PREDICTING VIDEO POPULARITY OF STREAMING SERVICES WITH MACHINE LEARNING APPROACHES

SINA SOLEIMANI ZARRIN

BOĞAZİÇİ UNIVERSITY

2023

PREDICTING VIDEO POPULARITY OF STREAMING SERVICES WITH MACHINE LEARNING APPROACHES

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts
in
International Trade Management

by
Sina Soleimani Zarrin

Boğaziçi University
2023

DECLARATION OF ORIGINALITY

I, Sina Soleimani Zarrin, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature:

Date: 03.04.2023

ABSTRACT

Predicting Video Popularity of Streaming Services with Machine Learning Approaches

Live or online video streaming services have gained immense popularity in recent years due to the expansion of the internet and the impact of Covid-19. However, delivering content to an ever-increasing user base poses challenges to online video streaming companies. To respond to user demands and preferences, a prediction model with high accuracy is needed. In this thesis, a predictive model is developed to anticipate a video's popularity as popular or unpopular by applying machine learning algorithms to metadata and textual features of each video from a prominent online video streaming provider in Iran.

The study shows that video popularity can be modeled with high accuracy based on video-related attributes and textual features in Persian language. Four classification models are applied, with the Random Forest model achieving the highest accuracy and F1-Score of 86% and 72%, respectively. The Support Vector Machine obtains the most accurate results when new attributes obtained through NLP are combined with metadata. Moreover, the inclusion of word embeddings of the video description as predictive features improves classifiers' performance significantly.

The study finds that the number of program episodes, video type, channel, and year of production are the most influential features in predicting video popularity. Predicting video content popularity in advance has enormous benefits for marketing purposes, network usage, and network cost reduction.

ÖZET

Akış servislerindeki video popüleritesinin makine öğrenmesi yaklaşımlarıyla öngörülmesi

Son yıllarda internetin genişlemesi ve Covid-19'un etkisi nedeniyle canlı veya çevrimiçi video yayın hizmetleri büyük bir popülerlik kazanmıştır. Ancak, her geçen gün artan kullanıcı kitlesine içerik sunmak, çevrimiçi video yayın şirketleri için zorluklar oluşturmaktadır. Kullanıcı taleplerine ve tercihlerine yanıt vermek için yüksek doğruluklu bir tahmin modeline ihtiyaç vardır. Bu tezde, İran'daki önde gelen bir çevrimiçi video yayın sağlayıcısının her video için metadata ve metin özelliklerine makine öğrenimi algoritmaları uygulanarak video popülerliğinin popüler veya popüler olmayan olarak tahmin edilmesi için bir tahmin modeli geliştirilmiştir.

Tez çalışması, Pers dilindeki video ile ilgili özellikler ve metin özellikleri temel alınarak video popülerliğinin yüksek doğrulukla modellenebileceğini göstermektedir. Dört sınıflandırma modeli uygulanmış olup, Rastgele Orman modeli en yüksek doğruluğu ve F1-Puanını, sırasıyla 86% ve 72% ile elde etmiştir. NLP yoluyla elde edilen yeni öznitelikler metadata ile birleştirildiğinde, Destek Vektör Makinesi en doğru sonuçları elde etmektedir. Ayrıca, video açıklamasının kelime gömbelemelerinin tahmin edici özellikleri olarak dahil edilmesi, sınıflandırıcıların performansını önemli ölçüde artırmaktadır.

Çalışma, program bölümlerinin sayısı, video türü, kanal ve yapım yılı gibi özelliklerin video popülerliğinin tahmininde en etkili özellikler olduğunu bulmuştur. Video içerik popülerliğinin önceden tahmin edilmesi, pazarlama amaçları, ağ kullanımı ve ağ maliyetlerinin azaltılması gibi büyük faydalar sağlarlar.

ACKNOWLEDGMENT

I am pleased to express my sincere gratitude and appreciation to all those who have supported and contributed to the completion of this thesis.

First and foremost, I would like to thank my thesis advisor Prof. Arzu Tektaş, for her invaluable guidance, insightful comments, and unwavering support. Without her encouragement, expertise, and patience, this thesis would not have been possible.

I would also like to thank the members of my thesis committee, Prof. Sona Mardikyan and Prof. Bilge Acar Bolat, for their valuable feedback and constructive criticism, which helped me refine my research and enhance the quality of my work.

I am deeply grateful to my family for their unconditional love, constant encouragement, and emotional support. Their belief in me and my abilities gave me the strength and motivation to persevere during these challenging times.

Lastly, I would like to acknowledge my friends for their encouragement, helpful suggestions, and moral support. Their positive attitude and willingness to listen made my academic journey more enjoyable and rewarding.

Thank you all for your contribution to my academic and personal growth. Your support means the world to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CDN	Content Delivery Network
CV	Cross-Validation
GBM	Gradient Boosting Machine
GC	Grammatical Construction
GWI	GlobalWebIndex
HD	High-Definition
ID	Identification
IMDB	Internet Movie Database
IP	Internet Protocol
IPTV	Internet Protocol Television
IRIB	Islamic Republic of Iran Broadcasting
KNN	K-Nearest Neighbor
ML	Machine Learning
MLP	Multi Layer Perceptron
NB	Naïve Bayes
NLP	Natural Language Processing
RF	Random Forest
SVM	Support Vector Machine
TV	Television
VOD	Video on Demand
WAPE	Weighted Mean Absolute Percentage Error

CHAPTER 1

INTRODUCTION

In this contemporary era, the Internet plays a critical role in every minute of people's daily lives. Its vast rise leads to changes in past traditional ways of dealing with life. Today, the Internet is integrated into every aspect of people's daily lives, including entertainment which has risen significantly. According to DataReportal, more than 5 billion people which is 63 percent of the world's total population now use the internet currently. Internet users have increased at a rate of 4 percent over the year 2021, representing almost 200 million users. Typical users spend more than 40 percent of their awake time online, and based on new research by GWI in Q4 2021, 50.3 percent of internet users state that the reason for going online is to watch videos, TV shows, or movies, which ranks the fourth among the main reasons of Internet usage.

Moreover, increasing internet connection speeds by 40.6 percent in mobile internet connections over the past 12 months (Ookla) leads to the ability to stream 4K videos on the mobile without any problem. As a result, 92.2 percent of internet users watch TV content via streaming platforms, e.g., Netflix, YouTube, and Apple TV, each month in Q4 2021(GWI).

These changes and the development of high-definition, 4K, and 3D videos, which have large data sizes, need more service traffic and have increased it more rapidly following the Covid-19 pandemic.

1.1 Problem statement

Recently, with the challenge of Covid-19, along with the reasons mentioned earlier, streaming video services like YouTube, Netflix, GloboPlay, Amazon Prime, and AppleTV have expanded along with the popularity of the Internet (Sá, Rocha, & Paes, 2021).

Increasing the popularity of HD (high-definition) and 3D technology, followed by the emergence and growth of IPTV platforms all around the world, leads

to making IP video traffic become a significant part of all consumer Internet traffic (Zhu, Cheng, & Wang, 2017).

With the development of these platforms and video services, their managers and operators try to provide high-quality services for their customers to fulfill their needs and interest and keep their quality of experience (QoE) high. QoE is used for measuring the quality of delivered services to end customers and their perception (Huang et al., 2018).

Video streaming platforms generate millions of hours of content and programs that get billions of views daily (daSilva & Winck, 2017). However, apart from the variety and the large number of TV programs that these platforms provide, user attention is not distributed normally for all the content. Just a few attract more than 85 percent of users' attention (Sá et al., 2021; Zhu et al., 2017).

Streaming platforms have complicated structures to deliver thousands of programs to millions of viewers (Sá et al., 2021). Users constantly evaluate the videos and services that are provided in the aspects of subscription cost, content quality, availability, and buffering stability (Huang et al., 2018).

Based on streaming services' properties and structure, traffic domination by popular content is increasing more than ever, and this growth in traffic is the main problem for providers of streaming services (Jeon, Seo, Park, & Choi, 2020). In this regard, there is a significant need for Content Delivery Networks (CDNs) development that can be able to handle this problem effectively (L. Chen, Zhou, & Chiu, 2015).

In general, the demand for video streaming programs is volatile, which means the demand for content becomes high just after they are released but quickly drops down after a certain period of time (Jeon et al., 2020). As a result, there is a vital need for video streaming platforms to manage these volatilities and changes in demand by responding to them proactively, which is provided by operating systematic services and a proper system architecture (J.-B. Chen & Chen, 2012; L. Chen et al., 2015).

Consequently, in order to provide quality network and efficiency in streaming service, which leads to customer satisfaction, these platforms need a solution that categorizes popular content that has gained more views and has also taken more network in advance, meanwhile keeping them in the storage with the most high-speed network accessible (Jeon et al., 2020).

Therefore, identifying what makes a video popular is crucial for online video streaming platforms. It may enable them to deliver better services and reduce information overload by prioritizing popular videos in prominent storage and network (Y.-L. Chen & Chang, 2019).

Developing an early prediction of the popularity of video content is crucial for video content providers, such as online streaming services, VOD providers and IPTV providers, both in terms of marketing as well as network usage (Jeon et al., 2020).

In addition, a reliable program popularity prediction should be able to enhance the entire systems of these platforms by optimizing CDNs and cache strategies (Zhu et al., 2017). This process improves the capability of CDNs to react by providing huge consumer demand for viral contents (Jeon et al., 2020).

Broadcasters, advertisers, and content providers benefit greatly from accurate and timely popularity predictions (Zhu et al., 2017). Contents can be optimally distributed among servers by accurately predicting users' preferences, ultimately reducing network costs (Jeon et al., 2020). On the other hand, with so many variables to consider, it is not easy to predict which videos will be popular in advance (Sá et al., 2021).

Developing strategies for online advertising and commercial marketing requires the ability to predict the popularity of videos on video sharing sites (Y.-L. Chen & Chang, 2019). Predicted popularity values can be used to determine what videos to recommend to users and the order in which to recommend them (Y.-L. Chen & Chang, 2019).

A challenge associated with web content is the ability to accurately forecast the amount of attention it will receive from users (popularity), which in turn promotes faster, more accurate suggestions (Moniz & Torgo, 2019).

By means of a reliable prediction system, participants in the streaming video service industry will have the capability of making decisions more quickly, precisely and objectively (Sereday & Cui, 2017). Furthermore, if a programmed process can generate accurate predictions, they can be applied to enable enhanced targeting in developing programmatic TV platforms (Sereday & Cui, 2017).

1.2 Motivation of the study

There are several practical uses in predicting Web content's popularity, which is beneficial and valuable for content producers and their infrastructure providers, marketers, and end users (Sá et al., 2021). It is worth mentioning some advantages of predicting content popularity as follows:

Firstly, the recommendation system of websites can be developed by utilizing a content popularity prediction system, which means they can recommend the best content to audiences according to their tendency. Despite standard recommendation systems, it is possible to recommend new videos, and video recommendations are no longer limited to old ones. Accordingly, it will be much easier for audiences to find valuable TV shows among all the video content, which will lead to a higher level of satisfaction and retention for users (Y.-L. Chen & Chang, 2019; Fernandes, Vinagre, & Cortez, 2015; Sá et al., 2021; Trzcinski & Rokita, 2017; Zhu et al., 2017).

Second, in commercials, especially in the field of advertising, with program popularity forecasting data, a company can choose TV programs with the most significant potential for advertising (Zhu et al., 2017). Consequently, tools with the capability of video popularity prediction will become highly valuable financially (daSilva & Winck, 2017). For example, In the United States, most of the TV networks' premium advertising inventory is sold at the "up-front" for the whole year (Sereday & Cui, 2017).

