



Department of Computer Science

MSc Data Science and Analytics

Academic Year 2020-2021

Sentiment Analysis of Social Media Usernames and Titles on YouTube and Twitch

Jef Ntungila
2046825

A report submitted in partial fulfilment of the requirement for the degree of Master of
Science in Data Science and Analytics

ABSTRACT

This Master Thesis investigates the influence of username and title sentiment on video performance in terms of viewership. This research is performed on Social Media Platforms YouTube and Twitch. Natural language processing data visualisation techniques, sentiment analysis and emotion detection were used to gain an in-depth understanding of the unstructured title and username data. Evidence was found to suggest that title sentiment and title emotions impact viewership.

This evidence was sourced through analysis of the median point estimates of the different emotion and sentiment effects, and their bootstrapped confidence intervals. Kruskal-Wallis Ranked 1-way ANOVA and Tukey Honest Significance Difference Test were used to further complement the analysis. In a consultancy setting where the goal is to increase viewership, the actionable advice would be to consistently optimise for the emotion 'trust' in titles.

These insights culminated in the creation of a Flutter web app which deployed the sentiment analysis models. This web app takes a username or title in and returns the sentiment classification. The findings in this paper can be used by actors that seek to understand sentiment and viewership performance on social media trending tabs.

Key Words: *Sentiment Analysis , Emotion Detection , Social Media , YouTube , Twitch*

ACKNOWLEDGEMENTS

I am grateful to my supervisor Professor Martin Shepperd who was incredibly helpful, insightful and resourceful throughout the whole research project. I am also grateful to my parents Stanislas Ntungila and Monique Mukunayi. Without their Love, Support and Financial Aid, I could have not dedicated a year of my life to the pursuit of the Master of Science in Data Science and Analytics.

I certify that the work presented in the dissertation is my own unless referenced

Signature:

A handwritten signature in black ink, appearing to be 'S. Ntungila', written over a horizontal line.

Date: 25-August-2021

TOTAL NUMBER OF WORDS: 12596

TABLE OF CONTENTS

CHAPTER 1: PROBLEM BACKGROUND, MOTIVATION AND RESEARCH QUESTION	1
1.1 Introduction	1
1.2 Research Aim and Objectives	2
1.3 Research Approach	2
CHAPTER 2: LITERATURE REVIEW	4
2.1 Sentiment Analysis of Social Media	4
2.2 Sentiment Analysis of Names	7
2.3 Sentiment Analysis of Titles	7
2.4 Comparison of Social Media Sentiment Analysis and Emotion Detection Models	8
2.5 Literature Review Summary	9
CHAPTER 3: DATA DESCRIPTION, PREPARATION AND CLEANING	10
3.1 YouTube Data Description	10
3.2 Twitch Data Description	12
3.3 Cloud Computing Data Collection	13
3.4 Processing and Cleaning Data for NLP	13
3.5 Refining Data Granularity	14
3.6 Data Description, Preparation and Cleaning Summary	15
CHAPTER 4: EXPLORATORY DATA ANALYSIS	16
4.1 Username Descriptive Statistics	16
4.2 Spearman Correlation Triangle of YouTube Numerical Features	17
4.3 Visualisation of Titles: Bar Plot	18
4.4 Visualisation of Titles: Word Cloud	19
4.5 Visualisation of Titles: Co-Occurrence Network	19
4.6 Exploratory Data Analysis Summary	22
CHAPTER 5: SENTIMENT ANALYSIS OF USERNAMES AND TITLES	23
5.1 Sentiment Analysis and Emotion Detection: Methodology	23
5.2 Sentiment Analysis and Emotion Detection: Findings	26
5.3 Sentiment Analysis and Emotion Detection: Limitations	30
5.4 Sentiment Analysis and Emotion Detection: Summary	31

CHAPTER 6: ANALYSIS OF SENTIMENT IMPACT ON VIEWERSHIP PERFORMANCE	32
6.1 Analysis of Sentiment Impact on Viewership Performance: Methodology	32
6.2 Analysis of Sentiment Impact on Viewership Performance: Findings	34
6.3 Analysis of Sentiment Impact on Viewership Performance: Limitations	38
6.4 Analysis of Sentiment Impact on Viewership Performance: Summary	39
CHAPTER 7: PROTOTYPE	40
CHAPTER 8: TECHNOLOGY STACK AND DATA MANAGEMENT	42
8.1 Discussion of Technology Stack	42
8.2 Discussion of Code and Data Management	42
CHAPTER 9: CONCLUDING DISCUSSION AND FURTHER WORK	42
9.1 Summary of the Dissertation	44
9.2 Future Research and Development	45
9.3 Personal Reflections	45
REFERENCES	46
APPENDIX A: ETHICAL APPROVAL.	48

CHAPTER 1: PROBLEM BACKGROUND, MOTIVATION AND RESEARCH QUESTION

This chapter introduces the aim and objectives of this research project. Motivation to the research question is provided. It finally discusses the research approach, methodology and dissertation outline.

1.1 Introduction

This research project aims to investigate how username and title sentiment impacts video viewership performance on YouTube and Twitch.

Social media usernames tend to be chosen with nominal realism in mind. Nominal realism is the idea that attributes described in the name extend to the bearer of the name (Kelly, 2020). Examples of these are the usernames 'NEUTRAL DROP' on YouTube and 'LoserFruit' on Twitch. Clickbait social media titles, which are found everywhere on the internet, take sentiment into consideration too. Having a fundamental understanding of what usernames and titles have the best performance in terms of views, is valuable information to all actors that seek to be at the top of trending pages.

Twitch ranks live channels by most viewed streams. The YouTube trending tab is a close reflection of videos on YouTube that are getting the most views. This allows for a quantitative non-probabilistic sampling approach of the best performing videos on both platforms.

An understanding of what usernames and titles result in the most views can be obtained through sentiment analysis. Asghar et al define sentiment analysis as 'the field of study related to the analysis of opinions, sentiments, evaluations, attitudes, and emotions' (Asghar et al, 2015). Emotion detection provides an in-depth understanding of the sentiment.

This understanding could have application potential to all actors that would appreciate their content to perform well e.g., media producers, social media personalities, marketing agencies etc. Second order application of this information are predictions and user-centric recommendations. These are mostly absent on Twitch. The recommendations can be based on user preference and personalities. The social media platforms could profile viewers based on emotions. They then could recommend videos and streams based on emotion preferences e.g., viewers in a melancholic mood could be shown either happy or sad videos.

1.2 Research Aim and Objectives

This project aims to investigate how username and title sentiment impacts video performance on YouTube and Twitch. Individual research on sentiment analysis of social media platforms, names and titles has been performed by various researchers. This is discussed in detail in the literature review section of the project (Chapter 2). Understanding how these different components interact with each other is the extension of previous work done in this field.

The following objectives have been identified that measure the achievement of the aim of the research project:

- Investigate the state-of-the-art of sentiment analysis methods
- Locate, process and clean data using relevant natural language processing (NLP) techniques in order to enable the application of relevant sentiment analysis methods
- Perform exploratory data analysis in order to achieve an in-depth understanding of the data
- Perform sentiment analysis on usernames and titles
- Critically reflect on the accuracy, limitations, interpretability of the data analysis
- Wrap the sentiment analysis methods into a prototype software tool

1.3 Research Approach

1.3.1 Overall Approach

The research question is defined as ‘How does username and title sentiment impact the viewership performance of a video on YouTube and Twitch?’. The approach used to answer the research question can be classified as causal quantitative conclusive research design. It is hypothesised that username and title sentiment is a factor in video viewership performance.

