Bachelor Thesis

CHARACTERIZING THE POPULARITY, VIRALITY AND SENTIMENTALISM OF VIDEO CONTENT CATEGORIES ON YOUTUBE AND TWITTER

Georgiou Zacharias

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

May 2017

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

Characterizing the Popularity, Virality and Sentimentalism of Video Content Categories on YouTube and Twitter

Georgiou Zacharias

Supervisor

George Pallis

Diploma project has been submitted for partial fulfillment of the requirements of the degree of Computer Science at the Department of Computer Science of the University of Cyprus

May 2017

Acknowledgement

Firstly, I would like to express my sincere gratitude to Assistant Prof. George Pallis and Associate Prof. Chrysis Georgiou for the continuous support of my BSc study and related research, for his patience, motivation, and immense knowledge. Their guidance helped me during the time of research and writing of this thesis. Last but not least, I would like to thank my family and friends for supporting me spiritually throughout writing this thesis.

Abstract

In this thesis we focus on the observable dependencies between the virality of video content on a micro-blogging social network and the popularity of such content on a video distribution service. We collected and analyzed a corpus of Twitter posts containing links to YouTube clips and the corresponding video metadata from YouTube. Our analysis highlights the unique properties of content that is both popular and viral according to its type, which allows such content to attract high number of views on YouTube (Popularity) and achieve fast propagation on Twitter (Virality). Previous work has shown that predicting popularity and virality of a video is possible with high accuracy and a few days of training.

We developed a framework to assist our data collection and analysis. The architecture of this framework follows a service-oriented design and the server-client model. We focused on the entire separation of the various functionalities, implemented as small and independent services, accessible via a Restful API.

Our results show the unique behavior of the popular and viral videos, in each category. Furthermore, we present the characteristics that contribute the most for the prediction of popular and viral videos according to their category. A key finding of this research is the contribution of the sentiment expressed in the comments of a video, in the prediction of popularity and virality, for news and politics videos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motive

Video sharing services have been seen to experience augment traffic growth during the last decade. In 2015, video Internet traffic constituted 70% of the total Internet traffic. According to CISCO [14], in 2020 the total Internet video traffic will rise to 82%, making content delivery networks (CDNs), carry about three-fourths of the total Internet traffic. This enormous amount of data needs to be distributed on a daily basis to the users, but not in uniform way. Some videos are popular, with views shooting over a billion while others barely attract the attention. Internet sharing strongly affects the popularity of a video. Its exposure on the Internet grows exponentially as more and more people discover and share it with others,making the video attract more views and eventually become popular.

The popularity of a video has also brought attention of marketing agencies. Knowing in advance through which video to advertise, grants a higher probability to attract an audience and thus a higher revenue. Several researches have studied the behavior of popular content

in order to find a mechanism that is able to predict the popularity of such content. A recent study by D. Vallet et al [24], has shown that an accurate prediction is possible. Their work analyzed the activity of popular and viral videos both on social media (Twitter) and a video content distribution channel (YouTube). Specifically, they proposed a model, able to accurately predict the potential of a video to become popular with a reasonably low amount of training data.

Interestingly, they have found some indicators about the effect of the different types of video content on popularity, specifically for music videos. However, this has not been thoroughly investigated. The video type is adapted to a topic, attracting a different audience and provoking different emotions. [15]. For example, the reasons to watch or share a music video, might differ from a news video or a film. To this end, we believe the influence of the video type, affects the video popularity and worths further investigation.

1.2 Objective & Contributions

Our research question is: "*How popularity and virality differentiates among the various video types*". Therefore the objective of this study, is to explore the characteristics of each video category and their impact on popularity and virality. In other words, we investigate how popularity and virality differentiates among the various video categories, by revealing the importance of the various features extracted both from Twitter—a social network for microblogging— and YouTube—a video content distribution service— for predicting the virality and popularity of the different types of video. In addition, we perform sentiment analysis on video comments, to capture the effect of sentimentalism expressed by users in respect to videos' popularity and virality.

The key contribution of this research is to provide a better understanding on how popular and viral videos behave regarding to their type. The categories we looked into, are *Music*, *Games*, *People & Blogs*, *Entertainment*, and *News & Politics*. We also verify the results and findings of a previous work [24], by applying their methodology on our data set. Finally we designed a framework, implemented as small and independent services accessible via a RESTful API, for the data collection and analysis of YouTube videos and Twitter posts.

1.3 Methodology

At the beginning, we study the methodology used in the aforementioned work [24] and customize it to our needs for studying the video's popularity and virality in regard to their type. An overview of the methodology used is depicted in Figure 1.1.

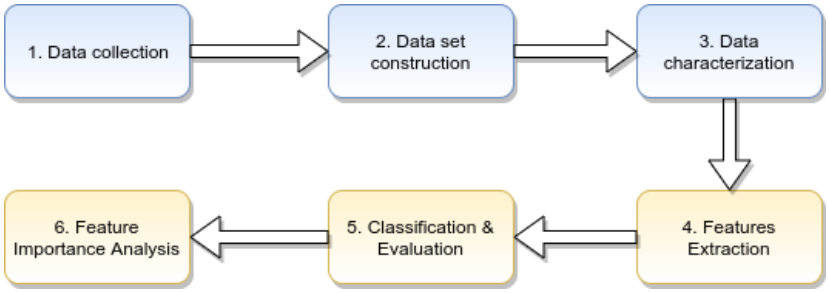


Figure 1.1: An overview of our methodology

In the fist step we collect our data, where we implemented a tool named "`Collector`" to observe Twitter for links to YouTube videos and store all the available and necessary information into a database. This information, includes also comments that refer to the videos. After data collection, we construct our data set used in characterization, in which we group the videos according to the number of views (popular group), the number of tweets (viral group) and their age(recent group), in order to produce plots that provide insights about their behavior.

After data characterization, we proceed with the feature extraction. In this step, we process attributes from videos, tweets and comments to produce a collection of features to train two classifiers; one for the prediction of popularity and the other for virality. Finally, we retrieve the importance of each feature and discuss their contribution in the prediction of popularity and/or virality for each video category.

1.4 Thesis Structure

This document consists of six chapters. In this chapter the purpose of this research is explained. We start with the motivation behind this research and explain its objectives. We then highlight the contributions and give a high level overview of the methodology followed. In Chapter 2, the background and related work is described. Here, we introduce the reader to

the necessary background knowledge for a better understanding of the context of this thesis. We also describe and present the results from the previous work, on which this dissertation is based.

We proceed with the description of the methodology used, in 3, where we give details about our approach and explain the setup of our experiment. In 4, the implementation is described. We give details about the framework we developed to conduct our experiment. In 5, we present and discuss our results. Firstly, we characterize our data set and show the results of the classification process as well as the results of the contribution of each feature for the different categories of the videos. Finally, in 6, we conclude and summarize the findings of our work and discuss the future work.

Chapter 2

Background & Related work

Contents

In this chapter, we explain the related background information that allows the reader to better understand the context of this research. Furthermore, we present the findings of a previous work, on which this research is based.

2.1 Background

2.1.1 User Generated Content

User-generated content (UGC) is defined as the content made publicly available over the Internet [8]. It can be in the form of blogs, wikis, posts, chats, tweets, digital images, video, etc. In this research as we are interested in the activity of videos and tweets, we base our analysis on the content available from YouTube and Twitter.

2.1.1.1 YouTube

YouTube [10] is a free video sharing website allowing the users to upload videos that anyone can watch. Originally created in 2005, YouTube is now one of the most popular sites on the Web, with visitors watching around 6 billion hours of video every month. Not only they can watch a video, but they can express their opinion using likes, dislikes or posting comments. Users are able to create their channel with their own videos and also follow other channels. Each video belongs to exactly one category, chosen by the user. In this study, we define "popularity" of a video, as its inherent propensity to attract views from YouTube. We also use the category of a video to serve as a proxy for its type.

2.1.1.2 Twitter

Twitter [4] is a free social networking microblogging that enables its registered users to share posts called tweets, of up to 140 characters. The users are not only able to broadcast tweets but also to follow other users and interact with tweets through multiple buttons; favourite[1], retweet [2] , and reply. Unless a user chooses private mode, then any user has view access to tweets as Twitter is public. According to Twitter's usage statistics, on 2016 [5] there were 313 million active users per month, justifying the fact that Twitter is an attractive social networking microblogging service.

A viral video is a video that becomes popular through a viral process of Internet sharing, typically through video sharing websites, social media and email [19]. Therefore, in this study we define "virality" of a video, as its potential to elicit Twitter posts from its viewers.

[1] A user can mark a tweet as a favourite to show appreciation or interest for it

[2] When a user wants to re-post someone else's tweet identically

Twitter's social network, can be represented as a directed graph, where nodes refer to the users and edges to their connections. For instance, consider a simple graph of 6 users in Figure 2.1. User A is being followed by 3 users (B, C, D) colored red and follows 2 users (E, F) colored blue. When user A posts a new tweet, his followers will receive that tweet but not his friends, however he receives the tweets posted by his friends.

Therefore, the in-degree of a node indicates the number of followers and the out-degree indicates the number of friends, i.e the number of people the user follows. The more followers a user has, the more popular he is, thus tweets reach more audience.

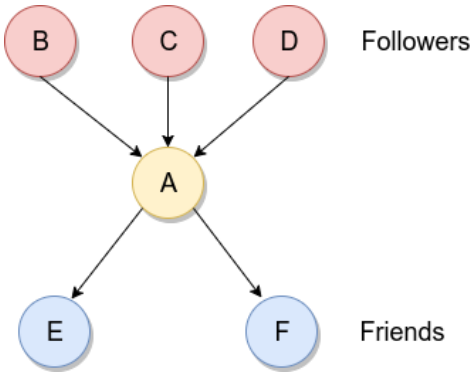


Figure 2.1: Twitter Followers & Friends

2.1.2 Machine Learning

Machine learning according to Arthur Samuel, gives computers the ability to learn without being explicitly programmed [22]. Machine learning is a method of data analysis that automates analytical model building, using algorithms that iteratively learn from data in order to find hidden insights.

Our analysis is based on supervised learning [20], which is the machine learning task of inferring a function from labeled training data. For better understanding, consider an example of a simple training data set, shown in Figure 2.2. The training data consist of a set of training examples, in which the desired output is known. Each data point on the graph represents a video with only two attributes, the number of likes on the x-axis and the number of views on the y-axis. The labels of these examples are "*popular*"(red) and "*not popular*"(blue). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

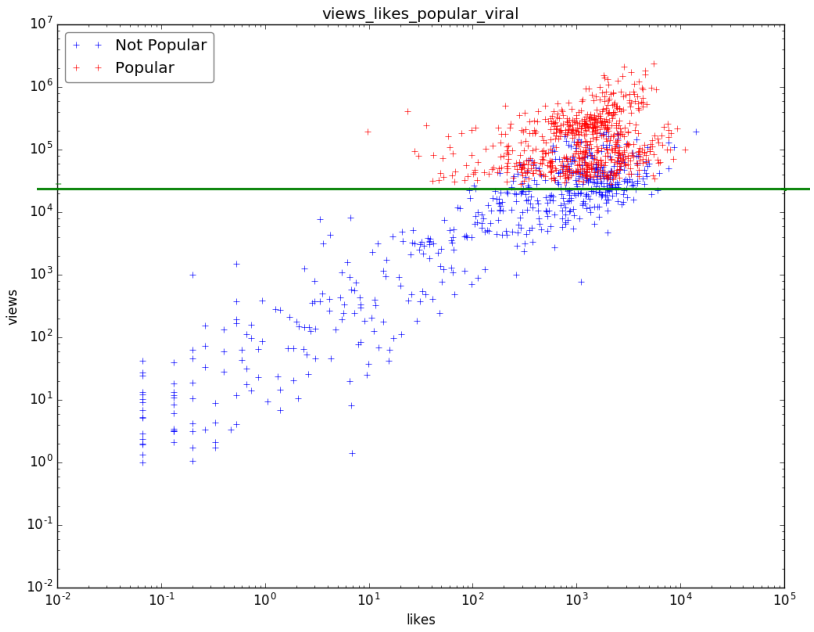


Figure 2.2: Supervised Learning Example

An example function that decides the label of a new data point is shown by the horizontal green line that separates the graph in two regions. The videos above will be predicted as *popular* while the others as not *popular*.