Another significant application is identifying the factors which are influential in video popularity. It enables companies to make a comparison and optimal trade-off regarding the cost of producing a TV program and its raised profit (Fukushima, Yamasaki, & Aizawa, 2016). Identifying these factors leads to investments in the programs that have the potential to return on marketing investment (Sá et al., 2021; Tatar, de Amorim, Fdida, & Antoniadis, 2014). Besides, by this perception of influential factors, companies can figure out how to produce more effective videos (Y.-L. Chen & Chang, 2019).

Finally, and most importantly, utilizing a popularity prediction model will allow broadcast TV operators to optimize their network configuration in advance by allocating enough transmission and storage resources to handle and deliver a popular program to users perfectly (Zhu et al., 2017). Such a model helps these platforms proactively allocate network resources by future demand adaptation (Trzcinski & Rokita, 2017). As such, by utilizing these models and successfully predicting users' tendencies, companies can benefit from network cost reduction by arranging contents among servers optimally (Jeon et al., 2020).

To take advantage of prediction models in this regard, a set of videos and programs from the most prominent streaming provider in Iran are examined, and popularity prediction models are implemented

This online streaming video platform is the first Iranian service for live broadcasting and archive of IRIB (Islamic Republic of Iran Broadcasting) programs and the largest professional video platform for Persian language users around the globe. There is more than 600,000 TV content on Telewebion, and thousands of the most up-to-date movies, series, and animations available worldwide without subscription and for free.

Considering the relevance of these functions in real-world environments and the challenges faced by these companies, such popularity prediction models would be beneficial for online video streaming platforms. The goal of this initiative would be to improve customer satisfaction, increase the number of subscribers, gain more

financial benefits from marketing activities, and, finally, provide a more advanced network infrastructure that significantly reduces network costs.

1.3 Research objectives

The main objectives of the research held in this study can be summarized as follows:

i Investigating main prediction methods used in related works and examining their advantages and disadvantages.

ii Identifying compelling features which are more influential in the popularity of video content.

iii Developing a popularity prediction model that has the ability to predict popularity before the content is published on the website.

iv Utilizing Natural Language Processing (NLP) and word embeddings in the prediction model.

v Assessing different machine learning (ML) algorithms for popularity prediction of Telewebion's programs.

1.4 Research questions

Research objectives will be fulfilled by working on a number of research questions.

i Can a TV program's popularity be predicted accurately prior to in advance publishing by using the related metadata?

ii Does the video's description contain information that a machine learning classifier can use to predict popularity or improve the prediction accuracy?

iii Among all of these methods and attributes, which provides higher prediction accuracy?

CHAPTER 2

LITERATURE REVIEW

2.1 Theoretical background

A significant amount of research has been conducted on the popularity of various types of content on the web, such as news, movie markets, advertising, streaming videos, and TV programs. The researchers must gather information about the users, historical usage data, and metadata regarding content for conducting these researches.

In today's world, modern organizations collect vast amounts of data. Data would be valuable to an organization when analyzed to extract insights that can be used to make better decisions. Data analytics is the process of transforming data into insights and making decisions based on those insights (Kelleher, Namee, & D'arcy, 2015).

In predictive data analytics, patterns are extracted from historical data, and models are built and used to make predictions. There are several applications for predictive data analytics: Risk Assessment, Diagnosis, Price Prediction, Medication Dosage Prediction, Tendency Modeling, and Document Classification (Kelleher et al., 2015).

Two facts are common to all of these examples. The first one is that a model is used in each case to provide a prediction to assist the person or organization in making a decision. Typically, the word prediction has a temporal connotation - it refers to predicting what will happen in the future. Generally, a prediction is a value assigned to an unknown variable in data analytics. Prediction can either refer to predicting the price at which something will be sold in the future or predicting the type of document. So, predictions may have a temporal aspect in some cases, but not always.

The second fact in common with the examples is that a model is trained using a set of historical examples to make predictions. In this stage, Machine learning comes into play in order to train and construct these models (Kelleher et al., 2015).

Generally, machine learning is defined as the automated process of extracting patterns from data. Supervised machine learning techniques automatically train a model to identify the relationship between a set of descriptive features and a target feature based on historical examples. On the other hand, ML algorithms capture the relationships between descriptive features and the target in a dataset automatically. Although it is possible in a simple dataset to create a prediction model manually, in a large and complicated dataset with too many input features, it is almost impossible to create a prediction model manually. Consequently, for building prediction models from a large dataset with multiple features, machine learning is the solution (Kelleher et al., 2015).

As a result, based on the concept and function of machine learning algorithms, the characteristics of the data, and the ultimate goal of this study, machine learning algorithms offer the most effective solution to answer the research questions.

Several approaches tackle the problem of predicting web content popularity, including standard classification and regression approaches. These approaches also consider various scenarios of data availability which are influential in predicting popularity before or after the item is published.

2.2 Empirical background

This section presents the literature related to the most applicable methods that are used in popularity prediction. To delineate the scope of the work, the set of methods and techniques are limited to those that use Machine Learning methods to predict the popularity of videos, images, and news articles. Therefore, machine learning-based approaches are reviewed, and their strategies to predict web content popularity are categorized and discussed.

Szabo and Huberman (2010) presents a model for predicting online content's long-term popularity (30 days), including news and videos. They utilize attributes such as view count, vote, and download in their model to predict future popularity based on early measurements of user access. As a first step towards that goal, Szabo

and Huberman (2010) discover a robust relationship between early popularity and log-transformed long-term popularity of content. They claim that their model would predict future popularity by linear regression if only the target and reference dates are known and not any specific information from the content itself. Their model can predict popularity in the first two hours of users' access to news content 30 days ahead with a 10% error and the same for video content after ten days of being published.

A Oghina, Breuss, Tsagkias, and de Rijke (2012) work on predicting movie ratings in IMDB with two sets of features from social media platforms (Twitter and YouTube). They extract two sets of surface and textual features from video-type content. Surface features include views, number of comments, number of favorites, number of likes and dislikes and their fraction, and number of tweets; textual features consist of tweets and comments on YouTube. They use linear regression with the combination of these features, which results in the fraction of likes and dislikes on YouTube, combined with textual features from Twitter, leading to the best performance model.

Another study by Pinto, Almeida, and Gonçalves (2013) predicts YouTube video popularity, based on metadata features through regression. In this study, by using historical information given by early popularity measures, the researchers develop two simple models (Szabo-Huberman (S-H) Model and Multivariate Linear (ML) Model) to predict Web content's future popularity. The popularity of content is derived from daily samples up to a specific reference date. According to their findings, their model is able to recognize videos with different popularity patterns when different weights are assigned to different popularity samples within the analysis period. Results depict a considerable reduction in the average prediction error of the model.

A study by Khosla, Das Sarma, and Hamid (2014) investigates what makes images uploaded by users popular on social media sites. The authors adopt a massive dataset from Flickr and find that image content and social context are two key factors

that affect an image's popularity. The main point of this research is that they use visual features with the advantage of the regression method for view count popularity. As a result, by combining various features, they come up with an approach that obtains more than 0.8 rank correlation in predicting popularity.

Visual features accompanied by the regression method are also utilized in another research by Trzcinski and Rokita (2017), with the difference being in the content type. This time, researchers work on video-type content. The authors suggest a regression model that can predict the popularity of an online video based on the number of views it receives. It is shown that by combining early distribution patterns with social and visual features, popularity prediction accuracy can be improved significantly. Among all these features, social features are the strongest signal when it comes to video popularity prediction. The results indicate that it is possible to predict future video popularity based only on visual features determined before the video's publication. Despite that, if a higher prediction accuracy is desired, temporal features such as view counts or social elements should be added to visual features to produce the best result.

Data-driven analysis of YouTube videos' popularity based on a large dataset is presented in the study by Hoiles, Aprem, and Krishnamurthy (2016). This study investigates how video features (meta-data) and social dynamics affect video popularity. In this paper, researchers, by using contents meta-data, social dynamics, and adopting a regression model, find out that the five most helpful meta-data features in the view count or video popularity are first-day view count, the number of subscribers, the contrast of the video thumbnail, Google hits and number of keywords. Among all these key features, it is interestingly observed that even the title length and the number of upper-case letters in the title are of great importance.

Another study is carried out to predict audience TV drama ratings before they are broadcasted by considering the involved cast and staff features (Fukushima, Yamasaki, & Aizawa, 2016). For this purpose, in addition to basic information such as date and station, they consider actors' popularity based on their Wikipedia page

and related tweets. The audience ratings of a dataset of 678 Japanese dramas are predicted with a correlation coefficient of 0.845. Moreover, about factors, their weights, and their impacts on the result, they discover that a combination of time and station greatly value audience ratings regardless of the content broadcasted.

All of the research summarized above utilize regression models from machine learning algorithms with various attributes on different content types such as images, news, or video content for popularity prediction. It is observed that there are various choices in selecting features for making predictions. These choices include using metadata or, regardless of it, textual features from social media or video itself, visual features, or staff popularity, making this field overly complex and somehow confusing. The publication time of web content and the reference date in the analysis can also affect the final results. Therefore, this wide range of varieties necessitates the utilization of Machine Learning algorithms to analyze and resolve them.

Another machine learning method, classification, is examined in the following studies.

In a study by Fernandes, Vinagre, and Cortez (2015), a binary classification model (popular/unpopular) is proposed to predict news popularity in the future prior to their publication with the vast extracted textual features that are known. These textual features are keywords, digital media content, number of words in the title, number of links, average word length, and many others. In the next stage, optimization is carried out to achieve a better prediction result by identifying the more convenient and accessible features to change. As a result, they improve their mean gain by 15%. In another study, a novel approach is proposed to predict the popularity of online video content shared on Facebook (Trzcinski & Rokita, 2017). Visual features in the video's representative frames are only used in their classification task. This study is conducted on a dataset of over 37'000 Facebook videos. It is seen that this method performs 30% better than traditional prediction approaches, which can be helpful in reaching insights for content creators.

Khan, Worah, Kothari, Jadhav, and Nimkar (2018) implement and compare eleven machine learning models in order to predict the popularity of news articles. Their approach is to predict the news popularity before it is released by using meta-data features of an article, such as time of publishing, the article's sentiment, channel, polarity, length, and keyword data. Textual features are utilized in the eleven classification algorithms to compare the results and reach the most accurate method. This research considers articles with more than 3395 shares as popular. The final result comes up with an accuracy of 79%.

As in the previous study conducted using textual features with a classification method on news content type, Jeon, Seo, Park, and Choi (2020) work with a different content type: video content. This paper proposes a hybrid machine learning approach for popularly predicting (unpublished) video content. Various combination of methods is adopted in this regard. They used meta-data attributes and textual features. The historical data is divided into type A, including previous work, and type B, without it. Structured data, including the data of previous contents, is utilized for type A. In addition, structured / unstructured data, including many kinds of text data such as actor names and keywords of content, is utilized for type B. The reason for separating historical log data and meta-data is that proper algorithms and models are different for each data type. Consequently, better results can be achieved in terms of accuracy.

Y.-L. Chen and Chang (2019) try to predict video popularity at the time of upload by using available data at that time. They propose a new classification model for the future popularity of a video. It uses textual features generated by users (video title, tag, description) and the access data of YouTube videos (thumbs-up count, thumbs-down count, number of comments, and subscriptions). They develop predictive models by adopting a machine learning approach, specifically the ensemble classification model. Experimental results show that this method can reach an accuracy of over 75%. Based on this successful result, indirect variables can be extracted from relevant information to predict future trends for a new object.