Extending existing research and good software engineering practice, further ensures replicability of research. An example of this is by not reinventing the wheel through the usage of existing machine learning models and lexicons e.g., emote-controlled lexicon developed by Kobs et al (Kobs et al, 2020).

1.3.2 Methods and Dissertation Outline

Data originating from Twitch and YouTube was collected from the 28th of May 2021 till the 28th of July 2021. The YouTube API was used to collect the top 50 daily trending videos worldwide on YouTube on a 6-hour basis. The Twitch API was used to collect the top 100 hourly Twitch streams worldwide. The ingested data rate reflects the API limits imposed by the respective platforms.

A comprehensive literature review is provided in chapter 2. The detailed description of the data pipeline provided in chapter 3 ensures the replicability of the project. Exploratory data analysis was subsequently performed in chapter 4. This was followed by the analysis of sentiment in usernames and titles. Emotion detection in titles was also performed to further the understanding of sentiment. Sentiment Analysis and Emotion detection are discussed in chapter 5.

Chapter 6 provides insight into the statistical techniques used to investigate the impact of sentiment and emotion on viewership performance. A prototype that embedded the sentiment analysis methods was produced. This is discussed in chapter 7. Chapter 8 is a record of the technology and data management used throughout the research project. Finally, chapter 9 provides a summary of this research project. It also discusses possible further work.

1.3.3 Limitations

Sentiment analysis inherently deals with subjective opinionated data. The usage of previous research is preferred throughout the whole project. This minimises the introduction of new bias. The data analysed in this project is public data. Hence confidentiality and anonymity concerns are very limited.

CHAPTER 2: LITERATURE REVIEW

Three types of research has been identified that relates to ‘Sentiment Analysis of Social Media Titles and Usernames on YouTube and Twitch’ namely

- Research that relates to Sentiment Analysis of Social Media as a whole focussing on user generated comments e.g., on YouTube, Twitch and Twitter
- Research that relates to Sentiment Analysis of Names focussing on nominal realism
- Research that relates to Sentiment Analysis on Titles focussing on news headlines

This chapter provides an overview of previous work done in those domains. It then proceeds by providing a comparison of sentiment analysis and emotion detection models covered in the overview.

2.1 Sentiment Analysis of Social Media

Sentiment analysis or opinion mining is the field of study related to the analysis of opinions, sentiments, evaluations, attitudes, and emotions (Asghar et al, 2015). Reagen et al state that online opinion mining has been used for stock prediction, marketing campaign response, global happiness monitoring, to inform government and industry officials. Reagen et al identify three types of sentiment analysis methodologies: dictionary (lexicon) rule-based approaches, machine learning approaches and deep-learning approaches (Raegen et al, 2017). Sentiment can be extracted as either classification of polarity (neutral, negative, positive) or a score of the valence of polarity (Hutto and Gilbert, 2014).

Emotion-detection is a sub-branch of sentiment analysis that seeks to extract finer-grained emotions from textual data. This is opposed to sentiment analysis that provides only a coarse overview of the general polarity of the text (Acheampong et al, 2020). Shivhare and Khethawat state that emotion detection is a content-based classification problem (Shivhare and Khethawat, 2012). Acheampong et al define two types of emotion detection methodologies: discrete emotion models and dimensional emotion models. Discrete emotion models classify emotion into distinct independent classes. Dimensional emotion models assume that emotions are not independent and that they need to be represented in dimensional space (Acheampong et al, 2020).

2.1.1 Sentiment Analysis of Social Media: Twitter

The nature of user generated reactions to ongoing events, in a relatively uniform standardised 280-character format, has made Twitter the most researched social media platform, in terms of sentiment analysis. Applications of sentiment analysis on Twitter include consumer insight discovery (Chamlertwat et al, 2012), stock prediction (Mittaland and Goel, 2012) and election opinion mining (Wang et al, 2012).

An example of work done on Twitter is the production of the sentiment analysis model GloVe-DCNN. Jianqiang et al were aiming to build a model that outperformed the results of a baseline model on Twitter sentiment classification. Jianqiang et al introduce their deep convolutional neural network GloVe-DCNN for sentiment analysis on Twitter. They trained their model on 20 Billion tweets. These tweets were embedded using contextual semantic features and co-occurrence statistical features. AFINN sentiment lexicon is used as labels for tweet sentiment. The sum of the sentiment score of each word in the tweets is used as the final valence polarity score. The researchers state-of-the-art model achieved an F1 score of 87.62%, outperforming the baseline model on five different tweet-based lexicons. This was done through the binary classification of Tweets as either negative or positive (Jianqiang et al, 2018).

2.1.2 Sentiment Analysis of Social Media: YouTube

Less research has been done on user generated text data from YouTube. Google Scholar returned 119,000 search results for the 2 Billion user platform YouTube. It returned 414,000 search results for the 330 Million user platform Twitter.

YouTube user generated text data has the same applications as Twitter user generated text data. This led Uryupina et al to develop a lexicon for YouTube comment sentiment analysis in 2014. The lexicon was restricted to processed comments from videos around the topic 'tablets and automobiles'. The sentiment analysis model that resulted from their research, is the lexicon-based model SenTube (Uryupina et al, 2014).

Uryupina et al argue that Twitter based lexicons are unstable due to the lack of context. They state that their lexicon can differentiate between comments regarding the video and comments about products in the videos. They state that the language used on Twitter is similar to the language used on YouTube. Hence their lexicon provides a reliable tool for sentiment analysis on Twitter (Uryupina et al, 2014).

Asghar et al wrote a technical engineering paper in 2015. It critically analysed and compared various methodologies for YouTube comment sentiment analysis (Asghar et al, 2015). More recently, a team of Italian researchers extended the work done on sentiment analysis on YouTube by analysing Italian YouTube vaccine comments. They were aiming to understand the Italian opinion on vaccines throughout 2017 and 2018.

The project quantified the education and campaigning efforts. The researchers used the lexicon-based sentiment analysis model NRCLex to classify the polarity of comments as negative, neutral or positive (Porreca, Scozzari and Di Nicola, 2020). Although the team only used NRCLex to classify the polarity of comments, NRCLex can also be used for discrete distinct class emotion detection (Acheampong et al, 2020).

2.1.3 Sentiment Analysis of Social Media: Twitch

Twitch (140 Million users) is a smaller platform compared to Twitter (330 Million users) and YouTube (2 Billion users). Hence less sentiment analysis research has been performed on data derived from the platform.

The live streaming platform Twitch has developed over the years its own emoji-like 'emote' based language. This language is referred to by media as twitch-speak. This language is heavily used in the comment sections on Twitch. It is comparable to the language used in a sports stadium. Kobs et al state that 'this language is very different from common English, combining internet slang and gaming-related language with abbreviations, intentional and unintentional grammatical and orthographic mistakes, and emoji-like images called emotes'. Hence traditional sentiment analysis methodologies are not directly applicable to data from Twitch (Kobs et al, 2020).

They introduce the domain specific, emote-controlled and lexicon-based sentiment analysis model, to navigate the intricacies of Twitch. This model is used in combination with the sentiment analysis model Vader. Precedence is given to the emote-controlled model. Where the emote-controlled model fails to return a valence polarity score due to lack of emotes, Vader is used (Kobs et al, 2020).