2.1.2.1 Gradient Boosted Decision Tree (GBDT)

A decision tree is a tree-like data structure where leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Following the example above, the simple function that decides if a video is popular or not, is modelled as a decision tree, shown in Figure 2.3.

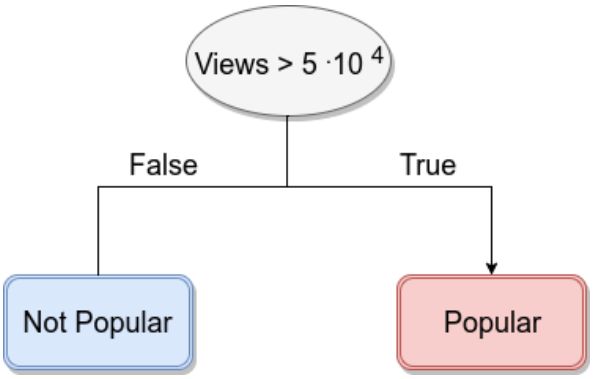


Figure 2.3: Decision Tree Example

Gradient boosting [17] is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. The key advantages of gradient boosting [23] is the robustness to outliers, its predictive power and the natural handling of heterogeneous features. In our study, we use GBDT in order to classify the videos as popular and/or viral, using the features extracted from YouTube and Twitter.

2.1.2.2 Classifier Performance - F1 Score

The are various ways to measure the performance of a binary classifier. In our study we chose the F1 score to measure the performance of our classifiers for the prediction of popular and/or viral videos. F1 score is interpreted as a weighted average of the precision and recall. It ranges from 0 to 1, with 0 the worst value and 1 its best. To better understand, consider the diagram in Figure 2.4. On the top-left corner the positive and negative labels of a data set are shown. A perfect binary classifier would have guessed all the positive labels as positive and all the negative labels as negative. However, errors can happen, leading to 4 possible outcomes(top-right corner), described below.

True Positive (TP)

It is equivalent with a hit. When the class is correctly predicted as positive.

True Negative (TN)

It is equivalent with a correct rejection. When the class is correctly predicted as negative.

False Positive (FP)

It is equivalent with false alarm(Type I error). When the class is wrongly predicted as positive.

False Negative (FN)

It is equivalent with a miss(Type II error). When the class is wrongly predicted as negative.

Recall is calculated using Equation 2.1 and gives information about a classifier's performance with respect to false negatives(how many did we miss).

$$Recall = \frac{tp}{tp + fn} \tag{2.1}$$

Precision is calculated using Equation 2.2 and gives information about a classifier's performance with respect to false positives(how many are actually positive)

$$Precision = \frac{tp}{tp + fp} \tag{2.2}$$

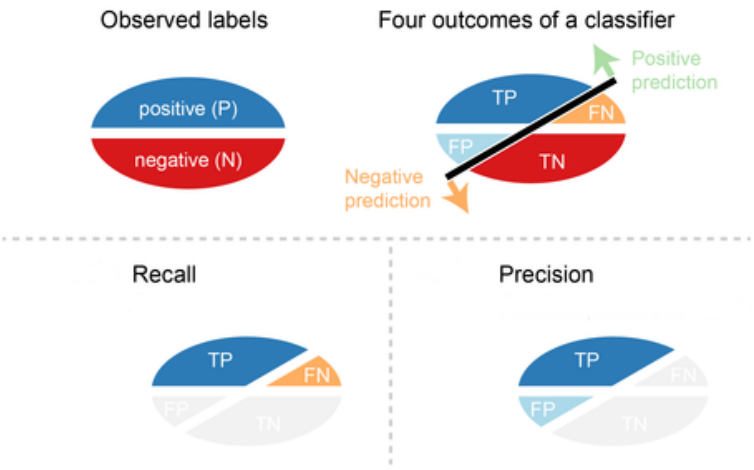


Figure 2.4: Precision-Recall Example

F1 score is calculated using Equation 2.3.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.3}$$

2.1.2.3 Feature Importance Analysis

Generally, feature importance provides a score that indicates how useful or valuable each feature was in the construction of a model. In our case, the model is an ensemble of decision trees. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. In our study, feature importance analysis has been used to reveal the contribution of each individual feature in the prediction of popularity and virality.

If we consider the example we described above, in Figure 2.3, the $views$ feature will have 100% contribution in the prediction of popularity label. That is happening, because it is the only feature used by the classifier to determine the video's label.

2.1.3 Sentiment Analysis

Sentiment analysis [21], also known as opinion mining or emotion AI, refers to the use of natural language processing, text analysis, computational linguistics, and bio-metrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject in respect to a topic or the overall contextual polarity or emotional reaction to a document, interaction, or event.

In our study, sentiment analysis has been used to capture how much negative, positive or neutral feelings are expressed in the comments of a video. We use VADER [18], a rule-based sentiment analysis tool, specifically attuned to sentiments expressed in social media. Specifically, it can detect emoticons[3], initialisms and acronyms, punctuations, slang[4] words and etc.

2.2 Related work

As we already mentioned, D. Vallet et al [24] studied the unique properties of content that is both popular and viral, i.e attracts high number of views on YouTube and achieves fast propagation on Twitter. They proposed a unifying approach for predicting video content virality and popularity, using features extracted from Twitter and YouTube logs, achieving high

[3]Is a pictorial representation of a facial expression using punctuation marks, numbers and letters.

[4]Consists of a lexicon of non-standard words and phrases implying particular attitudes.

levels of accuracy with reasonably low amounts of training data. In our study we followed a similar methodology, described extensively in Chapter 3.

Their analysis consists of a classification task where they aim to classify each video, based on a set of input features, as viral and/or popular. They trained two independent classifiers, one for popularity and one for virality using Gradient Boosting Decision Tree algorithm. As a comparison baseline, they chose a simple classifier that only uses two features: the number of original tweets and the number of views, reflecting a simple classifier that uses only the raw statistics on the uptake of the videos in both systems. The analysis splits the videos into two groups, according to their age; those uploaded less than 14 days prior to data collection (referred to as recent) and the rest (others).

They performed a cross-system feature importance analysis to find the contributions of each feature. That is, they computed the importance values when predicting video popularity on YouTube using Twitter features and also when predicting video virality on Twitter using YouTube features. They showed that for recently uploaded videos the rate of tweets and the number of users reached, are the most important for predicting popularity on YouTube. For predicting virality on Twitter, the most dominant predictive feature is the difference of ratings, capturing the number of likes per day, the number of views and the number of videos uploaded by the uploader. For predicting popularity and virality, YouTube features dominate Twitter features with the number of views added on last training day substantially outweighing other features.

In this research, we apply the same methodology with this work for each individual category of videos. We collect more attributes from YouTube and Twitter, including the sentiment, thus we extract more features for the classification.

Chapter 3

Methodology

Contents

This chapter explains in detail the methodology used in this research. We discuss extensively, the steps followed in order to study popularity and virality for each type of video.

3.1 Data Collection

The first step of this research involves the construction of our data set. To achieve this, we collect the information provided by YouTube and Twitter API. Specifically, we monitor the Twitter stream over a period of two weeks, for tweets that contain links to YouTube videos and extract all the available information. For this purpose we implemented a tool named *Collector*, described with detail in Section 4.1. We collected 135,000 videos with 8 million tweets and around 8 million comments.

We proceed with the separation of the videos according to their type. The type of a video is identified by the category it belongs to. Currently, there are 32 categories [12] as shown in Figure 3.1. From the tweets we obtained, we can see that two thirds of all the videos correspond only to 5 categories, which are described in the Table 3.1. Each of the remaining

categories contained a small number of videos, therefore we grouped them together and labeled them as others. Some of the rest categories,such as comedy, documentary and drama that refer to movie genres, were merged with the videos belonging to the entertainment category.

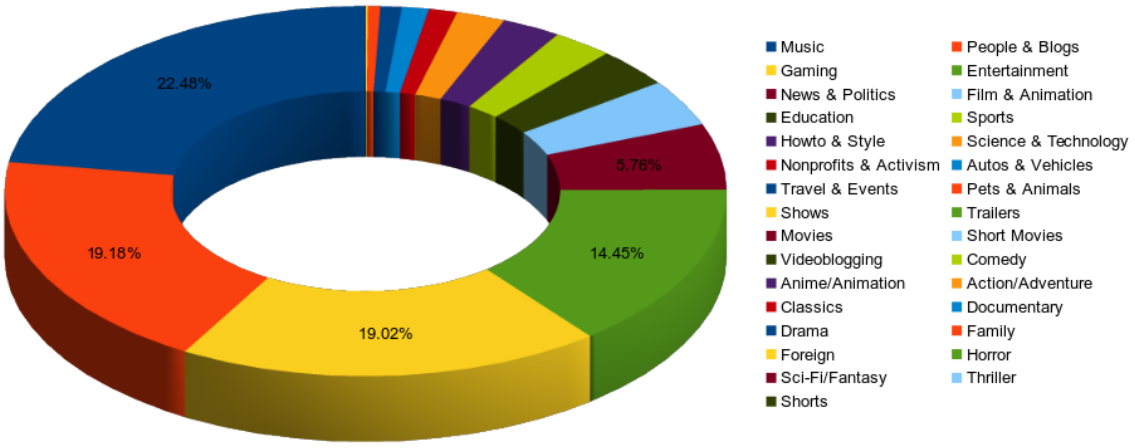


Figure 3.1: Categories for the 135,000 videos collected

Music	Videos featuring songs from a variety of genres
Games	Videos featuring news, reviews, playthroughs, and more.
People & Blogs	Videos with people talking, sharing ideas and opinions for any topic
Entertainment	Videos that contain any form of activity that holds the attention and interest of an audience, or gives pleasure and delight.
News & Politics	Videos featuring comprehensive up-to-date coverage on the latest top stories, sports, business, entertainment, politics, and more.

Table 3.1: Video Categories

Part of the data collection, is the sentiment analysis on YouTube comments for each video.

The average negative, positive and neutral sentiment is extracted for each video and it is used as a predictive feature for our classifiers.

3.2 High Level Characterization

After data collection, for each of the above categories, we define 7 groups of videos, described in Table 3.2. The last three groups derive from the intersections and differe of the videos in the first 4 groups. The videos in popular and viral groups, are based on YouTube views and Twitter mentions captured during the observation period. These are also the videos that are labeled as popular and viral, used in the classification process. The percentage used of 2.5% is chosen according to the related work [24].