A subset of a video data stream is analyzed using a simple data stream mining technique, namely Hoeffding Tree, in the research by daSilva and Winck (2017). Therefore, this work discusses data stream mining concepts, focusing on classification methods and, more specifically, Hoeffding Trees algorithms. By generating a classification model based on features humans cannot interpret, they propose a method for predicting video popularity before uploading to a video-sharing website. They define three categories of features for their model:

- video information and viewing statistics such as time of upload, tag count, video duration, and title length;
- image features such as frame rate, image quality, image ratio, and mean and standard deviation of colors;
- audio features include time domain zeros crossing, spectral roll-off, and Mel-frequency cepstral coefficients.

Sereday and Cui (2017) also examine TV ratings in the US media industry, which are metrics that evaluate past or at least present but are critical and functional to predict the future. They utilize entirely historical data as input for their prediction model. They find that the GBM method (specifically, xgboost optimized library) gives their project the highest level of accuracy and scalability, regardless of its advantages and disadvantages. The accuracy of their models is evaluated using WAPE (weighted mean absolute percentage error), which shows that the model is effective and its results are close to expectations concerning accuracy.

As reviewed in the previous studies, all these studies use classification method to achieve higher accuracy in video popularity prediction. The common characteristics of these studies are that they utilize metadata features and take advantage of textual features to reach higher accuracy. Also, they try to predict the content's popularity before release, which has several advantages in every aspect. Based on their results, classification methods using metadata features and textual features extracted from the content can produce a better outcome.

The literature contains two studies that use both regression and classification methods for the popularity prediction of news content. It is worth noting that both of these studies utilize meta-data attributes, specifically textual features, which are taken into consideration below:

Firstly, Bandari, Asur, and Huberman (2012), by using property-based features derived from articles, construct a multi-dimensional feature space and evaluate its effectiveness as a predictor of online popularity. In the analysis of articles, they look at factors within their content. Accordingly, four features are taken into account: the source of the article, the category, subjectivity in the language, and the name entities. Those features measure the popularity prediction of news on Twitter, and as mentioned, they examined both regression and classification algorithms for comparing results. Despite regression results that are not acceptable, classifiers attain an overall accuracy of 84%. Moreover, they find that the article's source is the most significant predictor of news popularity on the social web.

In the other study, Liu et al. (2017) explore variables influencing news popularity. They develop a novel methodology for predicting the popularity of online news before it is released. One of the most important findings of this study is that the grammatical construction (GC) of the titles is practical on news popularity by precisely 6.62% improvement in the regression, which is introduced for the first time. Besides this feature, the author's grade, publishing time, content length, and the score of categories are used as meta-data for popularity prediction. After all, based on experiments, it is discovered that the author's grade is the most significant feature in news popularity. Along with the regression model result, they check its validity by implementing the classification model.

Given the fact that both methods regression and classification methods have their own pros and cons in different scenarios, in order to benefit from them, some studies apply these two methods for popularity prediction. Some apply them to compare the results and others apply them to support the results. In certain cases,

only one of them performs adequately; whereas in others, they can validate the results of each other.

In conclusion, based on a large number of studies that examine and discuss the problem of video popularity prediction and the related literature in this field, classification is distinguished as the most common method for video popularity prediction. Related studies in the literature also show that in addition to metadata features, there is an opportunity to take advantage of textual features such as video descriptions to achieve better results.

Therefore, considering our company's data characteristics and the importance of video popularity prediction for the company, a classification method using available metadata is implemented. Also, it is attempted to take advantage of textual features to improve the results and make a comparison.

2.3 Foundation

2.3.1 Machine learning

Basically, machine learning refers to an automated process of extracting patterns from data. The purpose of Machine Learning is to give machines the ability to learn from experience and solve problems. The main reason is that an algorithm that follows a step-by-step pattern cannot model and solve every problem. As an example, distinguishing cats and dogs from their pictures is a complex task for a machine, while it is simple for humans. Since standard algorithms cannot be implemented in such situations due to many variables, machine learning methods, by learning from previous examples and experiences, are used to solve such problems and improve performance (Mitchell, 1997).

In machine learning, learning is usually done by identifying a target function to solve the problem. Based on previous data related to the task (the experience), the algorithms produce functions capable of carrying out the task independently. Experiences are called datasets, which consist of examples (individual experiences) and attributes (descriptive variables) (Mitchell, 1997).

The followings are some definitions of ML that have been used commonly in popularity prediction: Dataset: a dataset represents the list of historical instances. Every dataset involves descriptive features that describe the sample and a target feature that indicates the final condition of the sample.

In the context of Machine Learning, every row in the dataset represents a training instance, and the overall dataset is referred to as a training dataset. In this study, the dataset consists of video content and its attributes (Kelleher et al., 2015). Descriptive features: describes the content, whether it was obtained directly or indirectly (through some methods or calculations) (Mitchell, 1997). Predictive features: the features that are used as inputs to ML models. An attribute vector usually represents the entry (Mitchell, 1997). Target feature: It is the desired outcome of the model that is represented by target attributes or outputs, in this study popularity classes (Kelleher et al., 2015). Supervised learning: A supervised Machine Learning technique uses a set of historical examples or experiences to automatically generate a relationship model between a set of descriptive features and a target feature. In other words, the available target feature simulates a supervisor's activity, for instance, someone with knowledge of the correct answers (Cord & Cunningham, 2008).

The function or the model can be learned using a variety of ML strategies. These strategies may be different based on availability and kind of experiences, e.g., an array, input-output pairs, or just input, interrelation with the environment, how the function performs in learning, for instance, rules, probability, determining the exact output; and how these methods explore the environment to estimate the target feature (Mitchell, 1997)..

This study focuses on supervised techniques as defined above based on the type of gathered training instances or descriptive features.

2.3.2 Natural language processing

This area of research investigates how computational agents can communicate with humans by using methods and techniques at the intersection of artificial intelligence

and computational linguistics. The challenging part is that the computers use formal languages such as Java and Python programming with precise sentences and specific syntax to be understandable for machines. However, on another side, human communication is ambiguous and sometimes confusing (Sá, Rocha, & Paes, 2021).

There are two commonly used strategies to extract features from texts to feed ML methods. One way is manually engineering features based on linguistic cues and experts' experience and computing values to those features from the texts. The other way is representing the texts in a vector space relying on the distributional semantics (Harris, 1954).. It is possible to take two approaches in this case. First, the features are defined as the words in the vocabulary, while the values are determined by their frequency, known as bag-of-words. The other approach relies on a probabilistic or neural formulation to develop a language model from an extensive collection of texts (Bengio, Ducharme, Vincent, & Jauvin, 2003; Jurafsky, 2000).

Characters, words, sentences, and documents can be building blocks for language models. By using meaningful words, a language model is illustrated in this study.

The term word embeddings in natural language processing (NLP) refers to how words are represented in text for analysis, usually as a vector that encodes the word meaning in a way that is similar to the meanings of the words that are near each other in the vector space (Jurafsky, 2000). As a result of word embedding, a word with a similar meaning or influence in the sentence is given a similar value for a specific feature.

One-hot encoding is an important method for representing vectors in NLP and word embedding. Binary vectors are used to describe categorical variables in one-hot encoding. Binary vectors represent integer values with zero values except for the integer index, which is marked with a 1. Although it is easy to implement and can perform quickly, it loses the inner sense of the words (Li & Yang, 2018).

For the analysis of textual features, this method is considered in this study.

2.3.3 Popularity prediction

A predictive task can be classified as either a classification task or a regression task. A discrete output is produced in the first case, such as a tumor (benign, cancerous). Unlike the first one, the latter outputs a numerical value, such as temperature. Taking the dataset used and the research context into account, the definition of the popular class in a study differs from one to another (Sá et al., 2021).

Based on this categorization, methods of predicting popularity can be structured in the following ways according to the problem definition and prediction task:

- Regression methods. Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. As a method of predictive modeling, its objective is to predict a numeric value, such as the exact amount of an item's popularity. The most common target attributes in this context are the number of views, shares, tweets, and comments (Sá et al., 2021).

  In the real world, regression is used in many situations, such as forecasting weather conditions or temperatures, predicting sales in a company, marketing trends, forecasting health trends in the general population over time, predicting prices, and many others. It is also used for determining the causal-effect relationship between variables, the most important factor, the least important factor, how each factor affects the other factors, and how strongly the factors impact dependent variables ("Machine Learning Regression Explained", 2021).

  Based on a defined metric, these methods perform numerical predictions to quantify popularity. It is common to call these predictive methods regressors (Trzcinski & Rokita, 2017).

- Classification methods. Classification is defined as the process of recognition, understanding, and grouping of objects and ideas into preset categories. In Classification, a program learns from the given dataset or observations and then

classifies new observations into some classes or groups such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog. Classes can also be called targets/labels or categories (Sá et al., 2021).

There are two most common types of classifications: Binary Classifier and Multi-class Classifier. A binary classifier is used when there are only two possible outcomes to a classification problem such as male or female, benign or malignant. Conversely, a multi-class classifier is used when there are more than two outcomes in a classification problem, such as types of corps or types of music (Brownlee, 2020).

A typical application of classification is filtering email messages into "spam" and "non-spam," which are used by email service providers. Classification methods are also employed in many other real-life examples, such as customer behavior prediction, document classification, image classification, product categorization, image sentiment analysis, customer churn prediction, credit card fraud detection, sentiment analysis, and many others (Kumar, 2022).

2.3.4 Classification models and metrics

This section aims to present ML models that use classification and clarify concepts and definitions of them in the first step and their metrics in the next.

First, the top 5 machine learning algorithms for solving classification problems, including Neural Networks, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor, and Naive Bayes, which are used in this study, are discussed in this section. Moreover, the concept and the theory behind each of them and their applications are also discussed (Gong, 2022):

Neural Network. The neural network is a set of algorithms recognize patterns by replicating the human brain. They label or cluster raw input using machine perception to reflect sensory information. It is possible to classify and cluster using neural networks. When a trained dataset is available, they can classify unlabeled data based on similarities among the example inputs and group them. Additionally, neural

networks can extract features fed to other algorithms for clustering and classification (Bishop, 1994).

There are many potential applications of neural networks and Image recognition is one of the most well-known applications. Neural networks can be trained to detect objects in digital images, which can be used in many different fields, including security and search engines.

Advantages (Mijwel, 2018):

- A neural network can implement tasks that a linear program cannot.
- When an item of the neural network declines, it can continue without any problems with its parallel features.
- Continues learning which means it is designed to learn and improve its results continuously.

Drawbacks (Mijwel, 2018):

- It needed high processing time for big neural networks.
- Data-dependency. The more amount of data is used during training, the more accurate the results are.

Random Forest. As a meta-estimator, Random Forest uses several decision trees to fit a number of subsamples of datasets and uses averages to improve model prediction accuracy and prevent over-fitting. The Random Forest algorithm additionally uses bagging, which allows each tree to be trained on a random sampling of the original dataset and takes the majority vote from each tree (Cord & Cunningham, 2008).

Advantage:

- A Random Forest Classifier reduces the over-fitting of the model and is, in several instances, more accurate than a decision tree.

Drawback:

- Despite the fact that random forests are capable of real-time prediction, they are slow in nature. Additionally, they are challenging to implement and have complex algorithms.

Support Vector Machine. In support vector machines, training data is represented as points in space separated by a wide gap to categorize them into groups. The border that makes this gap is called a hyperplane, which maximizes the distance between data points from different classes. Using that same space, new examples are mapped and predicted to belong to a category based on their position in the gap (Cervantes, Garcia-Lamont, Rodríguez-Mazahua, & Lopez, 2020).

SVMs have a number of applications in several fields, such as Face detection, Text and hypertext categorization, Handwriting recognition, and many more.

