Survey On Text Mining, Its Techniques, And Applications," *Int. J. Mob. Comput. Appl.*, vol. 8, no. 1, pp. 1–6, 2021, doi: 10.14445/23939141/ijmca-v8i1p101.

[39] A. Shiri, "Introduction to Modern Information Retrieval (2nd edition)," *Libr. Rev.*, vol. 53, no. 9, pp. 462–463, 2004, doi: 10.1108/00242530410565256.

[40] M. Mujahid *et al.*, "Sentiment analysis and topic modeling on tweets about online education during covid-19," *Appl. Sci.*, vol. 11, no. 18, 2021, doi: 10.3390/app11188438.

[41] I. H. Sodhar, "Computer Science," no. December 2020, 2022, doi: 10.22271/ed.book.784-CITATIONS.

[42] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive

markers of depression," *EPJ Data Sci.*, vol. 6, no. 1, 2017, doi: 10.1140/epjds/s13688-017-0110-z.

[43] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[44] B. S. Ainapure *et al.*, "Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches," *Sustain.*, vol. 15, no. 3, pp. 1–22, 2023, doi: 10.3390/su15032573.

[45] S. Velickov and D. Solomatine, "Predictive Data Mining : Practical Examples Abstract : 2 . Data Mining – theoretical and practical aspects," *Neural Networks*, no. March, pp. 1–17, 2000.

[46] U. M. D. E. C. D. E. Los, *Data Mining Concepts, Models and Techniques*, vol. 12. 2011. doi: 10.1007/978-3-642-19721-5_1.

[47] P. Jaganathan, S. Vinothini, and P. Backialakshmi, "A Study of Data Mining Techniques to Agriculture," *Int. J. Res. Inf. Technol.*, vol. 2, no. 4, pp. 306–313, 2014.

[48] S. Mahdevari, K. Shahriar, S. Yagiz, and M. Akbarpour Shirazi, "A support vector regression model for predicting tunnel boring machine penetration rates," *Int. J. Rock Mech. Min. Sci.*, vol. 72, pp. 214–229, 2014, doi: 10.1016/j.ijrmms.2014.09.012.

[49] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008. doi: 10.1007/s10115-007-0114-2.

[50] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model

to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, 2021, doi: 10.1016/j.asej.2020.11.011.

[51] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, 2022, doi: 10.1016/j.engappai.2022.105151.

[52] S. Carta, A. S. Podda, D. R. Recupero, R. Saia, and G. Usai, "Popularity prediction of instagram posts," *Inf.*, vol. 11, no. 9, 2020, doi: 10.3390/INFO11090453.

[53] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[54] R. E. Schapire, "The Boosting Approach to Machine Learning: An Overview BT - Nonlinear Estimation and Classification," *Nonlinear Estim. Classif.*, vol. 171, no. Chapter 9, pp. 149–171, 2003, [Online]. Available: http://link.springer.com/10.1007/978-0-387-21579-2_9%0Apapers3://publication/doi/10.1007/978-0-387-21579-2_9

[55] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in *IEEE International Conference on Industrial Engineering and Engineering Management*, Feb. 2018, vol. 2017-December, pp. 319–323. doi: 10.1109/IEEM.2017.8289904.

[56] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir

characterisation predictions," *J. Pet. Sci. Eng.*, vol. 208, no. May, p. 109244, 2022, doi: 10.1016/j.petrol.2021.109244.

[57] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[58] R. Kohavi and S. Edu, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proc. 14th Int. Jt. Conf. Artif. Intell.*, vol. 2, pp. 1137–1143, 1993.

[59] R. Forests, "Chapter 15.1 Random Forest for Regression or Classification," pp. 587–604, 2001, [Online]. Available: http://www.math.usu.edu/

الخلاصة

لقد أحدث النمو المتزايد لمنصات الوسائط الاجتماعية ومحتوى الفيديو عبر الإنترنت ثورة في الطريقة التي نتواصل بها ونتفاعل ونستهلك المعلومات. مع مشاركة الملايين من مقاطع الفيديو يوميًا على منصات مثل يوتيوب، أصبح التنبؤ بشعبية مقاطع الفيديو جانبًا حاسمًا بالنسبة للمؤثرين ومسؤولي المنصات والمسوقين. كذلك إن فهم العوامل التي تساهم في التنبؤ ب شعبية مقطع الفيديو يمكن أن يؤثر بشكل كبير على تسويق المحتوى واستراتيجيات منصات يوتيوب.

تتناول هذه الأطروحة التحدي المتمثل في التنبؤ بدقة بشعبية الفيديو على منصات التواصل الاجتماعي، مع التركيز بشكل خاص على منصات يوتيوب. أقترح نهجًا مشتركًا جديدًا يستخدم تحليل البيانات الوصفية (بما في ذلك اختيار الميزات المؤثرة) وتحليل الصور المصغرة لإنشاء نموذج تنبؤ أكثر فعالية. ومن خلال استخراج الميزات ذات الصلة من هذه المصادر، أهدف إلى تحسين دقة التنبؤ بشعبية الفيديو. ولتحقيق هذا الهدف، تم استخدام مجموعة من خوارزميات التصنيف القوية، بما في ذلك دعم آلة المتجهات، تعزيز التدرج، الغابة العشوائية، تعزيز التدرج الشديد وأقرب جار.

تم إجراء بعض عمليات المعالجة المسبقة على البيانات لجعلها مناسبة لخوارزميات التعلم الآلي. وبعد ذلك تم استخراج ثلاث أنواع من الميزات: الميزات النصية، والميزات المرئية، والميزات المرتبطة بالزمن. وقد أدت هذه الميزات المضافة حديثًا، جنبًا إلى جنب مع الميزات الموجودة، إلى إثراء مجموعة البيانات بميزات مؤثرة إضافية قادرة على تعزيز دقة النموذج. من خلال تدريب النموذج باستخدام هذه الخوارزميات وتنفيذ تقنيات استخراج الميزات، تم تحقيق دقة كبيرة، خاصة مع خوارزميات الغابة العشوائية وخوارزميات تعزيز التدرج الشديد او الإضافي. تمت عملية التنبؤ على مرحلتين: التنبؤ ب استخدام الميزات الأصلية والمستخرجة، والتنبؤ بالميزات المستخرجة فقط.

تم تقييم أداء النموذج باستخدام مقاييس مختلفة، بما في ذلك الدقة، والتي تصل إلى معدلات ٩٦٪ و٩٧٪ لخوارزميات تعزيز التدرج الفائق وخوارزميات الغابة العشوائية، على التوالي. تكمن قيمة الجمع بين البيانات الوصفية وتحليل الصور المصغرة في اكتساب رؤية حول العوامل التي تؤثر على شعبية الفيديو. وهذا يمكّن المؤثرين والمسوقين من تخصيص المحتوى الخاص بهم بشكل أفضل، مما يلقى صدى لدى جمهورهم المستهدف ويزيد من احتمالية النجاح على منصات التواصل الاجتماعي.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جـامعـــة بـابـــل
كلية تكنولوجيا المعلومات
قسم البرمجيات

التنبؤ بشعبية الفيديو استنادًا إلى بيانات اليوتيوب الوصفية والصور المصغرة

رسالة مقدمة إلى
مجلس كلية تكنولوجيا المعلومات - جامعة بابل كجزء من متطلبات
نيل درجة الماجستير في تكنولوجيا المعلومات / البرمجيات

من قبل

هبة حسين عبد العباس

بإشـــراف

م.د وضاح رزوقي عبود حسن بيعي

١٤٤٥ هـ ٢٠٢٣م