Popular	2.5% of all the videos with the highest view count
Viral	2.5% of all the videos with the highest number of tweet mentions
Random	2.5% of all the videos, randomly chosen
Recent	2.5% of all the videos that were at least 2 days old from the date collected
Popular & Viral	Videos that belong both in popular and viral group
Popular & Not Viral	Videos that belong in popular, but not in viral group
Viral & Not Popular	Videos that belong in viral, but not in popular group

Table 3.2: Video Groups

We proceed with a high level characterization on these 7 groups in order to provide insights about the behavior of each group of videos, specifically the popular and viral. The characterization is performed for each of the 6 categories of videos and shows their differentiation. We analyze their behavior during a two week period, by producing various graphs such as video views and tweet mentions per day.

3.3 Classification & Feature Analysis

After the characterization we proceed with the part of Classification and Feature Analysis. For each video we extract its features depending on the size of the training window. The features are described in detail in Section 4.2.2.2. In Figure 3.2, the timeline of the classification is shown. We separate the period into the training and labeling window. The training window is the period that the classifiers are trained using all the available data during these days and the labeling window is the period in which only the label(popular and/or viral) of the videos is known or predicted.

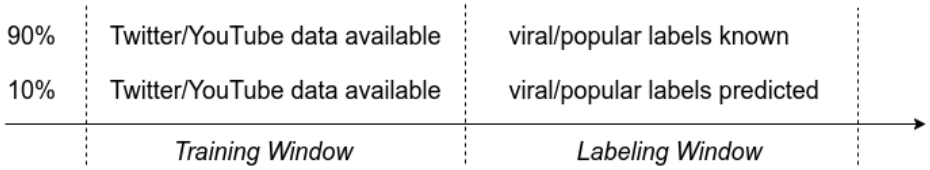


Figure 3.2: Classification Timeline

For classification, we train two binary classifiers capable of separately predicting the viral and popular label for any given video. We implemented the classifiers using a Gradient Boosted Decision Tree, widely used for general classification problems [17].

For the evaluation of the predictive accuracy of our classifiers we used 10-fold cross validation method. We randomly separate the data into 10 equal sized chunks, where 9 are used for the training and 1 for the validation. This process is repeated for 10 times and on each iteration the validation chunk changes, until all of them are used for validation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once, which is perfect for small data sets.

For the experiment we set the training window size to 2 days and the labeling window size to 7 days, using all the features extracted from YouTube and Twitter. For 2 days training, we used 39 YouTube features and 63 Twitter features. As a comparison baseline, we chose a simple classifier that only uses two features: the number of tweets and the number of views. This reflects a simple baseline classifier that uses only raw statistics from the videos. We also split the videos into two more groups according to their age. Those uploaded two weeks prior to data collection, referred as *recent*, and the rest, referred as *others*.

After the classification, we obtain the precision-recall graphs of our classifier for each category of videos, and calculate the F1 score. Finally, we perform Feature Importance Analysis to find out which of the features contributed the most in the prediction of popularity and virality. We analyze the contribution of each individual feature among the different categories of videos.

Chapter 4

Implementation

Contents

In this chapter we give details about the implementation of the methodology used. We explain the core tools that enable us to collect and analyze the data from YouTube and Twitter. In addition, we explain what data is collected and what features are extracted. The source code of all the developed tools, is hosted on Github [25].

A big part of this research is devoted into the implementation of a framework that enables the study of videos' popularity and virality. The framework consists of two core services,

the `Collector` and the `Analyzer`. The architecture of this framework follows a service-oriented design and the server-client model. We focused on the clear separation of the various functionalities, implemented as small and independent services, accessible via a Restful API. For the persistence storage we used MongoDB, a NoSQL document oriented and scalable database able to handle large amounts of data.

4.1 Collector

To construct our data set we implemented `Collector`. This application provides the means for collecting all the necessary information to conduct our experiment. `Collector` consists of 5 core services as shown in Figure 4.1, each of them running on its own thread. The interaction with this tool is achieved through the module that handles HTTP requests, providing several endpoints for monitoring and configuration.

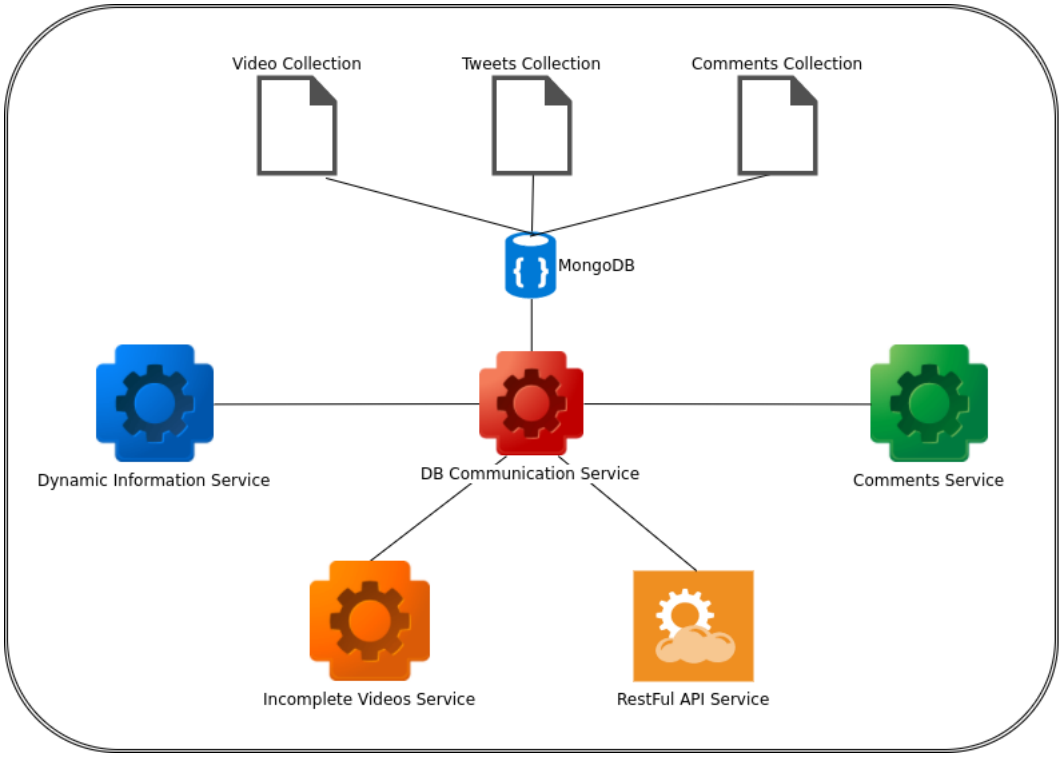


Figure 4.1: Collector Core Modules

Listing 4.1: Regular expression for YouTube links

```
(https?:\/\/youtu\.be)|(https?:\/\/www\.youtube\.com\/watch\?v=)
```

4.1.1 Streaming Service

Streaming service makes use of the Twitter's public stream endpoint [6]. In order to use this stream a user needs to create a Twitter Application to generate the 4 keys needed for the authentication. When the user connects to this endpoint, he obtains a sample of all the public tweets on real time. However, we are interested only for the tweets that contain a link to a YouTube video. This is achieved by filtering the tweets using the regular expression shown in Listing 4.1. This regular expression accepts only the links to `youtube.com` and `youtu.be` hosts. It is worth mentioning, that the links contained in the text, are shortened[1] in order to save characters, because as it was already said, Twitter has a 140 characters limit. Fortunately, *Twitter4j* library provides an easy way to retrieve the expanded url contained in a tweet.

After the collection of a valid YouTube link, the information extracted from the tweet is added into the database. There are two cases here. The first case, is that the mentioning video is already in the database for monitoring, thus a new record is inserted. If the video doesn't exist - the second case - we add the new video into the database. Video information is either static or dynamic. The video static information is collected only once and the fields are described in the Table 4.2. A description of all the information extracted from a tweet is given in the Table 4.1

[1] It is a technique in which a url is converted into a smaller one and direct to the original page

Field	Description
tweet_id	The id of the tweet
created_at	The time that was created
text_at	The text of the tweet
favorite_count	Indicates approximately how many times the tweet has been liked by Twitter users
lang	The language used in the text
is_possibly_sensitive	It is an indicator that the URL contained in the Tweet may contain content or media identified as sensitive content
is_retweeted	Whether this Tweet has been retweeted by the authenticating user
user_created_at	Time that the user account has been created
user_followers	The number of users that follow this account
user_friends	The number of users this account follows
user_favorites	The number of tweets this user has favorited in the account's lifetime
user_listed	The number of public lists that this user is a member of
user_statuses	The number of tweets (including retweets) issued by the user
user_verified	When true, indicates that the user has a verified account
user_lang	The BCP 47 * code for the user's self-declared user interface language

* BCP 47 RFC: https://tools.ietf.org/html/bcp47

Table 4.1: The information collected for a Tweet

When a new video is added to the database, we request its static fields of information from the YouTube API [11]. Meanwhile, for each video we store meta data, used to determine its state. The state diagram for the video class is illustrated in Figure 4.2.

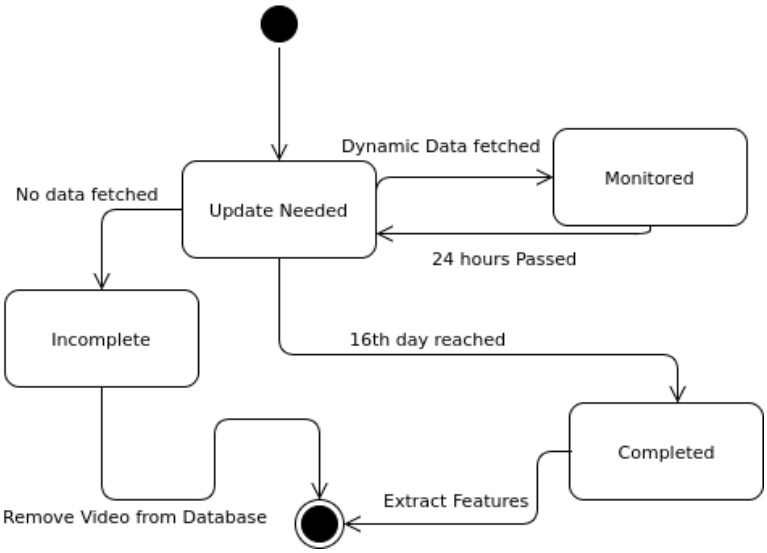


Figure 4.2: State diagram for the video class

Field	Description
video_id	The id of the video
title	The title of the video
channel_id	The id of the video's channel
description	A small description for the video provided by the uploader
category_id	The id of the category the video belongs to
published_at	The time that the video was uploaded
duration	The number of seconds of the video

Table 4.2: The static information collected for a video

4.1.2 Dynamic Information Service

This service provides all the functionality needed to collect the statistics of a video once per day from YouTube. Running on its own thread, it checks the database every 10 minutes for videos that need update. A video needs an update when 24 hours passed since its last update. On every update we insert to the database the dynamic information described in Table 4.3. It is critical for our purposes to retrieve the dynamic information every 24 hours. However, there are cases that the process of dynamic information retrieval will fail because the video is

removed by the user or due to network errors. In any case, this service will mark that video as incomplete and another service will remove it from the database along with its tweets and comments. Furthermore, this service is also responsible to identify if a video has reached its 16^{th} day, to set its state as completed.