Advantage:

- SVM is relatively memory efficient
- An SVM is effective when there is a clear separation between the classes.
- SVM is more effective in high-dimensional spaces.

Drawback:

- SVM algorithm is not suitable for large data sets.
- Noise in the data set (overlapping target classes) degrades SVM performance.
- The SVM classifier works by putting data points above and below the hyperplane, so there is no probabilistic explanation.

K-Nearest Neighbor. K-Nearest Neighbor algorithms represent data points as points placed in an n-dimensional space containing n features. After calculating the distance between two points, unobserved data points are labeled based on the labels of the nearest observed data points.

KNNs are used for pattern recognition, data mining, and intrusion detection. These KNNs are used in real-life scenarios that require non-parametric algorithms.

There are no assumptions about how the data is distributed in these algorithms (Mitchell, 1997).

Advantage (Cunningham & Delany, 2021):

- Easy Implementation
- There is no training period with KNN modeling because the data itself is the model that will be used as the reference for future predictions. This makes it very time efficient.

Drawback (Cunningham & Delany, 2021):

- Does not work well with large datasets
- Does not work well with high dimensions
- Sensitive to noisy data, missing values and outliers

Naïve Bayes. In naive Bayes, conditional probabilities are calculated based on the Bayes' Theorem, and each feature is assumed to be independent of the other. Unlike most machine learning algorithms, Naive Bayes performs relatively well even with a small amount of training data, whereas most require many training data to perform well.

This is one of the simplest and most effective algorithms for creating machine learning models that make rapid predictions. This is used for various purposes, including spam filtering and text classification (Kelleher et al., 2015).

Advantages:

- easy and quick way to predict the class of the dataset, especially in multi-class prediction.
- When the variables are independent, Naïve Bayes is more capable than the other methods
- Requiring less training data

Drawbacks:

- In naive Bayes, features are assumed to be independent. In real life, it is difficult to collect data that involves completely independent variables.
- Whenever a categorical variable falls into a category that was not included in the training set, the model gives it a probability of 0, which makes it impossible to predict.

Second, the evaluation of the model's performance is integral to any machine learning workflow. Classification performance can be measured in many ways. In this process, predictions are made using the trained model on previously labeled data that has not been seen before. In the case of classification, the model is evaluated according to how many correct predictions it makes (Vickery, 2021).

In classification problems, Accuracy, Precision, Recall, and F1-Score are some of the most commonly used metrics:

Accuracy. Model Accuracy can be calculated by dividing the number of correct predictions by the total number of predictions. On the other hand, Accuracy is defined as the ratio of correct predictions to total predictions. In simple terms, Accuracy refers to the number of predictions that the classifier makes correctly. There is a range of 0 to 1 in an accuracy score; a score of 1 indicates an ideal model (Mishra, 2018).

Advantages:

- Easy-to-use metrics
- Easy to understand and relate
- Gives proper effectiveness of model if data points are balanced.

Drawbacks:

- Doesn't take wrong predictions into consideration.
- Probability score is not considered.
- It cannot be understood where our model is making mistakes

Precision. A model's precision measures its ability to identify positive classes correctly. In other words, what percentage of all predictions for the positive class were correct? The false positive rate would be minimized when this metric is used alone to optimize a model. Precision is essential in music or video recommendation systems and e-commerce websites where incorrect results could lead to customer churn and damage the company (Mishra, 2018).

Advantage:

- Precision makes adjustment of error more straightforward

Drawback:

- Recurring measurements cannot improve Precision

Recall. It is defined as the total number of true positives divided by the total number of actual positives for a label. It measures how well the model accurately predicts all positive observations in a dataset. This metric is useful when False Negatives are of a more significant concern than False Positives (Mishra, 2018).

Advantage:

- Suitable for imbalanced data

F1-Score. An F1 score is a metric that measures a model's predictive ability based on class-level performance. F1-Score is the harmonic average of Precision and Recall. In general, the objective of this metric is to compare two classifiers' performance. It ranges from 0 to 1. The F1-score of 1.0 indicates perfect precision and recall, and in the F1-score of 0, either precision or recall is zero (Mishra, 2018). Advantages: Advantage:

- can be used for multi-class/multi-label problems by choosing the average method
- works well on imbalanced data

CHAPTER 3

METHODOLOGY

This chapter provides an overview of the research process. This study aims to "predict video popularity using machine learning approaches." In order to select appropriate features for the model, it is necessary to identify and reveal the features and attributes influencing the popularity of a video. The literature review is conducted to achieve this objective in the previous chapter. As discussed earlier, all the valuable features and attributes used in other studies are mentioned. In addition, it is found that textual features extracted from the title or description of a video or program can have information that is influential to the final popularity prediction results (Bandari, Asur, & Huberman, 2012; Y.-L. Chen & Chang, 2019; Fernandes, Vinagre, & Cortez, 2015; Hoiles, Aprem, & Krishnamurthy, 2016; Jeon, Seo, Park, & Choi, 2020; Khan, Worah, Kothari, Jadhav, & Nimkar, 2018; Liu et al., 2017; Oghina, Breuss, Tsagkias, & de Rijke, 2012). Considering these facts, the available data set is examined to identify features and attributes and extract and prepare them to be used as inputs for the popularity prediction models. The process consists of five main phases as follows:

i Collecting the data

ii Pre-processing the data

iii Extraction of features obtained from the metadata

iv Extraction of new features from existing ones, and

v Modelling and evaluation.

In this procedure, after features extraction from metadata, building an ML model, and observing its results, textual features extracted from the description of videos are added to features in order to compare models' results with different inputs. Figure 1 illustrates all the stages with their sub-stages in this study's methodology.

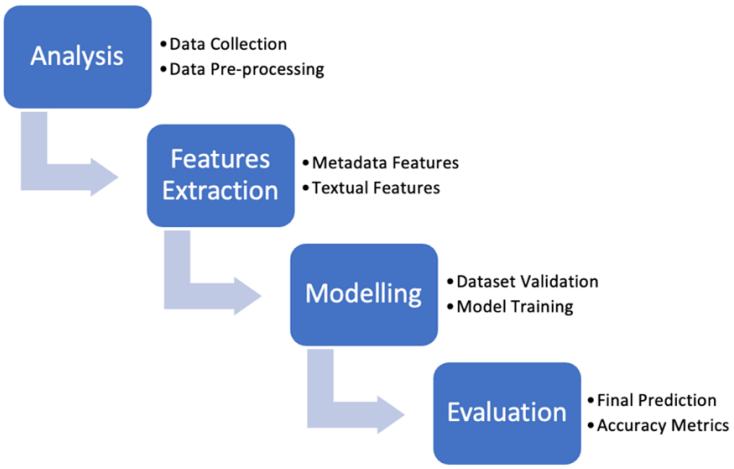

Figure 1. Conceptual model

3.1 Collecting the data

The dataset is acquired from an Iranian online streaming video service company, which is gathered from the streaming service database containing metadata and historical views. A six-month period of 23 August 2021 to 20 February 2022 is the storage period for our data. In light of the company's restrictions on sharing data, this 6-month period not only offers enough samples for the study, but may also cover any seasonality effects created by the data.

The dataset contains all the programs and the related episodes in a specific month separately. Except for the sequel episode count and related view, the features provided in advance by the service company include nine variables as mentioned in detail in the next section.

Machine Learning (ML) models are used to determine whether or not a video will be popular. For this purpose, attributes and features, which are influential in video popularity based on previous studies, are investigated in this dataset.

On average, almost 20000 videos, including their episodes, are contained in this dataset, which is divided into six main categories as movies, series, children, news, cultural and sports programs.

3.2 Pre-processing the data

3.2.1 Preparing the dataset

There has been a wide variety of raw data attributes for every sample, such as program ID, title, video creation date, video tags, episode ID, episode view counts, channel name, country ID, and program view counts. These nine attributes with various and sparse values are accessible in the dataset. It has been necessary in those cases to transform the data before inserting it as a feature in a classification model.

Firstly, the excel formatted, monthly datasets are analyzed and all the videos and programs are combined with their details and attributes into one file in order to work on the whole data more efficiently. At this stage, since there are so many episodes for each program, episodes are grouped by their program ID in order to have one row representing one program in the dataset. Therefore, not to miss any essential attributes, aggregate view counts and aggregate episodes count of all the programs are calculated, extracted, and saved in separate columns.

There are some missing attributes in some of the columns that is shown in Table 2 in detail. They are not removed and kept as in the original data set not to decrease the amount of data for the prediction stage, and not to affect the accuracy level negatively. Therefore, every missing value or null value is refilled with 0 in this dataset. However, this flaw might affect the final results and may be considered as a weakness of the study.

3.2.2 Extraction of features obtained from metadata

An essential stage of this study is selecting features, which involves testing various features and combinations.

A range of features can be applied to predict popularity, from those that describe items' popularity evolution to those that apply sentiment analysis or from feedback and reactions given by users or groups of users in social networks to the difference in time of publication of a video and its first impression at the time of uploading (Moniz & Torgo, 2019). In order to provide a basis for comparing diverse

feature sets used in previous works, the following categorization is proposed (Moniz & Torgo, 2019):

i Behavioral

ii Social-network

iii Content

iv Temporal

v Meta-data

vi External sources

By analyzing the samples and associated features in this dataset, it is possible to identify the following features:

i View count: the number of times that a program and its episodes are watched.

ii Episode count: the number of episodes for each program.

iii Tags: all the tags for each program including the program type, and program genres.

iv Year: the year that the program is produced.

v Channel: the channel from which the program is shown.

vi Country: the country which produced the program.

vii Clips: whether the video is a clip or not and the number of clips for each program.

viii Description: description of the program.

Even though many more features or attributes can be retrieved, these eight features are apparent in the preliminary examination based on the literature. According to Moniz and Torgo (2019) features categorization, the identified features

can be categorized into Content and Metadata groups. So, in this study, they are the features and attributes to be used in descriptive analysis, machine learning, and classification models.

3.2.3 Extraction of new features from existing ones

New features are created and extracted from the Tags column, and textual features are derived from the video's description. The following new attributes are also in the hands of video owners since they can manipulate those. Word embedding is used to extract these attributes from the videos' descriptions. The embedding of words is the transformation of individual words into numerical representations (vectors). It is a dense low-dimensional real-valued vector of a word that is learned from the data (Sá, Rocha, & Paes, 2021).

Firstly, each description is sent to text-mining.ir website to go through a comprehensive process in the Persian language, such as removing stop words, spaces, punctuation, and splitting words. Lastly, the remaining meaningful words present columns for each sample with the value of 1 for existing ones in its description and 0 for non-existing ones. The tags column for each program, as mentioned in 3.2.2, contains the genre and type of the videos. It is divided into two columns: genre and type, which contain information about videos or programs.

As a result, two columns (Genre & Type) plus word vectors are added to the data frame as one of the stages in the data pre-processing.

3.3 Modelling and evaluation

A classification is used to predict a nominal value, such as whether an item is "high" or "low," "hot" or "cold." in terms of its popularity. Popularity, in this context, is the relationship between a given item and its consumers (Sá et al., 2021). Despite the intuitive understanding of popularity, objective metrics are needed to compare two items and determine the most popular one. There are several measures that reveal which online content attracts the highest attention. This study uses the number of

views as a measure of popularity. Models can be categorized according to their distance (KNN), probability (Naive Bayes), ensemble (Random Forest), or function (SVM and MLP). This study attempts to include all these categories for comparison. The following five classifiers are used to determine whether a video will become popular before it is published: MLP, KNN, Naive Bayes, SVM, and Random Forest.