Vader is a rule-based sentiment analysis model introduced by Hutto and Gilbert in 2014. The rule-based model Vader had a F1 score of 96% classifying tweets as positive, negative or neutral (Hutto and Gilbert, 2014). This is compared to the deep-learning model GloVe-DCNN that had a F1 score of 87.62% on the same assignment (Jianqiang et al, 2018). Hence, rule-based approaches are often preferred for their simplicity of understanding, computational efficiency and extensibility over other methods.

Analogous to Kobs et al, Reis also proposed a rule-based lexicon approach to Twitch comment sentiment analysis. Both Reis and Kobs et al test their models on comments made during the livestream of the Blizzard Entertainment 2018 convention. Both Reis and Kobs et al provide a time series analysis of the valence polarity scores of comments during the live stream (Reis, 2020; Kobs et al, 2020). This provides a clear indication of sections of the livestream appreciated and less appreciated by viewers.

In 2019, Kim et al used sentiment valence polarity scores as a feature to predict whether a user was subscribed to a Twitch channel. They used the LIWC sentiment analysis model to produce the sentiment valence polarity scores (Kim et al, 2019). LIWC is a lexicon-based approach to sentiment analysis that was developed in 1996. Hutto and Gilbert describe LIWC as 'a comprehensive, high-quality lexicon' but 'used with little regard for its actual suitability to the social media domain'. They state that Vader improves on the benefits of LIWC. Hutto and Gilbert also state that Vader performs better on social media and that it generalises better to other domains than LIWC (Hutto and Gilbert, 2014).

In 2020, Kobs et al reproduced the work performed by Kim et al in 2019. They also used sentiment valence polarity scores as a feature in predicting the subscription status of a Twitch viewer. Unlike Kim et al that used the sentiment analysis model LIWC, they use the domain specific emote-controlled model and Vader (Kobs et al, 2020; Kim et al, 2019).

2.2 Sentiment Analysis of Names

Nominal realism is defined as the conviction that the name of an object affects the characteristics of an object. Nominal realism is observed in peoples naming traditions in names such as Brave or Strongman. It is also observed in place names traditions (Kelly, 2000) and in commercial setting names in e.g., Business and Sports.

Sentiment Analysis on US business names showed a ratio of seven to one positive to negative biased names. Kelly states that this bias translates to business performances. Businesses with a positive business name outperform business with a negative business name in terms of revenue (Kelly, 2017). In January 2021, Kelly extended nominal realism research by performing sentiment analysis on surnames. He proved that due to nominal realism, surnames with a negative sentiment occur far less than surnames with a positive sentiment (Kelly, 2021).

Kelly employed the sentiment analysis model AFINN, also known as ANEW, to classify the polarity of the surnames (Kelly, 2021). AFINN is also the model that was used by Jianqiang et al to label the 20 Billion tweets to produce the convolutional neural network sentiment analysis model GloVe-DCNN (Jianqiang, 2018).

AFINN is a lexicon-based model developed by Nielsen in 2011 that maps words to valence polarity scores (Nielsen, 2011). AFINN is hence suitable to sentiment analysis of names due to it taking only a single word as input. This is opposed to Vader that takes a sentence in as input. Hutto and Gilbert criticise AFINN for being insensitive to common sentiment-relevant lexical features in social text. Vader, unlike AFINN, can differentiate sentiment through capitalisation and punctuation (Hutto and Gilbert, 2014).

2.3 Sentiment Analysis of Titles

Sentiment Analysis is used to evaluate news headlines e.g., for financial prediction applications or recommender systems. Loureiro et al applied sentiment analysis to news headlines by assigning sentiment valence polarity scores to named entities. This was done by gathering data from their Wikipedia pages or blogs. This was an extension of research performed by Vinagre. Vinagre classified news headlines based on emotion to recommend news articles to users based on their preferences and personalities (Loureiro, Marreiros and Neves, 2011).

2.4 Comparison of Social Media Sentiment Analysis and Emotion Detection Models

	Vader	Emote-Controlled	Afinn	NRClex	SenTube	GloVe-DCNN
Type model	Lexicon rule-based Model (Hutto and Gilbert, 2014)	Lexicon rule-based model/ deep-learning based model	Lexicon rule-based model	Lexicon Rule-Based Discrete Emotion Model (Acheampong et al, 2020)	Lexicon rule-based model	Deep Convolutional Neural Network
Author	Hutto and Gilbert	Kobs et al	Nielsen	Mohammad and Turney	Uryupina et al	Jianqiang et al
Training of model	Crowd-sourced sentiment ratings through Amazon Mechanical Turk. Raters had training. Results verified through data quality checking. (Hutto and Gilbert, 2014).	Sentiment ratings were crowd-sourced through survey on gaming related twitter account and relevant reddit forums + supplemented by Vader (Kobs et al, 2020)	Words were manually scored by the author. Word list valence polarity score has a similar distribution to other reputable lexicons. It also has a Spearman correlation of 0.5 - 0.6 with their valence polarity scores (Nielsen, 2011)	Sentiment ratings were crowd-sourced through Amazon Mechanical Turk. Raters had training (Mohammad and Turney, 2013)	Four external raters; high inter reliability rate (Uryupina et al, 2014)	Afinn sentiment lexicon is used as labels for tweets sentiment. The sum of the sentiment score of each word in the tweets is used as the final valence polarity score (Jianqiang et al, 2018)
Dataset	- emojis - commonly used slang - punctuation nuances -4,200 tweets -10,625 film review sentences -3,708 sentence snippets from product reviews - 5,190 sentence snippets from the New York Times opinion editorials (Hutto and Gilbert, 2014)	3 Billion Twitch Comments (Kobs et al, 2020)	Tweets relating to the United Nation Climate Conference in 2009, extended with internet slang, internet swear words, Wiktionary synonyms	Terms from Roget Thesaurus that appear over 120,000 times in the Google n-gram corpus., amounts to 14,000 words (Mohammad and Turney, 2013), Expanded to 27,000 words using synonyms from WordNet (Bailey, 2019)	The lexicon was derived from 45,000 comments from YouTube videos around the topics Tablets and Automobiles (Uryupina et al, 2014)	20 Billion Tweets embedded using contextual semantic features and co-occurrence statistical features (Jianqiang et al)
Scale	-4 to 4 (Kobs et al, 2017)	-1 to 1 (Kobs et al, 2017)	-5 to 5 (Reagen et al, 2017)	-5 to 5 (Reagen et al, 2017)	Multi classification as positive neutral negative	Binary classification of tweets as positive or negative (Jianqiang et al, 2018)
Performance	F1 score = 96% at classifying tweets as neutral positive or negative (Hutto and Gilbert)	Rule-based method F1 score = 60.4%, CNN F1 score = 64.3% (Kobs et al, 2020)	Metrics are not directly provided, however extensive comparison and correlation metrics with other sentiment analysis methods are given	F1 score = 88.93% term level task, F1 score = 69.02% sentence level task (Mohammad and Turney, 2013)	66.35% accuracy on multi classification (Uryupina et al, 2014)	F1 score = 87.62% (Jianqiang et al, 2018)
Advantages	Performs extremely well on social media. Generalises well across various domains. Easily adaptable due to its rule-based approach. Computationally inexpensive. Users of LIWC, a sentiment analysis lexicon published in 1996, can readily use it (Hutto and Gilbert, 2014)	Domain Specific	Can provide a valence polarity score on only a word - hence useful for sentiment analysis on names	Dictionary based approaches are domain agnostic and generalises well (Raegen et al, 2017)	It can differentiate between comments regarding the video itself or the product reviewed in the video.	Detection of latent features
Limitations	Fails to detect misspellings, limited to words in lexicon, fails to detect sarcasm (Delancey, 2020)	Struggles with misspellings, limited to lexicon (Kobs et al, 2020)	Insensitive to common sentiment-relevant lexical features in social text (Hutto and Gilbert, 2014)	Could evaluate a word incorrectly out of context (Raegen et al, 2017)	Domain specific to certain products and hence will not generalise well. Valence of polarity is not considered.	Difficult to interpret and computationally expensive
Example of practical usage	Used to analyse Twitch comment sentiment (Kobs et al, 2020)	Used as part of feature in predicting paid subscription on Twitch (Kobs et al, 2020)	Kelly used Afinn to extract nominal realism from usernames (Kelly, 2021), Jianqiang et al used Afinn to classify tweets (Jianqiang et al, 2018)	Porreca, Scozzari and Di Nicola used NRClex in their research into sentiment regarding Italian vaccination videos on YouTube (Porreca, Scozzari and Di Nicola, 2020).		