Field	**Description**
timestamp	The time that dynamic data was collected
view_count	Current number of views
like_count	Current number of likes
dislike_count	Current number of dislikes
favorite_count	Current number of times a the video was listed as favored
comment_count	Current number of comments
channel_view_count	Current number of views in the video's channel
channel_comment_count	Current number of comments in the video's channel
channel_subscriber_count	Current number of subscribers in the video's channel
channel_video_count	Current number of videos in the channel that the video belongs to

Table 4.3: YouTube Dynamic Information collected

4.1.3 Comments Service

This service fetches the most 100 relevant comments for each video and it is executed once the video is added to the database. These are the comments that are shown more frequently to the user as it is the default option by YouTube. These comments are calculated primarily by Google+ quality factors, such as the user's YouTube channel age and the number of comments the user posted on his Google+ profile, especially the ones containing link to a YouTube video. Also, these are the comments that contain quality discussions and are published by popular personalities [13].

4.1.4 REST API

The user's interaction with this tool is performed using several endpoints, provided by this module. For security reasons all the requests need a token, set by the administrator. A description of the endpoints is shown in Table 4.4. Users can insert the parameter *help=1* on the request to retrieve a more detailed explanation for that endpoint.

Endpoint	Description
`GET /`	Retrieve the configurations and various statistics about the application, such as the number of videos being monitored or finished.
`PUT /`	Set the maximum number of videos being monitored. Requires *max_videos* parameter.
`PUT /youtubeApp`	Insert a YouTube API key into the database, which is necessary for requests to YouTube API. Requires *api_key* parameter.
`PUT /twitterApp`	Insert a Twitter application's keys into the database, which is necessary for requests to Twitter API. Requires the name for the Twitter application and it's 4 keys.

Table 4.4: Collector Endpoints

4.1.5 Technologies Used

In this application we make use of first-class technologies that provide useful functionalities for this program. For the communication with Twitter we used Twitter4j library and for the web server running our Restful API we used Spark micro-web framework. For the persistent storage we used MongoDB, to parse json documents we used Gson library, and finally for the sentiment analysis we used VADER. A brief description of these technologies is provided below.

Twitter4j

Twitter4j [7] is a Java library for the Twitter API. It provides an easy way to authenticate with Twitter and use its services to connect with the streaming API.

MongoDB

MongoDB [2], is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas. It supports map-reduce operations, widely used for processing large volumes of data.

Gson

Originally created by Google, Gson is a Java library that can be used to convert Java Objects into their JSON representation. Currently Gson is released under Apache Licence 2.0

Spark - Web Framework

Spark [3], is a micro framework for creating web applications in Java 8. Spark is mainly used for creating REST API's, but it also supports a multitude of template engines. Under the hood, Spark, runs an embedded Jetty server.

VADER

VADER [9] stands for Valence Aware Dictionary and sEntiment Reasoner. It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.

4.2 Analyzer

Analyzer is a tool that performs a multitude number of operations over the data collected by the Collector tool. It is accessible via its restful API. Firstly, it processes videos,tweets and comments to construct our data set. It also groups together the videos, produces graphs and statistics, and extract the features used in the classification and feature importance analysis. For processing, this tool utilizes the aggregation framework provided by MongoDB. We use a processing pipeline, to filter and tranform our data, according to our needs.

4.2.1 Core Operations

The core operations of this tool are illustrated in Figure 4.3 and described in the following paragraphs.

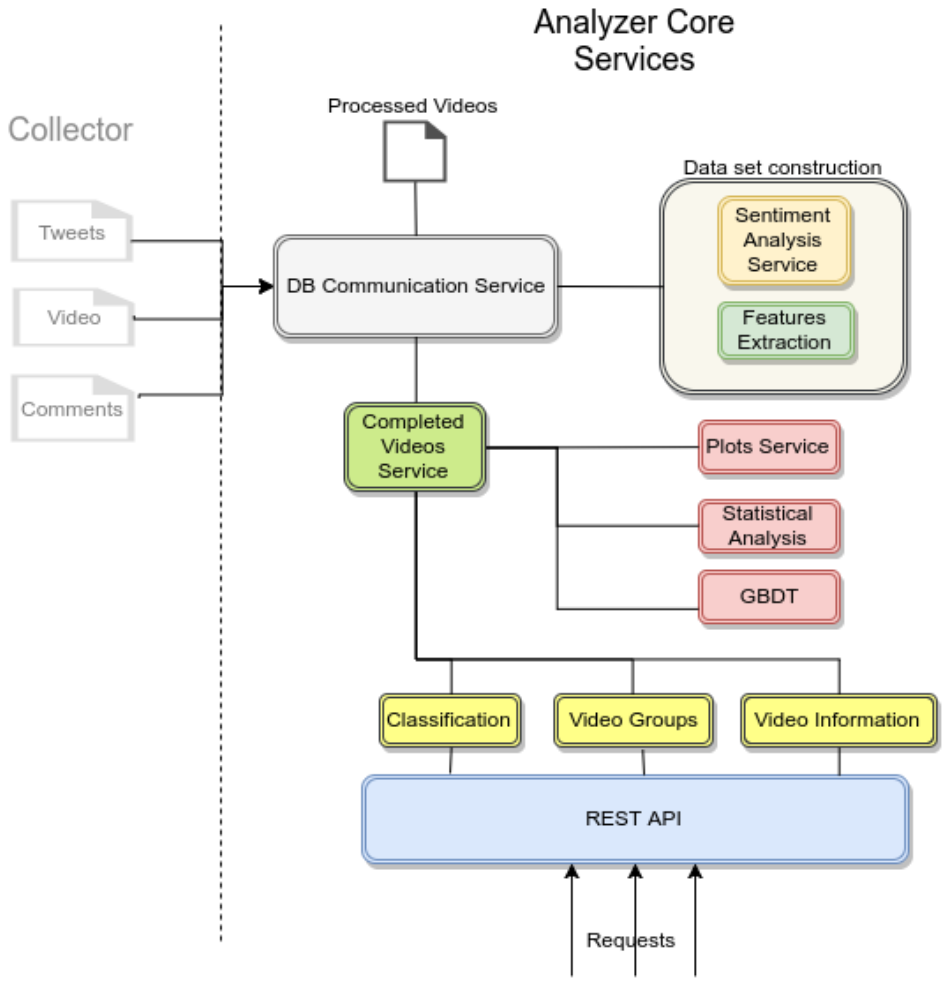


Figure 4.3: Analyzer Core Operations

Completed Videos Monitoring

This service constantly checks the database for completed videos. These are the videos that have reached their 16^{th} day. Once this service finds completed videos, it sends the batch of videos to another service for processing, that uses the aggregation framework provided by MongoDB in order to create our models—they are described in Subsection 4.2.2.

DB Communication Service

This service, acts as a middleware between the database and the main program. It

provides an abstraction layer for the data retrieval, offering various queries, needed for the analysis. These queries are builded as aggregation functions. MongoDB, provides map-reduce operations to perform aggregation. In general, map-reduce operations have two phases: a map stage that processes each document and emits one or more objects for each input document and a reduce phase that combines the output of the map operation.

Sentiment Analysis Service

This service, extracts the sentiment of the comments of each video. It uses the VADER tool, described earlier, to compute the average negative,neutral and positive sentiment found in the comments of each video.

Classification & Feature Importance Analysis Service

This service trains and evaluates two binary classifiers using Gradient Boosting Decision tree and produces the Precision-Recall graphs, F1 scores and the contribution of each feature. The classifier is implemented in python, using *sci-kit* package.

Plots Service

This service creates the desired plots. It uses *matplotlib* [1], a package for python, to create customizable 2D plots for the purposes of our research.

Statistical Analysis Service

This service offers various functions to perform statistic operations, such as computing the average, median and standard deviation in a list of numbers. It is used for the construction of our models.

4.2.2 Models & Features

4.2.2.1 Models

Processing of the videos, comments and tweets results to the video class model, depicted in Figure 4.4. These are the attributes extracted for the individual videos and are used during the analysis. Note that, the video class contains fifteen daily stats, one for each day of observation.

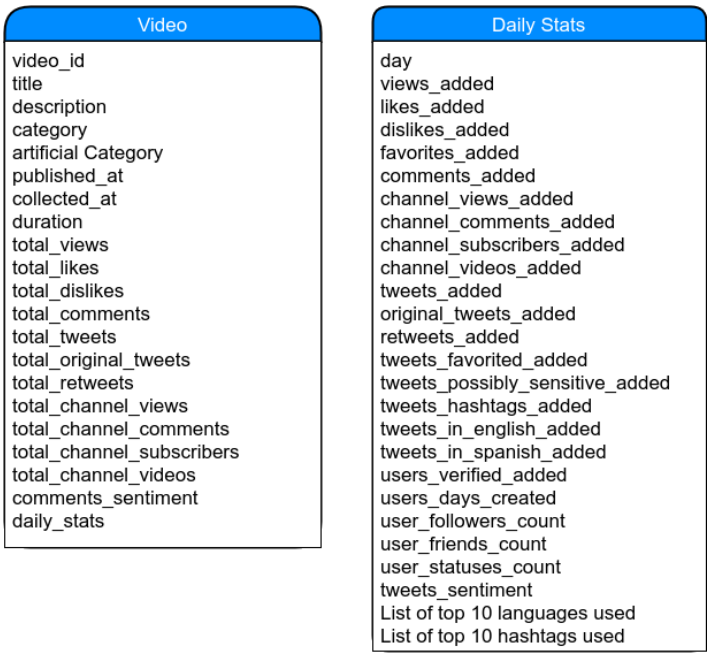


Figure 4.4: Video Modeling

In Figure 4.5, the Groups modeling is shown. Videos were grouped together for the purposes of our analysis. For each group we provide statistics that contain the average, median and standard deviation for each characteristic.

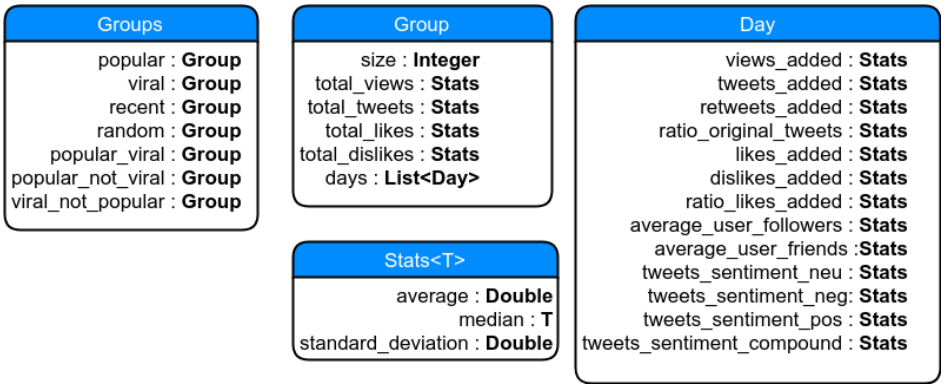


Figure 4.5: Groups Modeling

4.2.2.2 Features

In order to classify a given video as popular or viral, their features need to be extracted. The features have been extracted using a tool developed in another thesis [16], in which they compare different algorithms for the prediction of popularity and virality of a video. The

computed YouTube features used by our classifiers are described in the following tables ; 15 static and 12 that are calculated for each training day.

Feature	**Description**
duration	Video duration in milliseconds
negative_sentiment	The average negative sentiment of the video comments.
neutral_sentiment	The average neutral sentiment of the video comments.
positive_sentiment	The average positive sentiment of the video comments.
channel_uploads	The number of videos in the channel that the video belongs to.
channel_subscribers	The number of subscribers in the channel that the video belongs to.
channel_views	The total number of views in the channel that the video belongs to.