To achieve the desired results, it is necessary to apply data resampling methods such as Cross-Validation (CV) technique to the dataset before the determined features are inputted into the classifiers for training and testing. The purpose of this technique is to test the generalization ability of predictive models and to prevent overfitting different algorithms (Berrar, 2019), and also finding the best algorithm based on the available data by comparing two or more algorithms (Refaeilzadeh, Tang, & Liu, 2009). Generally, it is advisable to estimate the performance of models by using Cross-Validation before running them (Berrar, 2019).

Data resampling methods have various strategies, such as single hold-out random subsampling, K-fold cross-validation, leave-one-out cross-validation, etc. (Refaeilzadeh et al., 2009).

The K-fold cross-validation strategy involves randomly splitting the training dataset into K subsets of approximately equal size that do not overlap (Jurczyk, 2021). The model is trained on the K-1 training sets and then applied to the remaining validation set. The procedure is repeated K times with different training and validation subsets combinations. The average of the K performance measurements on the K validation sets is the cross-validation performance (Berrar, 2019). Since a real-world dataset is analyzed, stratified 10-fold cross-validation is recommended as it is the most effective model selection method with less biased estimation (Kohavi, 1995). Therefore, stratified 10-fold cross-validation is used in this study.

The term stratified in 10-fold cross-validation refers to sampling such that the class or target proportions in the individual subsets correspond to the proportions in the learning set (Berrar, 2019).

After applying cross-validation, the entire dataset is utilized to train the models and then test them. In this regard, as it is typical in machine learning, the data set have to be divided for training and testing purposes. For this purpose, 75% of the dataset are used for training and the rest for test set.

It is critical to split the data over different sets. Due to the fact that the entire data set is sorted by view counts, it is necessary to shuffle the dataset before dividing it into training and test sets and testing the model. Python's Scikit-learn library provides functions to help with this shuffling process and splitting the dataset. In order to make a reliable comparison between models, shuffling and splitting the dataset into training and test set is done once at first before models run. This ensures that, for each experiment, the same randomized training set is used for training, and also the same test set for prediction is used in all models.

In this study, Python-based machine learning frameworks are used to achieve the final results. Several open-source neural network libraries are utilized, including Scikit-Learn 7, and Keras. The neural net model is built using Keras. The experimental results can be evaluated through comparison of the model's predictions and the actual classification of the content.

For evaluating performance, the Accuracy, Recall, Precision, and the F1-score methods are used (See 2.3.4). Considering the imbalance in the dataset, the focus has been on accurate predictions for "hot" or "popular" content classification.

CHAPTER 4

DATA

In this section, the purpose is to provide readers with a general understanding of the statistics related to the platform's videos. After combining six months of data into one file, the dataset contains 109676 videos. It includes programs, their episodes, and their clips that have been uploaded on the platform. By extracting additional information from tags, episodes, and clip columns and aggregating view counts of every episode and clip for related programs, the dataset represents 3051 unique programs and all their essential attributes in separate columns, as shown in Table 1.

Table 1. A Sample of Final Dataset

	ProgramID	Title	ProductionYear	Description	No_Episodes	ProducerCountry	ViewCount	No_Clips	ChannelName	Genre	Type
0	0x1b2c124	کلبه عمو پورنگ	1400	داستان «کلبه عمو پورنگ» درباره فردی به نام «عم...	466	ایران	16862780	376	شبکه 2	انیمیشن	کودک و نوجوان
1	0x1b243e9	اخبار ورزشی 12:45	NaN	NaN	1098	NaN	13046744	866	شبکه 3	NoGenre	خبری
2	0x1b2c0f6	گاندو 2	NaN	مجموعه «گاندو 2» داستانی درباره مسائل امنیتی و...	16	NaN	9650433	0	شبکه 3	پلیسی_معمایی	سریال ایرانی
3	0x1b2ca33	افرا	1399	در سریال «افرا» شخصیت «عقیل» عاشق دختری به نام...	47	ایران	7082197	1	شبکه 1	ملودرام_اکشن	سریال ایرانی
4	0x1b2cbe9	سرجوخه	1399	سریال «سرجوخه» زندگی دو شخصیت اول یعنی «غلامرض...	41	ایران	6889469	0	شبکه 3	پلیسی_معمایی	سریال ایرانی
5	0x1b2ca4e	گلها و لحظات حساس لیگ اروپا 2022-2021	NaN	NaN	514	NaN	5530049	484	ورزش	NoGenre	ورزشی
6	0x1b2437e	اخبار ورزشی 18:45	NaN	NaN	824	NaN	5390003	615	شبکه 3	NoGenre	NoType
7	0x1b2cbbf	رویای فرمانروای بزرگ	2012	NaN	71	کره جنوبی	5340459	0	تماشا	تاریخی	سریال خارجی
8	0x1b2cf0d	کلیپ خندوانه(فصل هشتم)	NaN	فصل هشتم برنامه «خندوانه» برنامه ای شاد و کمدی...	426	NaN	4722435	368	نسیم	NoGenre	سرگرمی
9	0x1b2cf29	پلاک 13	1400	سریال «پلاک 13» داستان خانواده ای بزرگ در یک ع...	74	ایران	4545107	1	شبکه 3	کمدی_خانوادگی	سریال ایرانی

None of the retrieved videos are live broadcasts. This means that only static content uploaded to the website is analyzed. In the dataset, some of the samples (Programs or videos) do not have values in their columns, and there are some missing values in the dataset. Moreover, some videos in the dataset lack descriptions (See Table 2).

However, these flaws in the dataset are inevitable; they provide a ground to compare and measure their influences in the prediction results.

In Table 2, there are 3051 entries representing unique programs, with a total of 11 columns describing each program (samples' attributes). It can be seen from

Table 2. Missing Values

Feature	Available Values	Missing Values	Availability Percentage
Program ID	3051	-	-
Title	3051	-	-
Production Year	540	2514	17.68%
Description	343	2711	11.23%
No-Episodes	3051	-	100%
Producer Country	676	2378	22.13%
View Count	3051	-	100%
No_Clips	3051	-	100%
Channel Name	3051	-	100%
Genre	739	2315	24.19%
Type	2794	260	91.48%

each column count that some columns or features have missing values. For example, in the Genre column, 739 records contain the program's value out of the total 3051 records, indicating that there are no records for the 2315 videos

This is the final dataset used in this study for classification models containing 3051 Telewebion programs and their 11 related features. In the following, each of these elements, except ProgramID and title, which do not have any effect in the analyses, is examined in more detail.

4.1 Descriptive statistics of data

4.1.1 Type

It is important to note that some programs combine different types, so there are twenty unique video types in this dataset. Table 3 presents these data types.

Table 3. Video Types

Science	Social	Kids & Teenager	Wildlife	Local Series
Entertainment	Foreign Movie	Culture, Art & Media	Politics	New Year
Learning	Sports	Economy Knowledge	Religious	News
Music	Documentary	Foreign Series	Health	Local Movie

Some video types have higher frequency than others (see Figure 2). The types of the most videos are Social (16.9%), Religious (14.7%), Documentary (11%), and Culture, Art & Media (8.3%). The minor video types are Politics (0.1%), New Year (0.07%), and Learning (0.03%). However, some types do not exist in the data set (NoType: 8.51%). These videos have flaws in tagging additional information for users in the platform database. However, these videos do not have enough available information; some have high view counts among users.

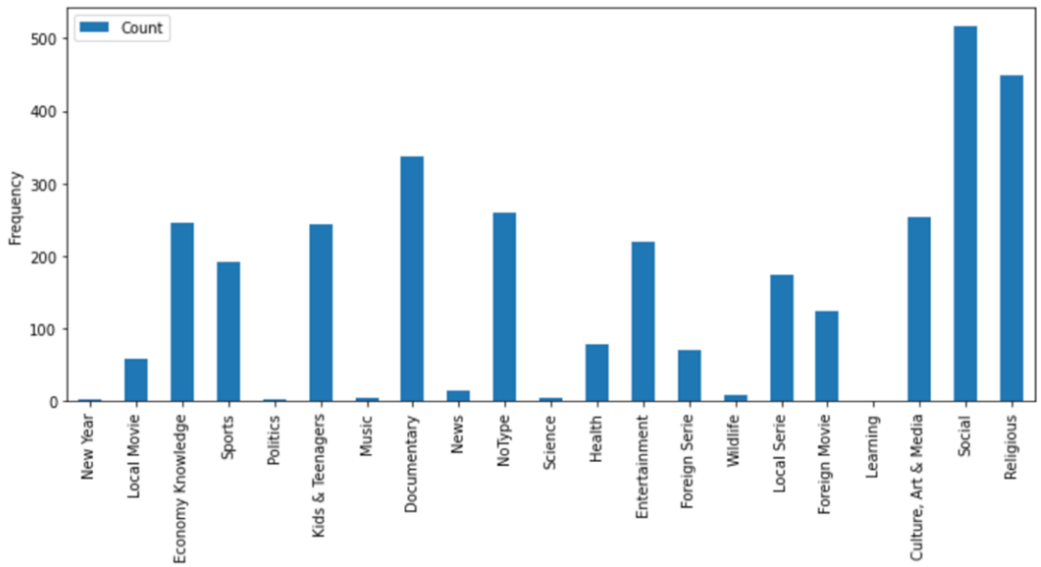

Figure 2. Frequency of the videos by type

4.1.2 Genre

Similar to types, video genres also have one tag or more than one tag with a combination of each other. Despite that, by implementing some processes, it is found that there are 14 unique genres for videos in total. The results are presented in Table 4.

Based on available data, the top three genres with the highest number of videos are: Animation (29.7%), Family (22.7%), and Melodrama (19.5%). On the other side, Fantasy (0.8%), Science Fiction (0.95%), and Romantic (1.6%) are the lowest genres for the videos (see Figure 3).

Table 4. Video Genres

Animation	Documentary	History	Holly Defense	Science Fiction
Family	Comedy	Mystery	Horror	Fantasy
Melodrama	Action	Crime	Romantic	No Genre (Null)

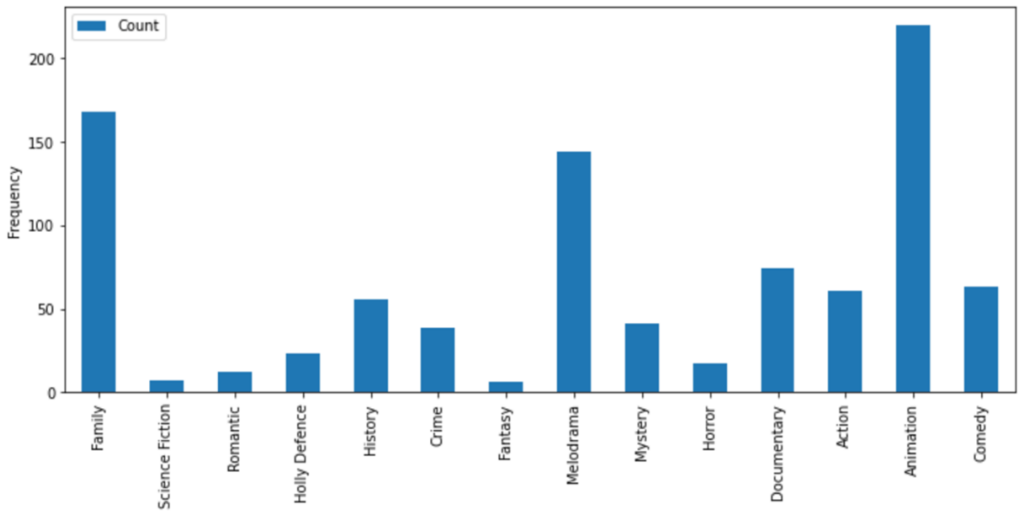

Figure 3. Genre frequency of videos

4.1.3 Episodes

Some programs have episodes available on the platform's website. Throughout this dataset, each program's number of episodes is identified as No_Episodes. It ranges from zero, meaning the program is a movie, to a maximum of 2062 episodes, which belongs to the Daily News program.