Table 1 - Comparison of Social Media Sentiment Analysis and Emotion Detection Models

2.5 Literature Review Summary

The literature review covers different applications of sentiment analysis on social media. Comment sentiment analysis on YouTube can be used to understand user sentiment regarding commercial products (Uryupina et al, 2014). It can also be used to understand marketing efforts (Porreca, Scozzari and Di Nicola, 2020). Another application of sentiment analysis is paid subscription prediction on Twitch (Kobs et al, 2020; Kim et al, 2019). Kelly appreciates nominal realism in surnames through sentiment analysis (Kelly, 2021). Kelly also states that nominal realism impacts business performance (Kelly, 2017).

It is observed that rule-based methodologies are preferred to machine learning and deep learning alternatives for sentiment analysis on social media. An explanation for this, is their state-of-the-art performances e.g., Vader; F1 score=96% on tweet polarity classification (Hutto and Gilbert, 2014).

CHAPTER 3: DATA DESCRIPTION, PREPARATION AND CLEANING

Data from YouTube and Twitch was ingested from the 28th of May till the 28th of July. 139,500 observations were ingested from Twitch. 11,783 observations were ingested from YouTube. Every 6 hours, the top 50 trending YouTube videos were sampled using the YouTube Data API. The top 100 Twitch streamers worldwide and their stream titles were sampled on an hourly basis. This was done using the Twitch API. The rate of ingestion reflects the limits set by the APIs on their respective platforms.

3.1 *YouTube Data Description*

This project investigates the impact of username and title sentiment on viewership. It is theorised that the YouTube trending tab would be a representation of the best performing videos in terms of viewership. YouTube defined the trending tab in 2015 as ‘Popular videos that were embedded in the web’s most popular websites.’ YouTube also stated that ‘a significant number of people needed to have viewed the video externally, in addition to views originating from youtube.com.’ (YouTube Trends, 2015). Google further clarified that it considered the following signals in defining a video as trending: view count, temperature (i.e., viewership pace), where views are originated from, the age of the video, and relative performance compared to other videos uploaded to the same channel.

The consideration of external viewership of videos embedded in other websites is analogous to the Google PageRank Algorithm. The PageRank Algorithm investigates the relevance of a page by verifying how frequently a page has been externally linked to. It considers a page more relevant if it has more external links referring to it (Rogers, 2002). This is comparable to the YouTube trending tab considering a video trending if views are originating from external sources. Hence a video with lower viewership, but with more external links might be ranked higher on the trending tab than a video with more viewership. A video with a high rate of viewership might not appear on the YouTube trending tab, if it does not have enough external viewership.

The youtuber Coffee Break found that the Associated Press channel had on average the least number of views before their channel appeared on the YouTube trending tab (Coffee Break, 2019). This is to be expected as other media refer to the Associated Press. Nevertheless, extreme high viewership on YouTube will still result in a video being considered for the YouTube trending tab. The high viewership will also result in external discourse. Hence a positive feedback loop is created. YouTube also reserves 50% of the YouTube trending tab for their own creators. These are people that primarily produce videos for the platform (Alexander, 2019).

YouTube does not publish a worldwide trending tab. Instead, they publish trending tabs for each individual country. Twitch on the other hand does not publish top trending streams on a country per country basis. They only list streams in their order of viewership worldwide. 'Kworbs YouTube worldwide trending tab' was used as reference in order to perform a comparative analysis with data from Twitch. Kworbs aggregates all the data that comes from the different individual YouTube trending tabs. It then orders the videos by considering the rate of viewership, and in how many countries the video is trending.

The YouTube API was employed to further complete the data with the channel name, publishing time, like count, dislike count, comment count and view count features. An exclusion criterion was used to filter for titles composed of only ASCII characters and English words.

Associated data

Feature Name	Feature Description	Example
username	The YouTube channel name.	EminemVEVO
video_title	The YouTube video title.	Eminem - Killer (Remix) [Official Audio] ft. Jack Harlow, Cordae
publish_time	Time stamp of when the video became available to the public.	2021-05-28T04:00:12Z
view_count	YouTube user has viewed the video for at least 30 seconds (Funk, 2020).	2,456,653
comment_count	How many comments are on a given video.	21,173
like_count	How many likes are on a given video.	230,404
dislike_count	How many dislikes are on a given video.	2,213
api_call_time	Time stamp of when was the data ingested.	2021-05-30T02:37:08.192447

Table 2 – YouTube Data Description

3.2 Twitch Data Description

The top Twitch streams are necessary to investigate the best performing streams. Twitch ranks streams by concurrent viewers. The platform sorts by default the channels from high to low viewership worldwide. The Twitch API was used to request the top 100 English streams ranked by viewership. The following features were retained: channel name, game played, stream title, viewer count, and finally the start time of the stream.

The observation is made that ‘twitch-speak’ is used in Twitch stream titles. Twitch-speak is different from standard English. It uses emotes i.e., spelled out case-sensitive emoji. Twitch-speak also uses internet and gaming slang combined with intentional and unintentional grammatical mistakes. It evolved from the need to rapidly communicate emotions and short sentences in chatrooms to the Esports athletes, gamers and streamers. Hence the comparison is often made between non-standardised twitch-speak and the language used by large audiences in stadiums (Kobs et al, 2020).