Table 4.5: YouTube Static Features

Feature	**Description**
views_diff	The difference between the accumulated views of a video at the first and last day of the training.
likes_diff	The difference between the accumulated likes of a video at the first and last day of the training.
dislikes_diff	The difference between the accumulated dislikes of a video at the first and last day of the training.
comments_diff	The difference between the accumulated likes of a video at the first and last day of the training.

Table 4.6: YouTube Difference Features

Feature	Description
views_acc	The average acceleration* of the views.
likes_acc	The average acceleration of the likes.
dislikes_acc	The average acceleration of the dislikes.
comments_acc	The average acceleration of the comments.

* Acceleration: The ratio of a feature between day n and day $n-1$

Table 4.7: YouTube Acceleration Features

Feature	Description
views_ratio$_i$	The ratio between the number of views on i-th day and the total views.
likes_ratio$_i$	The ratio between the number of likes on i-th day and the total likes.
dislikes_ratio$_i$	The ratio between the number of dislikes on i-th day and the total dislikes.
comments_ratio$_i$	The ratio between the number of comments on i-th day and the total comments.

Table 4.8: YouTube Ratio Features

Feature	Description
views_age_ratio$_i$	The ratio between the number of views on i-th day and the date it was uploaded.
likes_age_ratio$_i$	The ratio between the number of likes on i-th day andthe date it was uploaded.
dislikes_age_ratio$_i$	The ratio between the number of dislikes on i-th day and the date it was uploaded.
comments_age_ratio$_i$	The ratio between the number of comments on i-th day and the date it was uploaded.

Table 4.9: YouTube Ratio Features

Feature	**Description**
views_i	Views added on i-th day
likes_i	Likes added on i-th day
dislikes_i	Dislikes added on i-th day
comments_i	Comments added on i-th day

Table 4.10: YouTube Daily Features

In the following tables the Twitter features are described. The computed Twitter features used by the classifiers are described in the following tables ; 23 static and 20 that are calculated for each training day.

Feature	**Description**
users_followers	The average number of followers referring to a video
users_verified	The number of verified users referring to a video
users_friends	The average number of friends referring to a video

Table 4.11: Twitter Static Features

Feature	Description
tweets_diff	The difference between the total amount of tweets referring to a video at the first and last day of the training
orig_tweets_diff	The difference between the total amount of original tweets referring to a video at the first and last day of the training
retweets_diff	The difference between the total amount of retweets referring to a video at the first and last day of the training
tw_user_favorites_diff	The difference between the total amount of favorites referring to a video at the first and last day of the training
tw_eng__diff	The difference between the total amount of English tweets referring to a video at the first and last day of the training
tw_sp__diff	The difference between the total amount of Spanish tweets referring to a video at the first and last day of the training
tw_users_eng__diff	The difference between the total amount of English users referring to a video at the first and last day of the training.
tw_users_sp__diff	The difference between the total amount of Spanish users referring to a video at the first and last day of the training
tw_users_statuses_diff	The difference between the total number of statuses posted by users referring to a video at the first and last day of the training
tw_hashtags__diff	The difference between the total number of hashtags referring to a video at the first and last day of the training

Table 4.12: Twitter Difference Features

Feature	**Description**
tweets_acc	The average acceleration* of tweets
orig_tweets_acc	The average acceleration of the original tweets
retweets_acc	The average acceleration of retweets
tw_user_favorites_acc	The average acceleration of user favorites
tw_eng__acc	The average acceleration of English tweets
tw_sp__acc	The average acceleration of Spanish tweets
tw_users_eng__acc	The average acceleration of English users referring to a specific video
tw_users_sp__acc	The average acceleration of Spanish users referring to a specific video
tw_user_statuses_acc	The average acceleration of statuses of a user, referred to a specific video
tw_hashtags__acc	The average acceleration of hashtags referring to a specific video

* Acceleration: The ratio of a feature between day n and day $n-1$

Table 4.13: Twitter Acceleration Features

Feature	**Description**
tweets_added	The number of tweets added on i-th day
original_tweets_added	The number of original tweets added on i-th day
retweets_added	The number of retweets added on i-th day
user_favorites_added	The number of user favorites added on i-th day
english_tweets_added	The number of English tweets added on i-th day
spanish_tweets_added	The number of Spanish tweets added on i-th day
users_english_added	The ratio between the number of English users posted on i-th day
users_spanish_added	The ratio between the number of Spanish users posted on i-th day
user_statuses_added	The number of user statuses added on i-th day
hashtags_added	The number of hashtags used on i-th day

Table 4.14: Twitter Daily stats

Feature	Description
ratio_tweets	The ratio between the number of tweets on i-th day and the total number of tweets.
ratio_original_tweets	The ratio between the number of original tweets on i-th day and the total number of original tweets.
ratio_retweets	The ratio between the number of retweets on i-th day and the total number of retweets.
ratio_user_favorites	The ratio between the number of user favorites on i-th day and the total number of user favorites.
ratio_english	The ratio between the number of English tweets on i-th day and the total number of English tweets.
ratio_spanish	The ratio between the number of Spanish tweets on i-th day and the total number of Spanish tweets.
ratio_users_english	The ratio between the number of English users on i-th day and the total number of English users.
ratio_users_spanish	The ratio between the number of Spanish users on i-th day and the total number of Spanish users.
ratio_user_statuses	The ratio between the number of user statuses on i-th day and the total number of user statuses.
ratio_hashtags	The ratio between the number of hashtags used on i-th day and the total number of hashtags used.

Table 4.15: Twitter Ratio Features

4.2.3 REST API

The core functionalities of this application are served by the endpoints described in the Table 4.16. This API, executes the operations described earlier and retrieves the results in a json format. It also offers the ability to retrieve graphs concerning the classification or statistics about the groups, in png format.

Endpoint	Description
`GET /`	Retrieve various statistics about the application, such as the number of videos being monitored or finished.
`GET /videos`	Retrieve information about the number of videos per category.
`GET /videos/:id`	Retrieve information about the video with the value id.
`GET /videos/categories`	Retrieve information about the videos in each category.
`GET /videos/groups`	Retrieve information about the group of the videos.
`GET /videos/popular`	Retrieve the popular videos.
`GET /videos/viral`	Retrieve the viral videos.
`GET /videos/recent`	Retrieve the recent videos.
`GET /videos/random`	Retrieve random videos.
`GET /videos/classify`	Perform and retrieve classification results.
`GET /plots/:id`	Retrieve the requested plot.

Table 4.16: Analyzer Endpoints

Chapter 5

Results

Contents

5.1 High Level Characterization

In this chapter the results concerning the characterization of the video categories and groups are presented and discussed. The purpose of this characterization is to provide insights about

the behavior of the different groups of videos regarding their category. In addition, we show the results from the feature importance analysis, answering the primary question of this research; which characteristics from each video category, impact video popularity and virality.

5.1.1 Video Age Distribution

Video age distribution is illustrated in Figure 5.1. The graphs show a cumulative fraction of the videos collected, in comparison to the number of days passed since their release day. Observing the individual graphs, the popular and viral group of videos, in all the categories except *Music* and *People & Blogs*, hold a higher proportion of recent videos, compared to the random videos or the ones that are only popular or only viral.

By looking at all the graphs, we can see that the groups in the graphs for *Music* and *Games* categories show a completely different behavior. While the majority of the music videos are old, game videos are recent. If we take for example the videos that are 1 week old, the percentage of music videos is around 20% while for game videos is 80%. `News & Politics` , show a similar behaviour with the *Entertainment* category, with a slightly larger proportion of recent videos.

We can conclude that the users, post tweets mostly referring to recent videos for all the different types, except for music. This shows, that music videos despite their age, they still attract a great amount of tweets, compared to the other types of videos.

5.1.2 Views Daily Increase

The average daily increase in view count for each video category and group is depicted in Figure 5.2. The x-axis represents the days during the monitoring period and the y-axis shows the number of views added, in logarithmic scale. By looking at all the categories, we can see that the videos that are both popular and viral, attract the highest number of views for all days. It is also visible that videos that are popular but not viral, achieve fewer views than the ones that are also viral.

The most important result shown by these graphs, is that the music videos show a steady increase in view count for each day, while the videos in the other categories show a significant decrease in their views. This decrease is steeper in the first days, but it gradually stabilizes until the last day of observation. We believe that this behavior is not shown by the music videos, because of their age. These videos might had a similar behaviour on their first days, but during this period of observation, they managed to stabilize.

5.1.3 Tweets Daily Increase

Figure 5.3, shows the average daily increase on the number of tweet mentions for each video class and category. The y-axis represents the average number of tweets that correspond to each video class and the x-axis represents the days of observations.

The highest daily increase in tweet mentions for all the categories of videos is taken by viral videos. During the observed period, only for music videos, the popular and viral group attracts the highest number of tweets for all days. For *News & Politics*, *Games* and *Entertainment* the popular and viral group still attracts the highest number of tweets, but only for the first 5-7 days. After these days have passed, the viral but not popular group takes its place, attracting the highest number of tweets. This transition can be explained by the fact that some videos start as popular but through time they lose their popularity, thus they remain only viral. Indeed, videos about news and politics have their peak when an event takes place, but after it ends the popularity is gradually decreased. This didn't happen for music videos, because they don't experience this effect in such a degree.

In general, all the videos for all groups and categories experience a decrease in the number of tweet mentions as time passes. The decrease is steeper during the first days, but it slowly stabilizes in the last days. Specifically the *Games*, *Entertainment* and *News & Politics* categories, experience a dramatic drop in tweet mentions during the first days.

5.1.4 Original Vs Total Tweets Ratio

The graph in Figure 5.4, shows the average ratio between the original and total tweets for the different categories of videos over a two weeks period. Tweets are either original or retweets, thus a higher ratio means less retweets.

We see that popular but not viral videos attract the most original tweets for all the categories and videos which are viral but not popular attract more retweets than the other groups for all video categories. Videos marked as both popular and viral remain between those two groups, during the observed period. We can observe that popular videos attract more original tweets compared to the viral videos which attract more retweets. These graphs give an indication that viral videos highly depend on high retweeting rate. Nevertheless, a video must inspire users to post original tweets in order to be popular as well as viral.

5.1.5 Negative Sentiment

In Figure 5.5, the negative sentiment of the various classes is depicted. For all categories, the popular and viral videos have the highest negative sentiment. The comments from the videos that are classified as only popular or only viral, express a lower amount of negative sentiment. Interestingly, we observe that videos belonging to the *News & Politics* category, achieve the highest negative sentiment. Popular and viral videos from this category, have twice the negative sentiment of the other classes.

Indeed, in the Feature Importance Analysis, we observe that the negative sentiment contributes the most to the classification of videos belonging to News & Politics category, as popular and viral.

5.2 Prediction Accuracy

In this section, the accuracy of the classifiers for predicting popularity and/or virality of a video, is presented. We show the F1 score of each video cateory for the recent group of videos, i.e the videos that were younger than two weeks since the collection day.

5.2.1 YouTube Features

In Table 5.1, we show the F1 score of each category for predicting *popularity*, *virality* and *popularity and virality*, using all the available features from YouTube. In general, the prediction of a video popularity achieves a higher F1 score, than predicting *virality* and *popularity*

and virality. However, the most interesting is the cross dynamic predictions where the virality of a video using only YouTube features is predicted and achieves a reasonable F1 score, specifically for the music videos.

As for the categories, music videos achieve the highest F1 score with a comparable difference from all the other categories, while videos about news and politics achieve the lowest score. We believe that this is happening due to the number of videos used to train the classifiers. Also, another reason could be the fact that music videos showed a more stable behavior during the observed period compared to the other categories.