4.1.4 Producer country

Other than those samples without information about their producers, the dataset contains programs made by 31 unique countries. As shown in Figure 4, Iran (56.8%), the United States (12.43%), Britain (6.07%), France (3.4%), and South Korea (2.81%) have the highest number of programs, respectively (Figure 4).

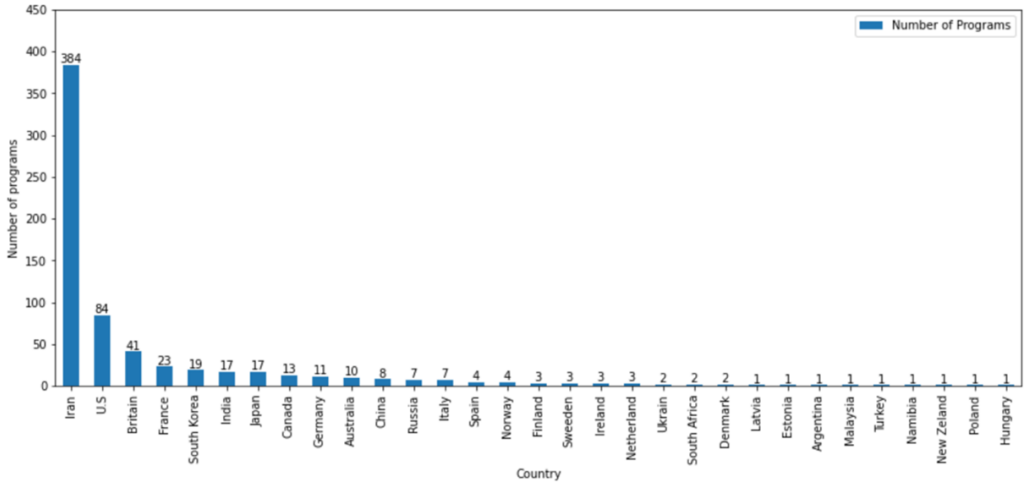

Figure 4. Distribution of programs by countries

4.1.5 Clips

This platform includes a section dedicated to short clips of different programs such as breaking news, funny moments from movies, impressive scenes from shows. This particular feature of a program appears in the No_Clips column in the dataset. There are a variety of values between zero and 1701 which determine how many clips have been cut and uploaded to the website. The Daily News program has the highest number of No_Clips, similar to Episodes. Further, most of the dataset has no value for this attribute, and only 171 samples have values, as it is mentioned in Table 2.

4.1.6 Channels

This dataset contains 25 different channels that broadcast programs and videos. Figure 5 shows how the channel shares of the dataset are distributed. The Shoma channel, with 263 programs, has the most programs, followed by Channel 1, with 222; Omid, with 216; and Channel 5, with 213 programs. Aside from that, there is just one program on the Khouzestan and Radio Telavat channels. In addition, there are four programs on the Ara channel and sixteen on the Sepehr channel at the bottom of the dataset.

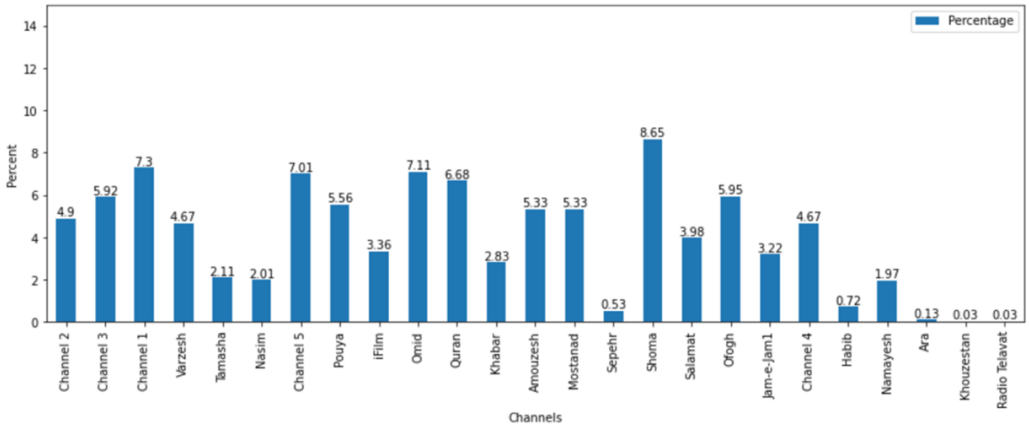

Figure 5. Distribution of programs by channel

4.1.7 Description

One of the most influential attributes of each program is its description, which is used to examine its effect on popularity prediction. However, this feature is the main focus of this study. The number of available descriptions of whole samples is too low in the database and website. Among all 3051 programs, there are just 333 programs with descriptions. The minimum number of words for these descriptions in the dataset is 12, and the maximum number of words is 67, with a mean of 39.26 words per program's description, which is extracted by exploring them.

4.1.8 View counts

The most viewed video has 16,862,780 views for the "Amoo Purang" program. Based on dataset analysis specifically for program views, resulting in Figure 6, Tables 5 & 6, it is evident that few videos tend to attract the most attention and demand from customers. This means these few popular video streaming contents consume most of the network resources. For example, the videos with more than 150,000 views represent just over 9% of all videos, referring to 277 videos out of 3051. Additionally, the sum of this class's view provides impressive information that this class accounts for 85.63% of all the view counts, as shown in Figure 6.

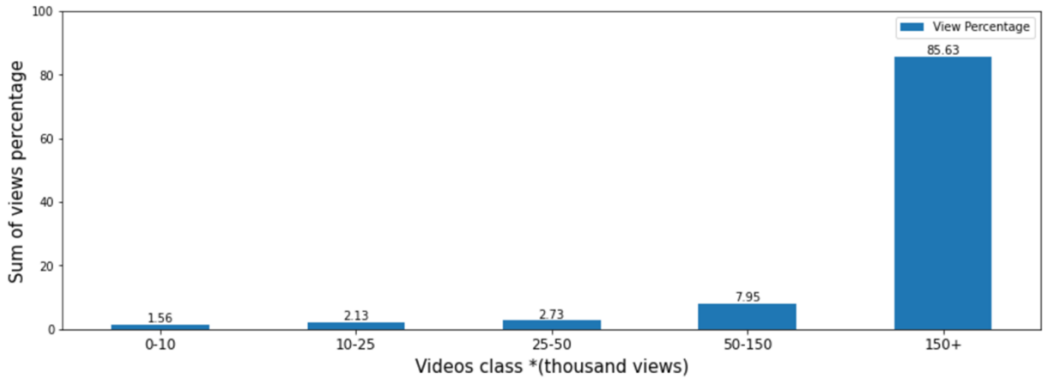

Figure 6. Percentage of total views by view class

Table 5 illustrates that 277 videos containing more than 150000 views (class 4) constitute almost 85% of the total views in the dataset. The second impressive fact is that adding the videos of classes 2 and 3 (25k-150k views) and class 4 (more than 150k views) indicates that 25.6% of videos in the dataset, which refers to 782 videos, accounts for 96.3% of the total number of views.

Table 5. View Classes Details

	Views	Videos	Video Percentage	View Percenatge
0	0-10	1873	61.39	1.56
1	10-25	395	12.95	2.13
2	25-50	232	7.60	2.73
3	50-150	273	8.95	7.95
4	150+	277	9.08	85.63

Besides, the quartiles of the view counts of videos are also measured, with the third quartile equal to 26,007 (See Table 6). In result, this means only 25% of the videos have more than 26,007 views (accounts for 96% of view-counts), while 75% of the rest have less than 4% of the total number of views.

The distribution of the contents and related view counts reveals the criterion for determining the threshold of labeling the popular and unpopular contents.

Thus, predicting videos with more than 26k views, equal to the third quartile, implies predicting videos that gain more than 96% of users' attention; on the other

Table 6. Descriptive Statistics of View Counts

	Descriptive statistics
Count	3051.00
Mean	97311.64
std	572927.36
min	0.00
25%	859.00
50%	4974.00
75%	26007.00
max	16862780.00

hand, videos that occupy 96% of the infrastructure and network of streaming services. For this reason, the popularity class in the experiments is defined by the third quartile value.

As the binary classification is used for the prediction in this study, videos or programs with more than 26000 views are considered popular content with the value of 1, and those with fewer views are considered unpopular content with the value of 0.

The boundary is set at the top 25% (third quartile) to define popular and unpopular content. Accordingly, this dataset contains 25% of popular and 75% of unpopular videos, and the objective is to find 25% of the most popular videos.

CHAPTER 5

FINDINGS

This chapter presents three experiments with various combinations of samples and features. Five classification models are utilized to predict the class of views (popular and unpopular) related to the prepared dataset. Online content that attracts the most attention can be identified through several measures. The number of views is used as a metric of popularity in this study due to the fact that quite several other research considers view count as the most important metric related to popularity among others, such as comment count, like count, and dislike count (Chatzopoulou, Sheng, & Faloutsos, 2010).

The data are collected as described in the methodology chapter, resulting in a data set of 3051 Telewebion videos to be used in the classification models. As mentioned in section 3.4, the data set split into training and test set. The details of train and test sets are shown in Table 7.

Table 7. Allocation of Samples to Training and Test Sets

N = 3051	n	Percentage
Training data set	2288	75%
Test data set	763	25%

5.1 Experiments

In order to evaluate the effectiveness of these models related to the popularity prediction of the streaming platform's videos, three experiments are carried out. In the first experiment, all the samples (3051 video contents) with their metadata obtained from those extracted features from the raw dataset described in section 3.2, without textual features, are utilized as input. In the second, the predictive features are the sample's metadata plus the word embeddings of descriptions for only those videos that contain the description. This means, 333 samples with related attributes (metadata and textual features) are added to the models. In the third experiment, all

the samples (3051 video contents) and features concatenate. In other words, the features extraction from the metadata are combined with the word embeddings of descriptions for all the samples.

5.2 Results

The experimental modeling uses five classifiers and four quantitative metrics to analyze the results. Four metrics are applied to each classification result to evaluate it. The Scikit-learn library is used to implement the entire implementation in Python. Feature extraction results in two datasets. The first dataset contains all samples, and the 11 predictive metadata attributes, excluding the Title, ProgramID, and Program's description. Due to the lack of available and reliable libraries for Persian language processing in Python, the one-hot encoding method is utilized in this study. Therefore, the other dataset includes related metadata attributes and one-hot encodings of descriptions for only samples that contain the description.

In this stage, the dataset is validated by the Stratified 10-Fold Cross-Validation technique, and its results for these three different input features and experiments are shown in Table 8.

Table 8. Stratified 10-Fold Cross-Validation Results

	Ex1		Ex2		Ex3	
	Mean accuracy	Mean std	Mean accuracy	Mean std	Mean accuracy	Mean std
MLP	0.83	0.018	0.81	0.050	0.83	0.017
SVM	0.81	0.014	0.87	0.050	0.81	0.015
KNN	0.84	0.017	0.86	0.062	0.84	0.019
NB	0.81	0.018	0.75	0.098	0.81	0.015
RF	0.86	0.021	0.86	0.080	0.86	0.018

Across all experiments, the cross-validation results indicate adequate average accuracy and insignificant standard deviations. These results suggest that the models have a reasonable degree of generalization when dealing with new unseen data. Furthermore, Random Forest performed better than the other classifiers in the first

and third experiments. However, Support Vector Machine performed better in the second experiment.