An example of a stream title found in the dataset is ‘{NO-PIXEL} RANK 1 CEO BUSINESS OWNER DOMINATES THE PLEB WORLD AND EBCOMES THE TRUE GRINDFATHER (GONE WRONG) (LAPTOP BUSTAH LULW) (EPIC FAIL)’. This was used by Twitch streamer xQcOW for his livestream of the game Grand Theft Auto V that started at 07:25:11 BST on the 8th of June 2021. An analysis of the title is given below:

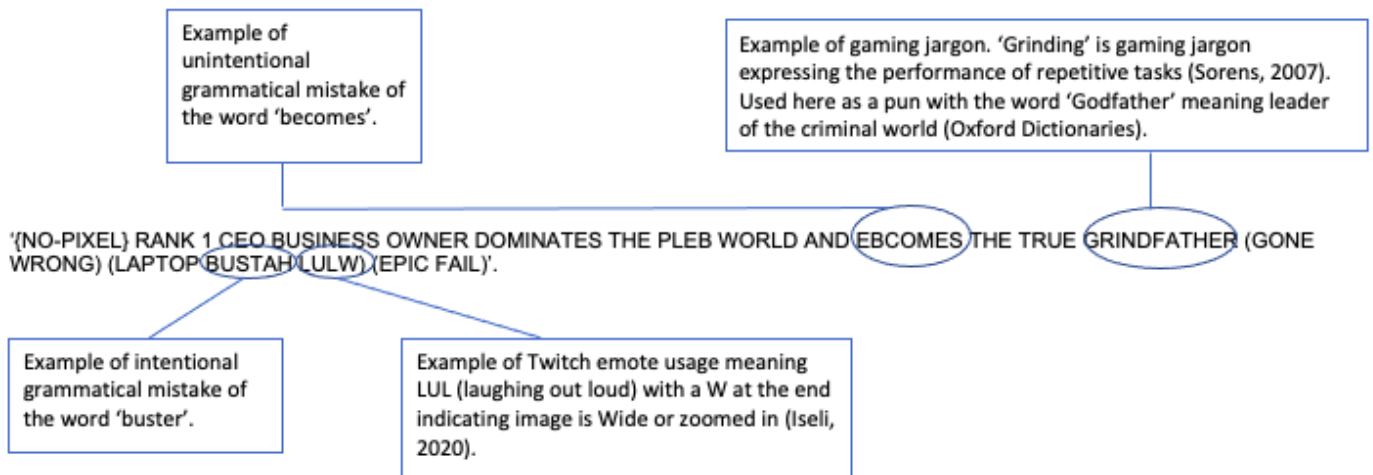


Figure 1 – Example of Twitch Stream Title

Associated data

Feature Name	Feature Description	Example
user_name	The Twitch channel name.	cloakzy
game_name	The game title currently being played by the streamer.	Call of Duty: Warzone
title	Stream title given by the user.	verdansk with my famous rich friends 5.4KD+ !loadouts
viewer_count	Concurrent viewers watching the stream (Twitch, n.d.).	11,161
started_at	Time stamp of when the stream went live.	2021-05-28 15:56:45
api_call_time	Time stamp of when was the data ingested.	2021-05-28T18:06:02.606835

*Table 3 – Twitch Data Description***3.3 Cloud Computing Data Collection**

Cloud computing on the Google Cloud Platform was performed to automate the data collection process. The data collection schedule was programmed using CRON and Google Cloud Scheduler. The scripts prepared in the development environment were updated for production. They were then deployed as cloud functions on the Google Cloud Platform. These cloud functions ran at a time set by Google Cloud Scheduler. The cloud functions pulled the data from YouTube and Twitch and pushed the data to GitHub for storage before processing for NLP.

3.4 Processing and Cleaning Data for NLP

A rule-based approach was employed to process YouTube and Twitch usernames. This was performed in order to enable the extraction of nominal realism and sentiment from usernames. An exclusion criterion was set to omit stop words found in usernames.

The English words were extracted from usernames by finding the set intersection between English words from the Natural Language Toolkit (NLTK) dictionary and the usernames. E.g., The username 'AdeptTheBest' and the NLTK dictionary set intersection returned eleven words namely: ['st', 'th', 'theb', 'de', 'bes', 'adept', 'depth', 'best', 'ade', 'es', 'ad'].

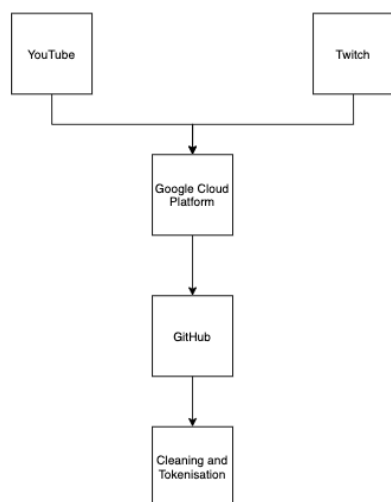


Figure 2 – Data Processing Flow Chart

YouTube and Twitch titles were processed by removing html characters from the documents. An exclusion criterion was set to keep only alphanumeric characters. The documents were subsequently tokenised. Stop words were omitted for data visualisation purposes, but kept for sentiment analysis and emotion detection. Finally, leading and trailing spaces were removed from tokens. The decision was made to not stem or lemmatise the tokens. Stemming and lemmatisation can improve NLP applications when dealing with sparse data, but exploratory data analysis demonstrated the richness of the YouTube and Twitch title corpus.

3.5 Refining Data Granularity

YouTube and Twitch data was refined in order to obtain a better understanding of the data. Data from Twitch was merged with 2019 games sales data. This game sales data was sourced from vgchartz.com. The game sales data contained a game genre feature. The game genre feature was used to label the genre of games played on Twitch. The remaining top video games from 2021 were manually labelled. This allowed the data to be filtered based on the genre of game played. This ensured better analysis and visualisations.

In 2018, the youtuber Coffee Break analysed 40,000 videos on the YouTube trending page. 70 people were crowdsourced to label YouTube channels by genre. Six different video genres were used to label 2,195 YouTube channels. The genres are: Youtuber, Commercial, Trailer, Music, Traditional Media, Viral Youtuber. A 'viral youtuber' is a youtuber with less than 10,000 subscribers (Coffee Break, 2019). The viral youtuber and youtuber genre were merged for this research. The crowdsourced labels were merged with the recently ingested YouTube data. This resulted in the labelling of channels that appeared in both datasets.

The remaining YouTube channels were labelled using a supervised extreme-boosted-decision-tree machine learning algorithm. The boosting algorithm was initially developed and documented by Chen and Guestrin. They describe its state-of-the-art performance on sparse data (Chen and Guestrin, 2016). The algorithm was trained and tested on the sparse count vectorised matrix of the text used in YouTube titles.

The matrix had each word that appeared in the titles as column vectors. The rows of the matrix consisted of word count row vectors. These row vectors contained the counts of the words appearing in each YouTube video title. The training data was 80% of the labelled YouTube titles, for the period 28th of May 2021 till the 28th of July 2021. The testing data was 20% of the labelled YouTube titles for that same period. The

target was the genre of the video title. This algorithm was configured to use 100 epochs and 100 decision trees. The multi classification algorithm correctly classified 95% of the genres of a video based on its title.

The algorithm was used to classify the remaining unlabelled YouTube videos by genre. 55% (6,512) of the observations from a total of 11,783 observations on the YouTube trending tab were labelled with genre 'youtuber'. This is in accordance with the comments made by YouTube CEO Susan Wojcicki in 2019. She stated that half of the videos on the YouTube trending tab would be from creators that make their money primarily from YouTube (Alexander, 2019).

Additional numerical features such as view count, likes, dislikes and comment count could have been used to further improve the model. The performance of the model on textual data was satisfactory enough for the purposes of this research. An unsupervised learning algorithm could have been used as an alternative methodology for refining the granularity of data. A supervised methodology was chosen ahead of an unsupervised methodology due to the availability of labelled data.

3.6 Data Description, Preparation and Cleaning Summary

Username and title data from YouTube and Twitch was ingested from the 28th of May till the 28th of July. 11,783 observations were ingested from YouTube. 139,500 observations were ingested from Twitch. The ingestion of the data was automated through the usage of Cloud Functions and CRON jobs on the Google Cloud Platform. The raw data was then pushed to GitHub for storage. The granularity of the data was refined by adding the genre feature in order to better analyse the data. This chapter found that YouTube had kept their promise in allocating half the trending tab to their creators. Finally, the data was cleaned and tokenised for NLP visualisation and analysis purposes.