	Popularity	**Virality**	**Popularity & Virality**
Music	0.94	0.75	0.84
Games	0.87	0.64	0.67
People & Blogs	0.90	0.61	0.62
Entertainment	0.89	0.62	0.69
News & Politics	0.85	0.58	0.57
Others	0.89	0.65	0.58

Table 5.1: F1 Score using only YouTube features

5.2.2 Twitter Features

In Table 5.2, we show the F1 score of each category for predicting *popularity*, *virality* and *popularity and virality*, using only Twitter features. In general, predicting both popularity and virality achieves a high F1 score for all the categories, with *Music* category achieving the highest and *Others* the lowest score. For predicting only the popular videos, the score is around 0.64 for all categories except for *Music* which is 0.76. This changes when predicting viral videos. The performance of the classifier gets better because virality is associated with the Twitter features. The score increases above 0.82 for all categories, except for the *News & Politics* videos, which increases up to 0.75.

We can clonclude that Twitter features can give good predictions for both popular and viral videos, especially for *Music* videos.

	Popularity	**Virality**	**Popularity & Virality**
Music	0.76	0.88	0.80
Games	0.63	0.83	0.67
People & Blogs	0.65	0.82	0.64
Entertainment	0.64	0.84	0.70
News & Politics	0.64	0.75	0.68
Others	0.63	0.83	0.62

Table 5.2: F1 Score using only Twitter features

5.2.3 All Features

In Table 5.3, we show the F1 score of each category for predicting *popularity*, *virality* and *popularity and virality*, using all the available features during the training window. We see a dramatic increase on the F1 score in the prediction of both popular and viral videos. Again, *Music* category achieves the highest F1 score and *News & Politics* the lowest.

	Popularity	**Virality**	**Popularity & Virality**
Music	0.94	0.88	0.90
Games	0.87	0.84	0.78
People & Blogs	0.90	0.87	0.80
Entertainment	0.89	0.86	0.82
News & Politics	0.85	0.79	0.72
Others	0.90	0.85	0.73

Table 5.3: F1 Score using all features

5.3 Feature Importance Analysis

In this section we present the contribution of the 3 most important features during the classification process.

5.3.1 YouTube Features

Table 5.4, shows the 3 YouTube features contributing the most in the prediction of popular, viral and both popular and viral videos, in each category. Their rank is represented by their order on each cell, with the top feature on the first line. Predicting popularity using features from YouTube is trivial and it confirms that the dominating feature is the number of views added on the last day of training. What is more interesting, is the cross system predictions. The dominant feature for predicting virality using only YouTube features varies among the categories. The duration of the videos in *Games* and *News & Politics* category is the feature that contributed the most, the views ratio on the last training day for *People & Blogs* and *Others*, and the likes for *Music* and *Entertainment*. We observe three groups of categories. The ones that depend on likes, the ones that depend on views and the last group that depends on the duration of its videos.

This behavior changes when predicting both popularity and virality. The video duration is substituted by the likes for *Games* category and by the negative sentiment from the comments for *News & Politics* category. Likes are a type of sentiment in which users show their interest on a video, therefore, sentiment dominates the number of views on *Games*, *Entertainment* and *News & Politics*.

5.3.2 Twitter Features

Table 5.5, shows the 3 Twitter features contributing the most in the prediction of popular, viral and both popular and viral videos, in each category. Their rank is represented by their order on each cell, with the top feature on the first line. In contrast to the prediction of popularity using YouTube features, the Twitter features used to predict virality differ among the categories. The dominating feature for *Music* and *Games* is the tweets acceleration, for *People & Blogs* and *Entertainment* is followers and for *News & Politics* is friends.

When it comes to the prediction of popularity we see that the dominating features for all categories switch to the number of friends and followers, except music videos. For the prediction of both popular and viral videos the original tweets dominate among the top features, except games which remain with the number friends feature. We can conclude that music videos depend more on the original posts from the users than the other categories.

Category	Popular	Viral	Popular & Viral
Music	Views added day 2 Views Difference Age Ratio views 2	Likes ratio day 2 Views ratio day 2 Likes difference	Views ratio day 2 Views Difference Views added day 2
Games	Views added day 2 Views Difference Age Ratio Likes 2	Video Duration Likes Difference Likes added day 2	Likes added day 2 Likes Difference Video Duration
People & Blogs	Views added day 2 Views Difference Comments Ratio 2	Views Ratio day 2 Age Views Ratio 1 Comments Acceleration	Views added day 2 Likes Difference Age Ratio Views 2
Entertainment	Views added day 2 Views Difference Age Ratio views 2	Likes added day 2 Likes Difference Video Duration	Likes Difference Likes added day 2 Negative Sentiment
News & Politics	Views added day 2 Views Difference Negative Sentiment	Video Duration Views day 1 Age Ratio Likes 2	Negative Sentiment Age Ratio Likes 2 Likes Acceleration
Others	Views added day 2 Views Difference Views Ratio day 2	Views Ratio day 2 Age Comments Ratio 1 Age Ratio Views 2	Comments Ratio day 1 Likes added day 2 Views Acceleration

Table 5.4: Features contributing the most for predicting virality and/or popularity using only YouTube features

5.3.3 All Features

The table 5.5, shows the 3 features contributed the most in the prediction of popular, viral, and both popular and viral videos in each category. Their rank is represented by their order on each cell, with the top feature on the first line. For predicting only popularity the number of views added on the last training day is the dominating feature, while for predicting only virality the dominating feature concerns tweets, either their count, their acceleration or their difference. Specifically, we see that for most of the categories, the acceleration of tweets has a significant contribution in the classification of viral videos.

For predicting videos that are both popular and viral, the features differ among the categories For *Music* and *Others* category the number of views dominate the tweets acceleration in the prediction of both popularity and virality. For *Games* the number of views and tweets acceleration are substituted by the likes ratio on the last day of training, while the rest categories substitute their features with the original tweets attract on the last training day.

In this table we can see that all categories use different features to predict popularity and/or virality. All the categories use mostly Twitter features to predict both popularity and virality,

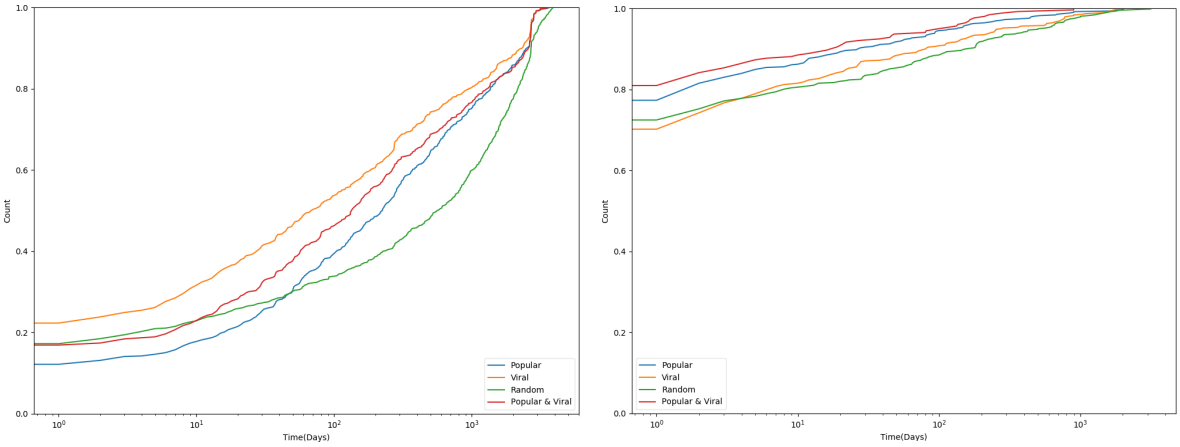
Category	Popular	Viral	Popular & Viral
Music	Original tweets day 1 Original Tweets Acc. Retweets Ratio day 2	Tweets Acceleration Friends Original Tweets Acc.	Original tweets day 1 Original Tweets Acc. User Statuses day 1
Games	Friends Followers User Statuses Acc.	Tweets Acceleration Friends Tweets Difference	Friends Original Tweets Acc. Followers
People & Blogs	Friends Followers Original Tweets Acc.	Followers Tweets Difference User Statuses day 2	Original Tweets Acc. Ratio Original Tweets 2 User Statuses day 2
Entertainment	Followers Friends Original Tweets day 1	Followers Tweets Acceleration Tweets day 2	Original Tweets Acc. Original Tweets 2 Original Tweets Diff.
News & Politics	Friends User Statuses day 1 User Statuses Acc.	Friends User Statuses Diff. User Statuses day 2	Original Tweets Diff. Tweets Ratio day 2 Original Tweets day 2
Others	Friends User Statuses day 1 User Followers	Tweets Difference Tweets Acceleration User Statuses day 2	Original Tweets Diff. Original Tweets day 2 Friends

Table 5.5: Features contributing the most for predicting virality and popularity using only Twitter features

except music videos which use mostly featrues from YouTube. Therefore, we can conclude that music videos depend less from the activity on Twitter. We believe, that this happens due to the fact that music videos are more personal to a user (more original tweets), thus sharing through social media doesn't affect in such a degree their popularity.

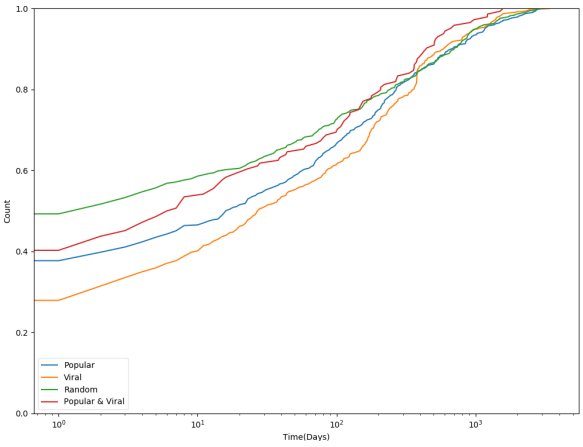
Category	Popular	Viral	Popular & Viral
Music	Views added day 2 Age Ratio views day 2 Views Difference	Tweets Acceleration Tweets added day 2 Views Ratio day 2	Views added day 2 Age Ratio views day 2 Views Difference
Games	Views added day 2 Views Difference Age Ratio Likes 1	Tweets Acceleration Video Duration User Statuses day 2	Likes ratio day 2 User Statuses day 2 Negative Sentiment
People & Blogs	Views added day 2 Views Difference Comments Ratio day 2	Tweets Difference Tweets added day 2 User Statuses day 2	Original Tweets added 2 Age Views ratio 2 Ratio Views day 2
Entertainment	Views added day 2 Views Difference Age Ratio views 2	Tweets acceleration Tweets added day 2 Tweets Difference	Original Tweets added 2 Tweets ratio day 2 Original Tweets Acc.
News & Politics	Views added day 2 Views Difference Original Tweets 2	Tweets added day 2 Views day 1 Followers	Original Tweets Diff. Original Tweets added 2 Retweets day 2
Others	Views added day 2 Views Difference Age Ratio Comments 2	Tweets acceleration Tweets added day 2 Tweets Difference	Views ratio day 2 Original Tweets Acc. Original Tweets 2

Table 5.6: Features contributing the most for predicting virality and/or popularity using only all features