It is worth mentioning that the performance differentiation among all classifiers is not very significant and probably will not have a significant difference in their final results.

5.2.1 Experimenting all samples with their metadata features

For the first experiment, the first dataset and the five classifiers are used to test the performance of the metadata features in predicting popularity. As a baseline, this experiment is used in the analysis. A summary of the results can be found in Table 9.

The results of the first experiment are gathered in Table 9. As mentioned earlier, cross-validation results indicate that the RF model performs better than the other classifiers in this experiment. Its highest Accuracy, F1-Score, Recall confirms the validation analysis. Following RF, MLP and KNN classifiers perform almost the same with minor differences in their metrics such as Accuracy, Recall, and F1-Score. On the other side, SVM suffers the lowest results in Accuracy, Recall, and F1-score.

Since this experiment is the base experiment for the study analysis, it can be concluded that regarding input features' characteristics, which are just metadata of the videos, and the number of samples (3051 samples), based on the results, this model is able to predict video popularity with acceptable accuracy on unseen new samples.

Table 9. Classification results with metadata attributes

	Precision	Recall	F1-Score	Accuracy
MLP	82.3	48.69	61.18	84.53
SVM	77.94	27.75	40.93	79.95
KNN	77.95	51.83	62.26	84.27
NB	65.31	50.26	56.8	80.87
RF	77.01	70.16	73.42	87.29

In addition, another output that can be derived is the five most compelling features of RF, as shown in Figure 7.

Examining these results shows that the number of episodes, the type of video, where the video is played (Channel), and the number of video clips for a program affects more than other input features on the final results in the RF model. Therefore, as a preliminary conclusion based on this figure, it turns out that the platform's customers are inclined to programs that have more episodes or, in other words, long series. Moreover, they show interest in programs from specific channels.

From the point of view of video providers, the type of video tagged in the program's metadata, which is in hand for them, has the second rank importance in popularity prediction of a program. Besides, they can benefit from creating more clips for their programs to attract more viewers; as can be seen, this feature has a prominent effect on attracting users' attention.

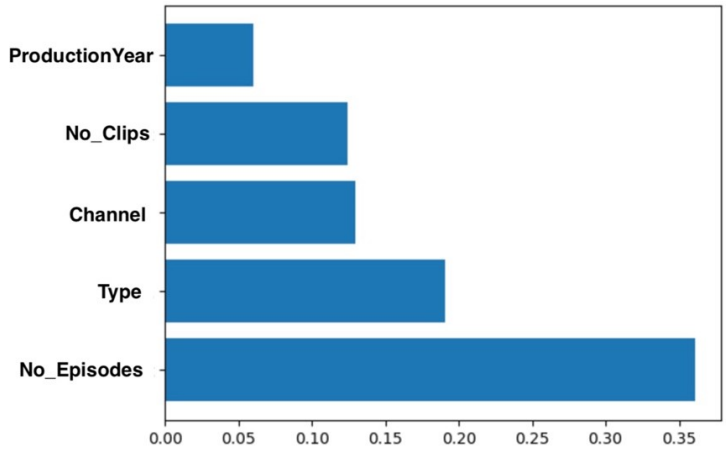

Figure 7. Top five most influential features in RF model

5.2.2 Experimenting samples with their metadata & word embeddings of description

In the second experiment, in order to examine the effect of textual features, word embeddings of videos' descriptions are added to the input features, and those videos without descriptions are dropped from the dataset. Consequently, in this experiment, the dataset consists of 333 samples along with their metadata and word embeddings of descriptions.

The results of five classifiers for this experiment are shown in Table 10. Based on the cross-validation results for the second experiment, SVM, with an accuracy of 87% on average, is expected to perform better among the other classifiers. However, the RF classifier achieves the highest performance overall, SVM and other classifiers revealed significant improvement. Given the fact that accuracy scores for all of them have suffered a reduction a bit, significant improvements result in other metrics such as precision, recall, and F1-Score in all classifiers, which is more meaningful in this case study.

According to the concept of metrics used in this study (Precision, Recall, F1-Score, and Accuracy), improvement in precision in comparison to the base experiment indicates that this model is able to predict more popular videos correctly and reduce the rate of unpopular videos predicted as popular which negatively affects resource management in the video streaming service's infrastructure. On the other side, higher recall scores imply that the more popular videos are identified as popular among all the actual popular videos, which is also desirable for this case study.

Finally, compared to the base experiment, the F1-Score (harmonic mean of precision and recall) also indicates a significant improvement. It means that these results support the fact that textual features and metadata features are better at building the predictive model and positively impact the final prediction results.

Table 10. Classification results with word embeddings and metadata attributes

	Precision	Recall	F1-Score	Accuracy
MLP	73.81	72.09	72.94	72.62
SVM	77.5	72.09	74.7	75
KNN	77.5	72.09	74.7	75
NB	83.33	46.51	59.7	67.86
RF	75.51	86.05	80.43	78.57

Regarding the influential factors or attributes in prediction results, the top five effective features are shown in Figure 8. Notably, the number of episodes, video type, and channel influence the final results similar to the base experiment; however, their contribution to the results has significantly decreased due to the words' vector

existence and their collaboration in the prediction process. As seen in the following rankings, the words "Film" and "Series" demonstrate that textual features effectively predict. Considering that the words "Film" and "Series" are the most common ones in the descriptions and the one-hot encoding method was used to analyze the textual features, these words become highly important in the prediction process.
Accordingly, this is one of the flaws of one-hot encoding method, which does not include the sentiments of the text, and a limitation of this study regarding the Persian language analysis.

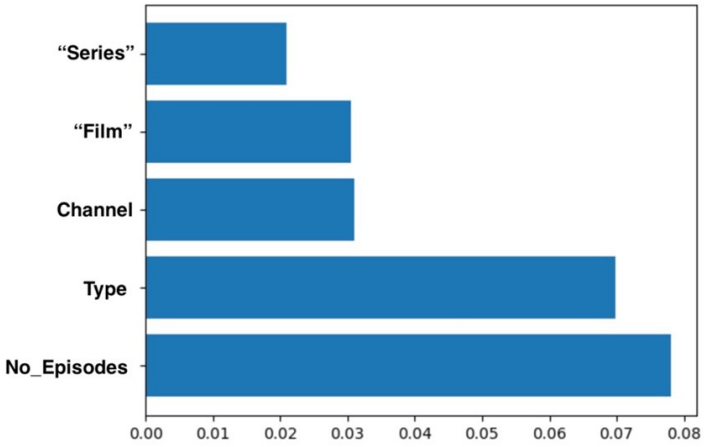

Figure 8. Top five most influential features in RF model

5.2.3 Experimenting all the samples with all metadata & textual features
In the third and final experiment, two datasets are concatenated. In this experiment, all the samples, whether they have descriptions or not, are analyzed with their metadata features and word embeddings of those that contain the description.

The final results are shown in Table 11. Based on the evaluation results, the first impression is that all the classifiers perform satisfactorily in all the metrics. For instance, KNN achieved the highest score in precision among the others. In the following, SVM, MLP, and RF also have high precision, meaning that most classifiers classified popular videos correctly at 74% on average. Suppose recall and F1-Score are taken into account. In that case, it figures out that scores can be

acceptable but less reliable than a trustworthy model. Although the accuracy scores are high, video popularity prediction requires higher performance in metrics such as precision and recall, which its efficiency reflects in the F1-Score metric.

Apart from all these conclusions, the RF model performs way better than other classifiers, as it is expected based on cross-validation analysis with an accuracy of 86% and F1-Score of 72.33%.

Compared to the base experiment, it is interesting to note that none of the classifiers or metrics differ significantly in their scores. Some of them improve a little, while others perform poorly a bit. Based on this comparison, first, the results are almost the same as the base experiments, and second, textual features do not significantly contribute to the prediction popularity. Considering the fact that just 333 samples retain descriptions and the rest are null values, this might be the reason for the resulting similarities in these two experiments.

In conclusion, based on the essential characteristics of the third experiment, such as the number of samples and the minority of textual features, adding textual features to the model does not indicate any meaningful performance improvement.

Table 11. Classification results with all samples and all attributes

	Precision	Recall	F1-Score	Accuracy
MLP	72.44	59.16	65.13	84.14
SVM	77.94	27.75	40.93	79.95
KNN	80	50.26	61.74	84.4
NB	64.86	50.26	56.64	80.73
RF	75.86	69.11	72.33	86.76

Accordingly, as expected, the top five effective features in RF are the same as in the base experiment. The number of episodes, the channel that broadcasts programs, the number of clips, and the type of video influence users' attention more than other features in this study (See Figure 9).

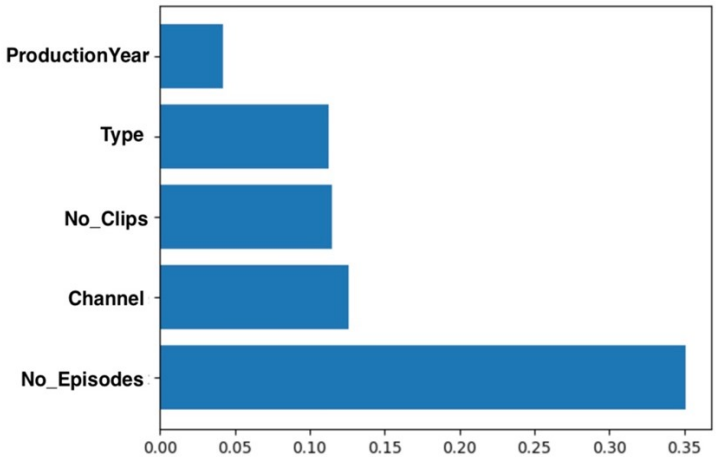

Figure 9. Top five most influential features in RF model

CHAPTER 6

CONCLUSION AND SUGGESTIONS

Intending to predict the video popularity of streaming services, this study examines existing research and presents a real-life case study using machine learning to predict popularity. In order to determine which content catches users' attention on the platform website, several strategies and models are reviewed, developed and implemented. Moreover, discussion of the theoretical foundations related to the concepts, algorithms, methods, and results are provided.

Classification models, such as MLP, KNN, Naïve Bayes, SVM, and RF, and metrics including Accuracy, Precision, Recall, and F1-Score, are applied according to the attribute and data characteristics available. In addition, NLP is used to extract features from textual data in order to strengthen the model. Among the word embedding methods, One-hot encoding is adopted based on the limitations of trained languages libraries and packages in Python. Despite its ease of implementation and speed, it loses the inner meaning of words due to this process.

Models used in this study, take advantage of both textual information and metadata provided by the website publishing the content.

The following sections provides the discussion and evaluation of the results of the models. It outlines the significant findings of the research, followed by a discussion of their possible implications. Afterward, the research limitations identified during the course of this study are presented and explored. As a final point, suggestions are made for future research.

6.1 Main findings

Findings show that using metadata related to a TV program enables us to predict its popularity. The methods used in this study led to an average accuracy of 83%, with the highest F1-Score of 73% in RF. Regarding the characteristics of the descriptive features and target feature, the results show that it is possible to predict the popularity of a video with an acceptable and reliable accuracy before it is published.

Results of the empirical study also demonstrate that the video's description and textual attributes extracted through Natural Language Processing methods and word embeddings can provide important information for classifier models that lead to more accurate predictions in terms of identifying popular videos and the rate of its correctness. Consequently, other elements of a video, such as its title or caption, can be examined and analyzed in the future.