Please find code used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/code/Sentiment_Analysis_of_Usernames_and_Titles_on_YouTube_and_Twitch_DATA_WRANGLING.ipynb

Please find data used for this chapter at:

https://github.com/JefNtungila/master_thesis

CHAPTER 4: EXPLORATORY DATA ANALYSIS

The Twitch raw data for this analysis had 139,500 observations. The YouTube raw data for this analysis had 11,783 observations. The data was from the period 28th of May 2021 to the 28th of July 2021.

This research investigates the influence of sentiment on video performance in terms of viewership. The data was hence filtered to keep the observation of a video with the highest viewership. This corresponds to the last observation of a video on the YouTube trending tab. This corresponds to the observation with the highest viewership and same username, title and stream start on Twitch. Username summary statistics and title visualisations were produced as part of the exploratory data analysis.

4.1 Username Descriptive Statistics

Username descriptive statistics were produced in order to obtain a high-level understanding of the username data. Unique usernames were counted. This was followed by the production of statistics that counted usernames that contained English words. Median statistics were produced that described the average words contained in usernames.

	count unique usernames	count usernames with words	percentage usernames with words	median words per username
YouTube	1,303	1,231	94%	8
Twitch	4,062	3,672	90%	6

Table 4 – Username Statistics Table

94% of usernames on YouTube contained words. 90% percent of usernames on Twitch contained words. Median words per username is higher on YouTube than on Twitch. YouTube usernames contained a median of 8 English words. Twitch usernames contained a median of 6 English words. It is observed that the right-skewed distribution of count of words in usernames is very similar for both platforms.

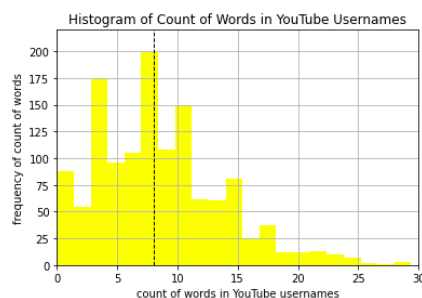


Figure 3 – Histogram of Count of Words in YouTube Usernames

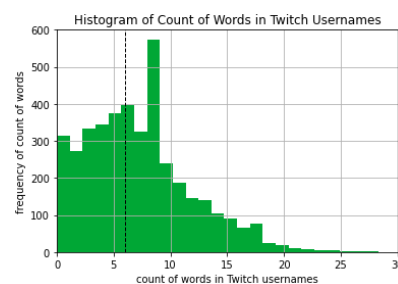


Figure 4 – Histogram of Count of Words in Twitch Usernames

4.2 Spearman Correlation Triangle of YouTube Numerical Features

A Spearman correlation triangle was produced of the view count, comment count, like count and dislike count features. Spearman rank order correlation was chosen ahead of Pearson because count features rarely have a symmetrical distribution. Correlation is observed between all variables. Hence the decision was made to only analyse the impact of sentiment on the viewership feature.

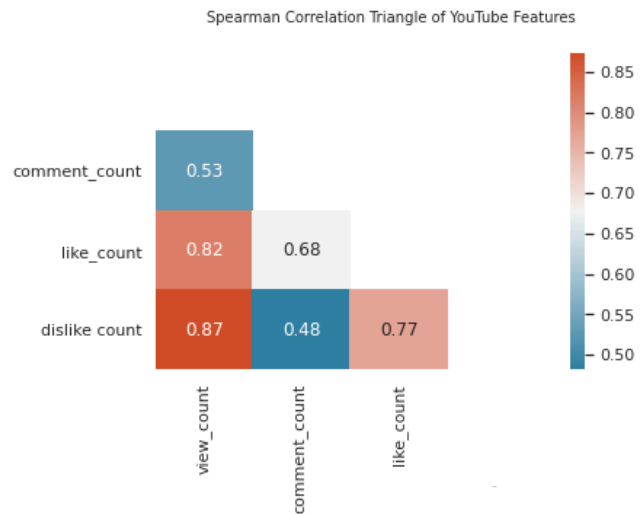


Figure 5 – Correlation Triangle of YouTube Features

Strong Spearman correlation is found between the view count and like count features: Spearman correlation = 0.82, p-value = 0. View count and dislike count also share strong correlation with Spearman correlation = 0.87, p-value = 0. This finding is in accordance with the results observed by Niture at the Dublin Business School, as part of their analysis of the YouTube trending tab (Niture, 2021).

It is appreciated that there is a lower correlation between view count and comment count (Spearman correlation 0.53, p-value = 0). Deep diving into the relationship of viewership and comment engagement for the different genres reveals interesting findings. There is a Spearman correlation of 0.63 between viewership and comment engagement for the 'Traditional Media' genre (p-value = 0). There is a Spearman correlation of 0.81 between viewership and comment engagement for the 'Trailer' genre (p-value = 0). The strong correlation of 'Trailer' and 'Traditional Media' viewership with comment engagement on the YouTube trending tab is explained by their novelty factor.

There is a Spearman correlation of 0.49 between viewership and comment engagement for the 'Youtuber' genre (p-value = 0). It is also observed that there is weak correlation between viewership and comment engagement for the 'commercial' genre (Spearman correlation 0.32 and p-value = 0.01). The genre 'Music' had a viewership and comment engagement Spearman correlation of 0.55 (p-value = 0).

Although p-values are used in this research to understand the significance of correlation, it is noted that p-values are limited in explaining statical validity of results (Nuzzo, 2014). The decision was made to include p-values due the difference in values between the genres that had a high correlation and genres that had weak correlation between the view feature and comment feature.

4.3 Visualisation of Titles: Bar Plot

Bar plots of the most frequently used words in YouTube and Twitch titles were produced. The most frequent word in popular YouTube videos was 'official'. The most frequent word in popular Twitch streams was 'day'.

Emotes were not found in the most frequent words used in Twitch titles. This is inconsistent with research performed by Kobs et al on a corpus of 3 Billion unlabelled Twitch comments. They found that 10 of the top 20 words were twitch-speak emotes (Kobs et al, 2020). The conclusion can be made that emotes are used at a higher rate in Twitch comments than in Twitch titles. The word 'best' was used more than 1000 times in Twitch titles. This can be interpreted as indication of positive sentiment polarity in titles.