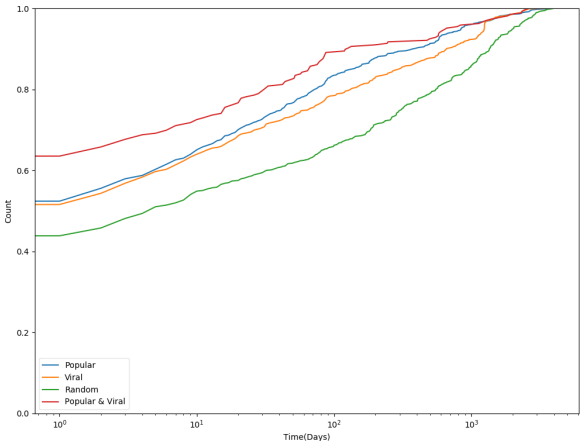


(a) Music

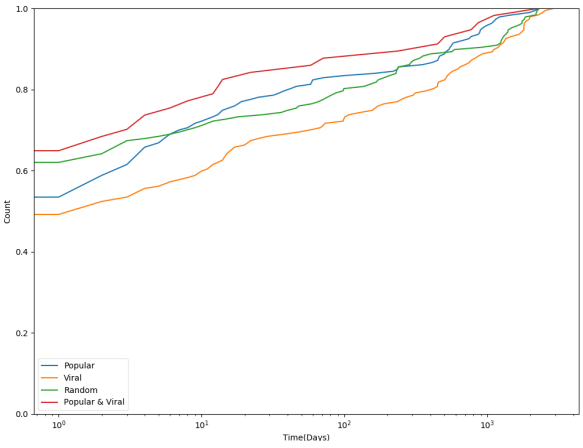
(b) Games



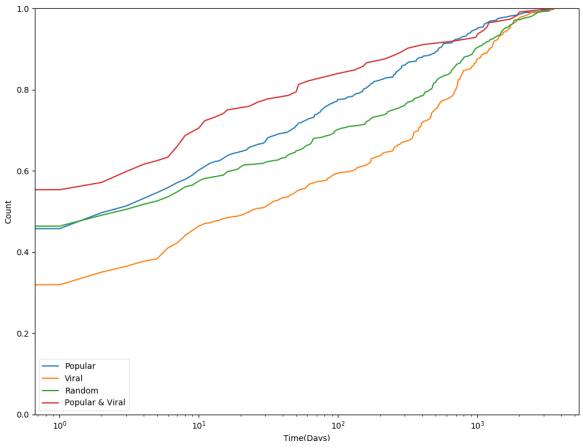
(c) People & Blogs



(d) Entertainment

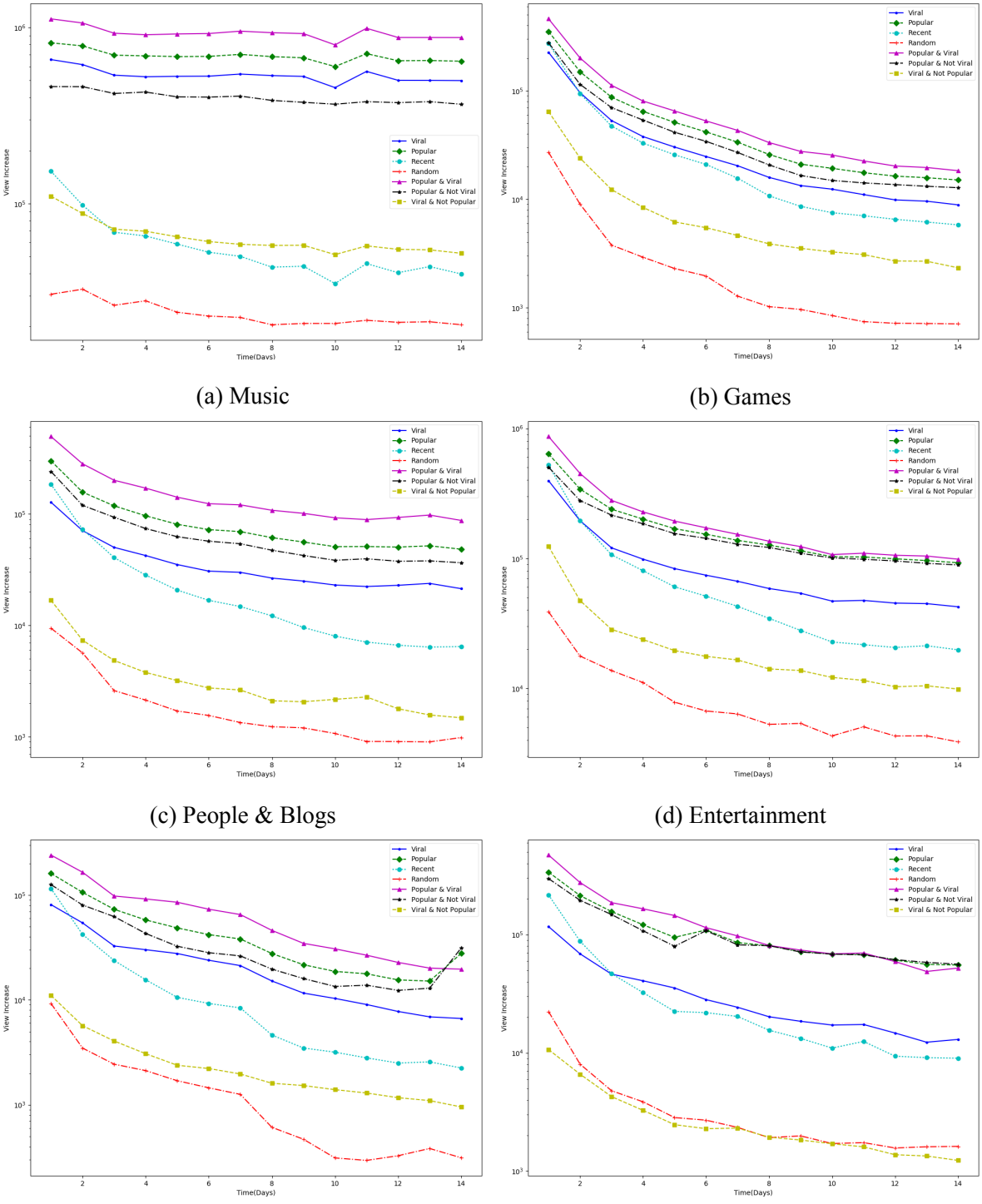


(e) News & Politics



(f) Others

Figure 5.1: Video age distribution for the various categories of videos



(a) Music

(b) Games

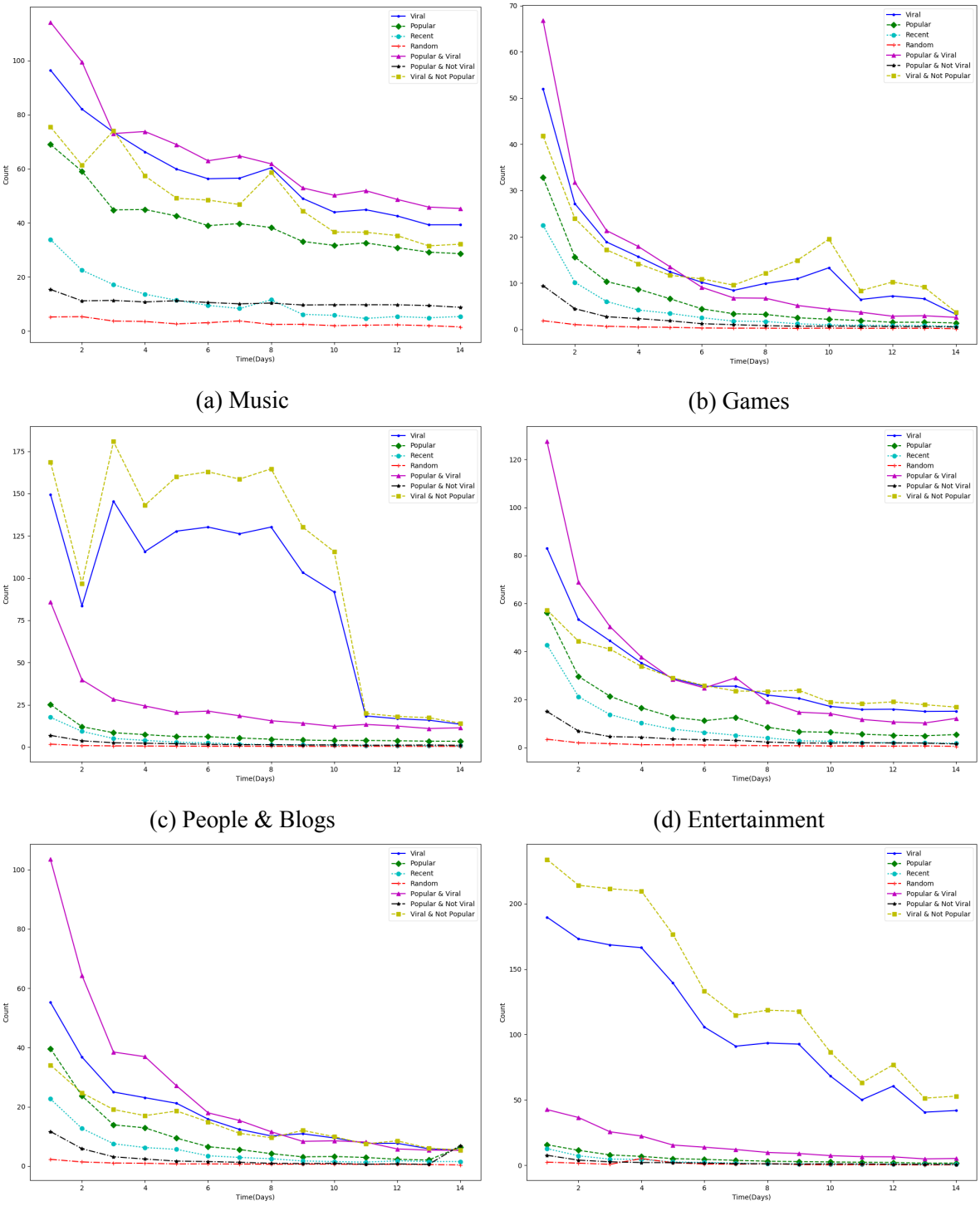
(c) People & Blogs

(d) Entertainment

(e) News & Politics

(f) Others

Figure 5.2: Average daily increase in view count of the various categories of videos