This study attempts to utilize as many different classifiers as possible for popularity prediction in order to compare them and find the most accurate performing classifier. Considering the results of the study, the Random Forest classifier perform significantly better than other classifiers in all three experiments based on accuracy and F1-Score's results. Therefore, it can be concluded that the Random Forest classifier can be a good option when the metadata is accessible as input, and the dataset is not too big. In Addition, when textual features come to play for popularity prediction, RF can take advantage of textual features well and increase the performance by 7% in F1-Score.

Besides, by implementing cross-validation on the data in the first step, the generalization and reliability of the models were measured. Cross-validation results with high mean accuracy and low mean Standard Deviation approved that this predicting model satisfies the generalization and reliability of the model in confronting new unseen data.

By considering the aim and findings of this study, this video streaming company, as well as other similar companies, in order to enhance the platform performance, increase the quality of the user's experience, and expand its market even on an international scale, should establish a standard for the company or even content providers to follow and fill out and determine every descriptive feature for each program. Besides, since a video's textual features significantly impact its popularity, textual features such as descriptions should gain more attention before publishing their content. It is not only beneficial for further analysis but also, based on the results of this study, that manipulating these attributes can affect the video's

exposure. Moreover, incorporating machine learning techniques into its structure enable it to optimize infrastructures, storage, and network management, resulting in more appealing content with a shorter buffering time. It would also help to add interactive elements to the website so that users can give feedback by liking, disliking, commenting, or sharing specific videos and programs.

The Entertainment & Media industry contributes significantly to the local economy. As a result of increasing globalization, this industry has experienced significant changes toward being more global that have led to a more competitive market. The high level of competition has caused uncertainty in the market. Therefore, consumers have a potentially unlimited number of film and streaming service choices. Consequently, companies need to increase quantity and quality to attract and retain as many customers as possible to increase value. In order to achieve this objective, video streaming services are adopting advanced technologies, such as machine learning. Due to the application of advanced technologies in the E&M industry, the role of this industry in the global economy has increased. This is because the E&M industry is one of the major drivers of globalization, and major players of this industry have been expanding globally to attain competitive advantages.

As a result of economic developments various countries, as well as increasing use of electronic and digital means such as internet, the entertainment industry is expanding along with rising demand. This video streaming service company, as well as any similar companies, can benefit from this study's results to gain competitive advantage concerning identifying and distributing popular content more effectively than their competitors and stay ahead of trends at local and global levels. By utilizing popularity prediction models, it can be possible to select the programs that may receive high attention for different segments based on characteristics such as age, gender, education level as well as culture. These may lead to higher demand and sales in local and global markets and increase their revenue. Referring to the company example analyzed in this study, it is evident that users demand foreign films, and the

company should import foreign films to retain users and fulfill their needs. So, for this company, the opportunity to purchase films with features similar to those of popular films is provided.

Overall, video popularity prediction can have significant contributions to the Media Entertainment industry by improving content creation, marketing strategies, advertising, competitive advantages, user experience, content distribution, revenue generation, and decision-making.

Furthermore, applying popularity prediction models makes the company's network more efficient at allocating storage and sources among networks. For instance, by understanding which videos are likely to be popular, CDNs operators can ensure that they have enough resources available to deliver those videos to end-users while freeing up resources for less popular content. Predicting the popularity of videos can help companies in capacity planning and delivering content quickly and efficiently to end-users. Also, it can help in network optimization by managing traffic usage for popular videos. On top of all that, the benefits of all these advantages and actions ultimately result in significant cost reductions for streaming services.

As companies benefit from these opportunities, this would facilitate the development and further globalization of the entertainment industry.

6.2 Limitations and suggestions

During the course of this research, some limitations are encountered: Access to the data has been limited to six months and the period has covered the Covid-19 outbreak. Having access to longer period of data with low levels of missing values can improve accuracy and covering an ordinary time period with no pandemics can provide results that can be generalized with higher reliability. Furthermore, Telewebion is an Iranian platform with only local users which may affect generalization of results adversely. Finally, the lack of NLP packages for Persian in Python constrains the utilization of more advanced NLP techniques.

In future research, textual features can be combined with visual features, such as thumbnails from videos, to predict video popularity.

Furthermore, it is possible to explore new methods for calculating the word embeddings of descriptions in order to investigate other aspects of textual feature's effectiveness. Applying more advanced word embedding methods, such as the Continuous Bag-of-Word, sentiment analysis of the description.

Besides description, even the title of the video or video's comments can be examined by NLP techniques such as sentiment analysis to investigate their influence.

Building and developing language packages and libraries in Persian language would be another value adding action which enable researchers to extract sentiments from Persian language in their future works.

A further analysis could be creating a model incorporating information from social networks such as the number of shares in Facebook or the number of tweets related to a program in Twitter.

REFERENCES

Bandari, R., Asur, S., & Huberman, B. (2012). *The pulse of news in social media: Forecasting popularity. Proceedings of the International AAAI Conference on Web and Social Media,* 6, 26-33. Doi: 10.1609/icwsm.v6i1.14261

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2023). A neural probabilistic language model. *Journal of Machine Learning Research*.

Berrar, D. (2019). Cross-validation. doi: 10.1016/B978-0-12-809633-8.20349-X

Bishop, C. M. (1994). Neural networks and their applications. *Review of Scientific Instruments, 65,* 1803-1832. Doi: 10.1063/1.1144830

Brownlee, J. (2020). 4 types of classification tasks in machine learning. *Machine Learning Mastery*. https://machinelearningmastery.com/types-of-classification-in-machine-learning/

Cervantes, J., Garcia-Lamont, F., Rodriguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing, 408,* 189-215. doi: 10.1016/j.neucom.2019.10.118

Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010). A first step towards understanding popularity in youtube. *2010 INFOCOM IEEE Conference on Computer Communication Workshops,* 1-6.

Chen, J.-B & Chen, C.-C. (2012). Using particle swarm optimization algorithm in multimedia cdn content placement. *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming.* doi: 10.1109/paap.2012.15

Chen, L., Zhou, Y., & Chiu, D. M. (2015). Smart streaming for online services. *IEEE Transactions on Multimedia, 17(4), 485-497.* doi: 10.1109/tmm.2015.2405343

Chen, Y.-L., & Chang, C.-L (2019). Early prediction of the future popularity of uploaded videos. *Expert Systems with Applications, 133,* 59-74. doi: 10.1016/j.eswa.2019.05.

Cord, M., & Cunningham, P. (2008). *Machine learning techniques for multimedia.* Berlin, Heidelberg Springer Berlin Heidelberg.

Cunningham, P., & Delany, S.J. (2021). K-nearest neighbor classifiers – a tutorial. *ACM Computing Surveys, 54,* 1-25. Doi: 10.1145/3459665

DaSilva, V., & Winck, A.T. (2017). Video popularity prediction in data streams based on context-independent features. *Proceedings of the Symposium on Applied Computing,* 95-100.

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. *Semantic Scholar*. doi: 10.1007/978-3-319-23485-4_53

Fukushima, Y., Yamasaki, T., & Aizawa, K. (2016). Audience ratings prediction of tv dramas based on the cast and their popularity. 2016 IEEE Second International Conference on Multimedia Big Data (BigMM). doi: 10.1109/bigmm.2016.24

Gong, D. (2022). Top 6 machine learning algorithms for classification. *Medium*. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Harris, Z. S. (1954). Distributional Structure. *WORD,* 146-162. doi: 10.1080/00437956.1954.11659520

Hoiles, W., Aprem, A., & Krishnamurthy, V. (2016). Engagement and dynamics and sensitivity analysis of youtube videos. *arXiv preprint arXiv: 1611.00687.*

Huang, R., Wei, X., Gao, Y., Lv, C., Mao, J., & Bao, Q. (2018). Data-driven qoe prediction for iptv service. *Computer Communications, 118,* 1965-204. doi: 10.1016/j.comcom.2017.11.013

Jeon, H., Seo, W., Park, E., & Choi, S. (2020). Hybrid machine learning approach for popularity prediction of newly released contents of online video streaming services. *Technological Forecasting and Social Change, 161,* 120303. doi: 10.1016/j.comcom.2017.11.013

Jurafsky, D. (2000). *Speech and language processing.* Pearson Education India, Noida.

Jurczyk, T. (2021). Clustering with scikit-learn in python (A. Wermer-Colan, Ed.). *The Programming Historian.* doi: 10.46430/phen0094

Kelleher, J. D., Namee, B. M., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. The Mit Press.*

Khan, A., Worah, G., Kothari, M., Jadhav, Y. H., & Nimkar, A. V. (2018). News popularity prediction with ensemble methods of classification. *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT*). doi: 10.1109/icccnt.2018.894095

Khosla, A., Das Sarma, A., & Hamid, R. (2014). What makes an image popular? *Proceedings of the 23rd International Conference on World Wide Web.* WWW '14, 867-876.

Kohavi, R. (1995). *A study of cross0validation and bootstrap for accuracy estimation and model selection* (Vol. 14).

Kumar, A. (2022). Classification problems real-life examples. Data Analytics. https://vitalflux.com/classification-problem-real-world-examples/

Li, Y., & Yang, T. (2018). World embedding for understanding natural language: A survey. *Guide to Big Data Applications,* 83-104.

Liu, C., Wang, W., Zhang, Y., Dong, Y., He, F., & Wu, C. (2017). Predicting the popularity of online news based on multivariate analysis. *2017 IEEE International Conference on Computer and Information Technology (CIT).* doi: 10.1109/cit.2017.36

Machine learning regression explained. (2021). Seldon. https://www.seldon.io/machine-learning-regression-explained

Mijwel, M. M. (2018). Artificial neural networks advantages and disadvantages. *Linkedin.com.* https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/

Mishra, A. (2018). *Medium.* https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Mitchell, T. M. (1997). *Machine learning.* Mcgraw Hill.

Moniz, N., & Torgo, L. (2019). A review on web content popularity prediction: Issues and open challenges. *Online Social Networks and Media, 12*, 1-20. doi: 10.1016/j.osnem.2019.05.002

Oghina, A., Breuss, M., Tsagkias, M., & de Rijke, M. (2012). Predicting imdb movie ratings using social media. *ECIR 2012: 34th European Conference on Information Retrieval. Springer-Verlag, Barcelona, Spain: Springer-Verlag,* 503-507.

Pinto, H., Almeida, J. M., & Goncalves, M. A. (2013). Using early view patterns to predict the popularity of youtube videos. *Proceedings of the sixth ACM International Conference on Web Search and Data Mining* – WSDM '13. doi: 10.1145/2433396.2433443

Rafaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database Systems,* 532-538. Doi: 10.1007/978-0-387-39940-9_565

Sá, S. L. d., Rocha, A. A. d. A., & Paes, A. (2021). Predicting popularity of video streaming services with representation learning: A survey and a real-world case study. *Sensors, 21(21),* 7328. doi: 10.3390/s21217328

Sereday, S., & Cui, J. (2017). Using machine learning to predict tv ratings. *Nielsen Journal pf Measurement, 1,* 3-12.

Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communication of the ACM, 53,* 80. doi: 10.1145/1787234.1787254

Tatar, A., de Amorim, M. D., Fdida, S., & Antoniadis, P. (2014). A survey on predicting the popularity of web content. *Journal of Internet Services and Applications, 5*. doi: 10.1186/s13174-014-0008-y

Trzcinski, T., & Rokita, P. (2017). Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia, 19,* 2561-2570. doi: 10.1109/tmm.2017.2695439

Vickery, R. (2021). 8 metrics to measure classification performance. *Medium.* https://towardsdatascience.com/8-metrics-to-measure-classification-performance-984d9d7fd7aa

Zhu, C., Cheng, G., & Wang, K. (2017). Big data analytics for program popularity prediction in broadcast tv industries. *IEEE Access, 5,* 24593-24601. doi: 10.1109/access.2017.2767104