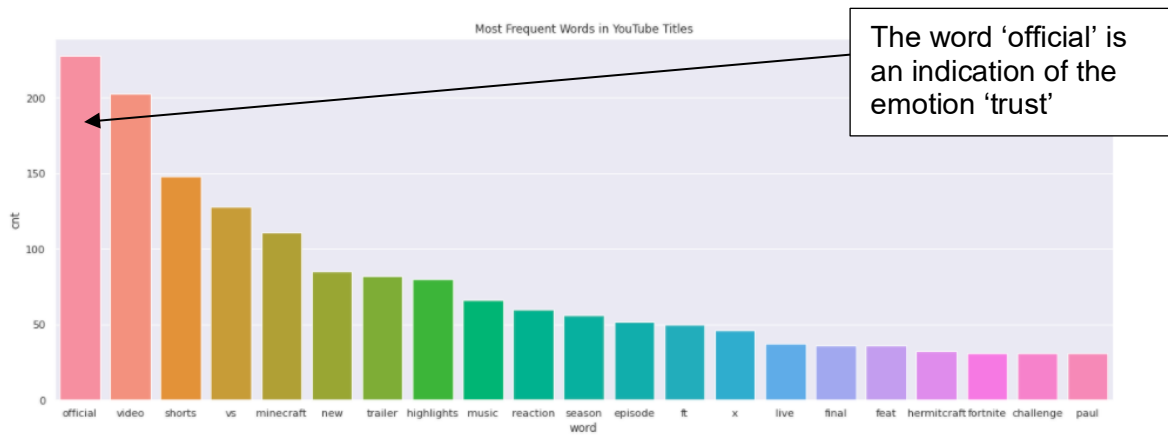


Figure 6 – Most Frequent words in YouTube titles

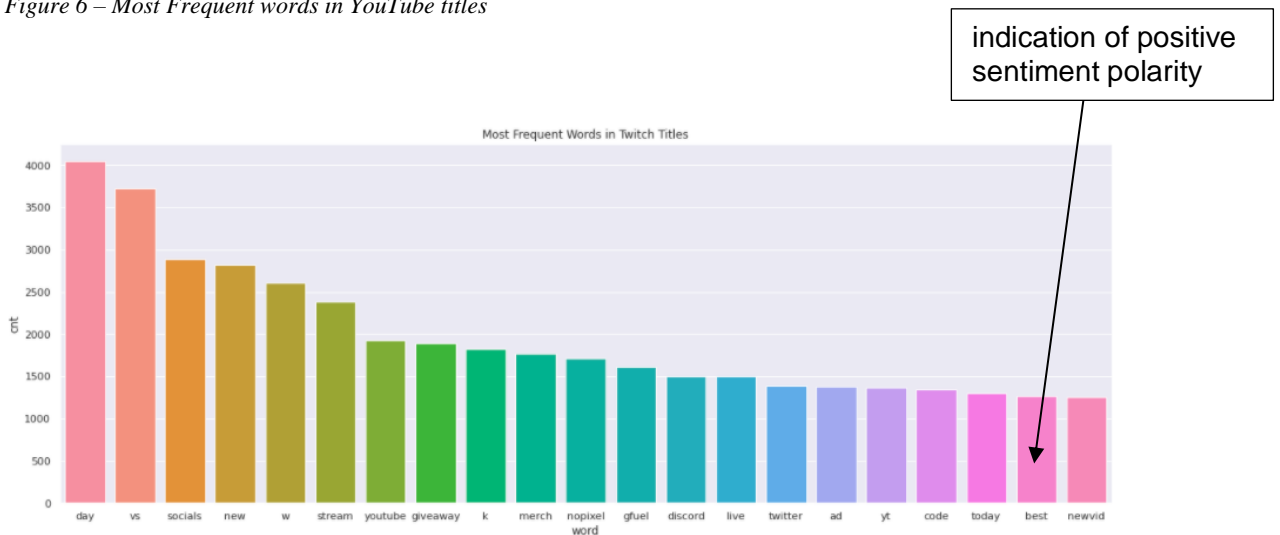


Figure 7 – Most Frequent words in Twitch titles

4.4 Visualisation of Titles: Word Cloud

Word clouds of the most frequent words used in YouTube and Twitch titles were generated. Inspiration was taken from Kobs et al. They created a word cloud in the Twitch logo shape. The team used Twitch subscription data for their word clouds (Kobs et al, 2020). Words such as 'new' and 'highlight' indicate the presence of surprise, an emotion that can be algorithmically detected.



Figure 8 – Word Cloud of Most Frequent Words used in YouTube Titles

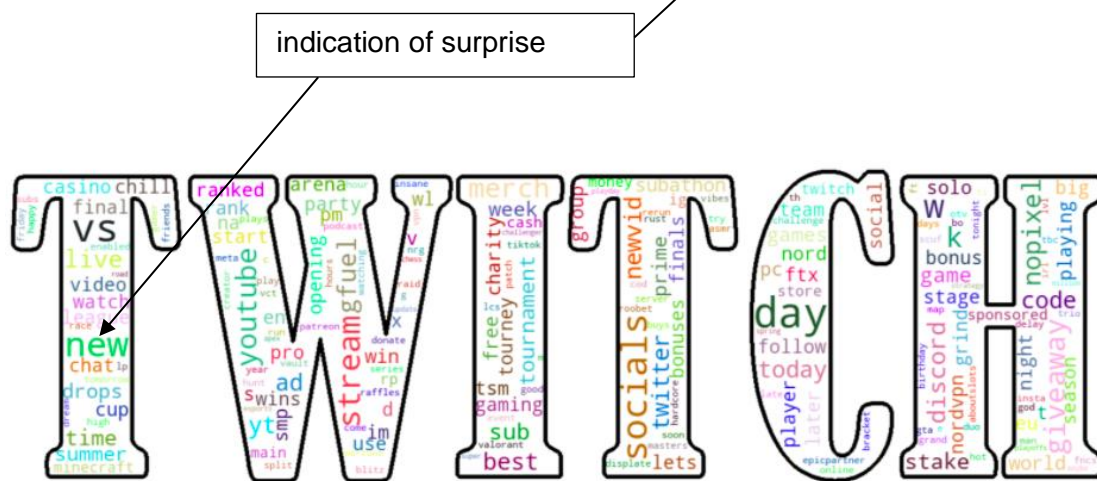


Figure 9 - Word Cloud of Most Frequent Words used in Twitch Titles of the 'Just Chatting' type

4.5 Visualisation of Titles: Co-Occurrence Network

Porreca, Scozzari and Di Nicola produced co-occurrence networks in their research into sentiment regarding Italian vaccination videos on YouTube. The team defined the degree of co-occurrence using the spatial distance computed by the Jaccard index. The values were then stored in a co-occurrence matrix (Porreca, Scozzari and Di Nicola, 2020). The co-occurrence matrix was then converted into a network with algorithmically positioned nodes. Bigger nodes mean higher frequency of word usage. Thicker lines between nodes mean stronger connection between word usage.

This research followed the same strategy as part of exploratory data analysis into YouTube and Twitch titles. The observation is made that usernames are frequently mentioned in YouTube titles e.g. 'Paul'. Another observation is that titles on Twitch often contain creator social media information.