(a) Music

(b) Games

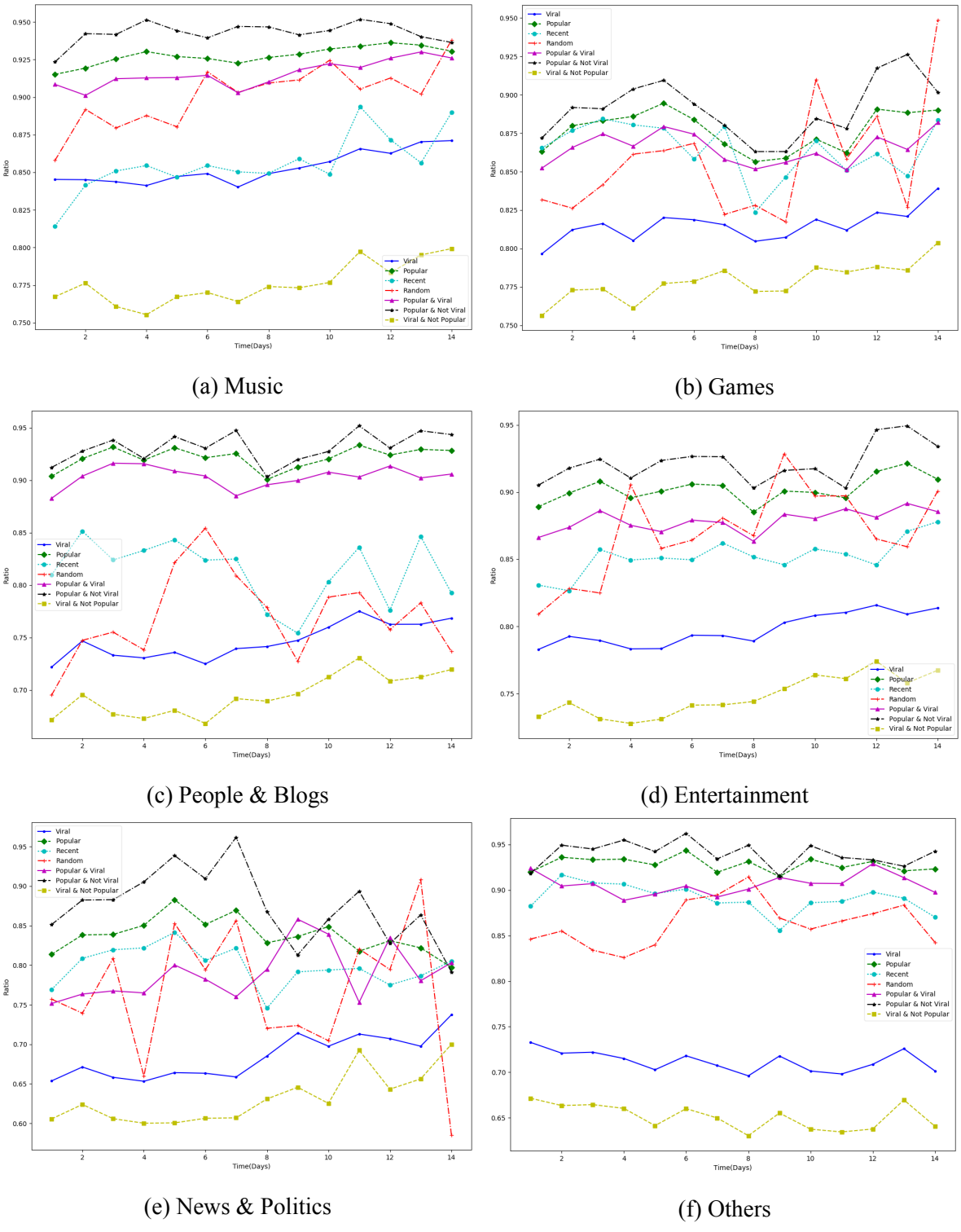
(c) People & Blogs

(d) Entertainment

(e) News & Politics

(f) Others

Figure 5.3: Average daily increase of tweet mentions for the various categories of videos



(a) Music

(b) Games

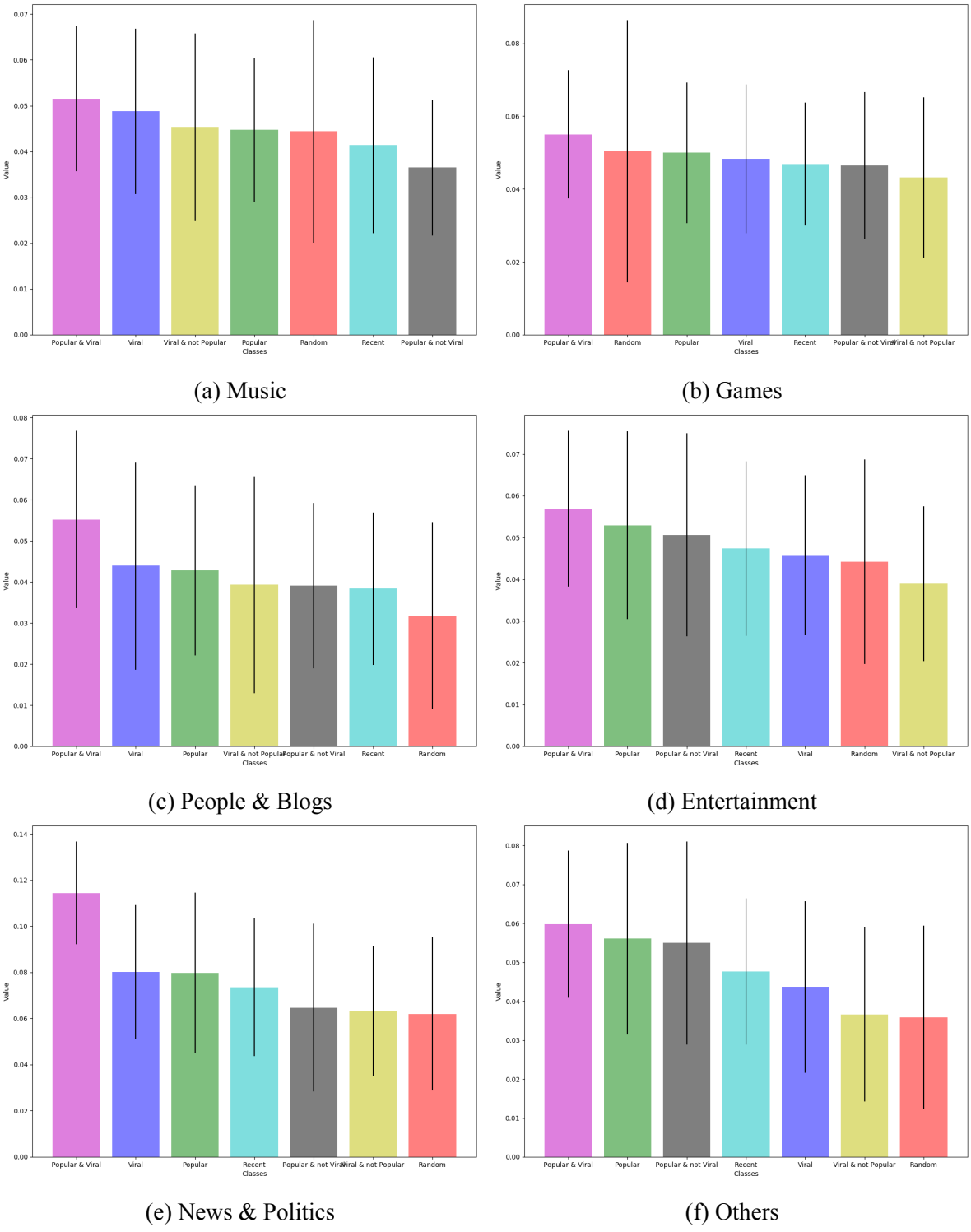
(c) People & Blogs

(d) Entertainment

(e) News & Politics

(f) Others

Figure 5.4: Average ratio between original and total tweets for the various categories of videos



(a) Music

(b) Games

(c) People & Blogs

(d) Entertainment

(e) News & Politics

(f) Others

Figure 5.5: Comments average negative sentiment and error bars for the various categories of videos

Chapter 6

Conclusions & Future Work

Contents

In this chapter, we summarize our findings and discuss the possible future work.

6.1 Conclusions

Our research question has been answered, revealing the differences in the behavior of popular and/or viral videos in regard to their category. The results of this research, confirmed the findings of the previous work [24] and through further investigation we enriched their results with our new findings regarding the behavior of each type of video.

High level characterization, showed a unique behavior for popular and viral videos in each category, specifically for music videos. Despite their age, these videos showed a stable increase in their view count during the period of observation, compared to the other categories that had a gradual decrease in their view count, showing that this type of videos experience popularity for longer periods. Music videos are highly depend on the number of original tweets they gain, showing that these videos are more personal and do not depend on sharing as much as other types of videos. Another key finding, is the contribution of the negative sentiment in the identification of a video as popular and viral for news and politics. Our characterization has shown that the amount of negative sentiment expressed in the comments

of such type, is twice as much as it is found in other types of videos. Finally, our results verified the assumption that videos with different type of content, become popular and viral in a different way.

6.2 Future work

This work scratched the surface of this topic, paving the way for further research. For example, collecting and analyzing data for a longer period can give deeper understanding in the behavior of popular and viral videos. Extending the period, it is possible to capture things that couldn't be analyzed during a two weeks period. In addition to this, another possible future work is to repeat this experiment using a different classifier for predicting popularity and virality. In another thesis [16], it has been shown that Ada Boosting Decision Tree provides better accuracy than Gradient Boosting Decision Tree, for predicting popularity and virality of video content.

Finally, another possible future work is to extend the **Collector** (cf. Section 4.1) in such a way that it retrieves the geo-location of tweets. Having this information, we will be able to to identify popular and viral videos within an area. This is extremely important for CDNs as they can proactively find and store popular videos, close to the area of their customers, resulting to lower latency, and cheaper transmission. This can be achieved using multiple stream listeners with different bounding box.

Bibliography

[1] Matplotlib. https://matplotlib.org/. [Online; accessed May-2017].

[2] MongoDB. https://www.mongodb.com/. [Online; accessed May-2017].

[3] Spark Framework. http://sparkjava.com/. [Online; accessed May-2017].

[4] Twitter. http://whatis.techtarget.com/definition/Twitter. [Online; accessed May-2017].

[5] Twitter Company. https://about.twitter.com/company. [Online; accessed May-2017].

[6] Twitter Stream. https://dev.twitter.com/streaming/reference/get/statuses/sample. [Online; accessed May-2017].

[7] Twitter4j. http://twitter4j.org/en/. [Online; accessed May-2017].

[8] UGC. http://www.webopedia.com/TERM/U/UGC.html. [Online; accessed May-2017].

[9] VADER. https://github.com/cjhutto/vaderSentiment. [Online; accessed May-2017].

[10] YouTube. http://www.gcflearnfree.org/youtube/what-is-youtube/1/. [Online; accessed May-2017].

[11] YouTube API. https://developers.google.com/youtube/v3/. [Online; accessed May-2017].

[12] YouTube Categories. https://gist.github.com/dgp/1b24bf2961521bd75d6c. [Online; accessed May-2017].

[13] Youtube comments. http://edition.cnn.com/2013/09/24/tech/social-media/youtube-comment-upgrade/. [Online; accessed May-2017].

[14] Cisco visual networking index: Forecast and methodology, 2015 to 2020, 2016.

[15] Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.

[16] Giorgos Demosthenous. Prediction of virality and popularity of youtube videos: Machine learning algorithms analysis. *Diploma Thesis*, Department of Computer Science, University of Cyprus, 2017.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[19] Lu Jiang, Yajie Miao, Yi Yang, Zhenzhong Lan, and Alexander G Hauptmann. Viral video style: A closer look at viral videos on youtube. In *Proceedings of International Conference on Multimedia Retrieval*, page 193. ACM, 2014.

[20] Mehryar Mohri. *Foundations of machine learning*. MIT Press, Cambridge, MA, 2012.

[21] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[22] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[23] Scikit-learn. Gradient boosting. http://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting. [Online; accessed 27-May-2017].

[24] David Vallet, Shlomo Berkovsky, Sebastien Ardon, Anirban Mahanti, and Mohamed Ali Kafaar. Characterizing and predicting viral-and-popular video content. In *Proceedings of the 24th ACM International on Conference on Information and*

Knowledge Management, CIKM '15, pages 1591–1600, New York, NY, USA, 2015. ACM.

[25] Georgiou Zacharias. Collector and Analyzer Framework. https://github.com/UCY-LINC-LAB/YouTube-Twitter-Analysis. [Online; accessed May-2017].