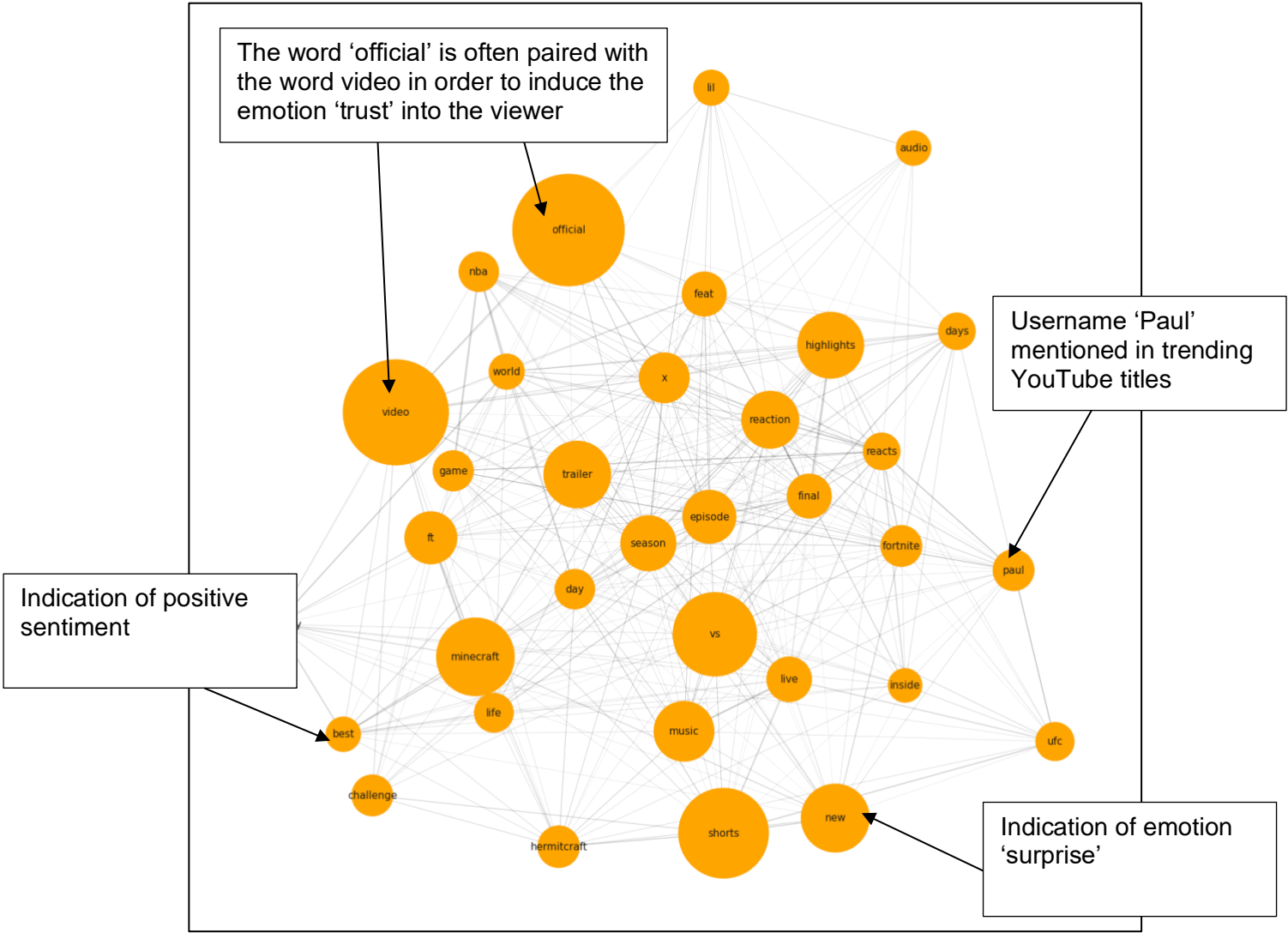


Figure 10 – Co-Occurrence Network of Frequently used Words in YouTube titles

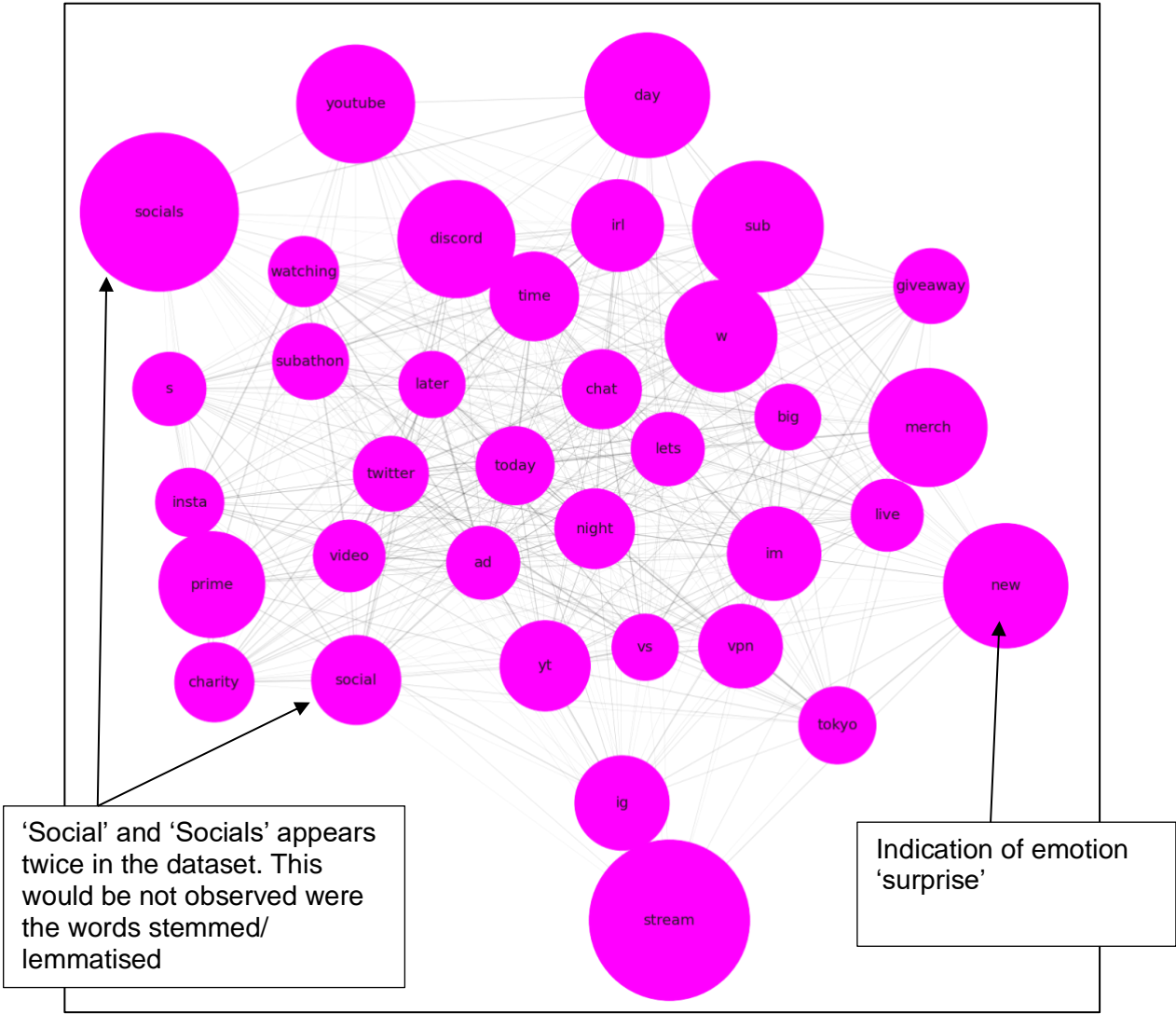


Figure 11 - Co-Occurrence Network of Frequently Used Words in Twitch ‘Just Chatting’ Titles

4.6 Exploratory Data Analysis Summary

The Twitch raw data for this analysis had 139,500 observations. The YouTube raw data for this analysis had 11,783 observations. The data was from the period 28th of May 2021 to the 28th of July 2021. This data was filtered to retain the observations with the highest viewership.

This chapter found that most usernames consist of existing English words. The words in usernames could potentially have sentiment polarity and hence be an indication of nominal realism. This chapter also found strong Spearman correlation between like count and view count on YouTube. Early exploration indicates that ‘surprise’ and ‘trust’ are dominant emotions in YouTube and Twitch titles. Valence sentiment polarity and emotion will be discussed in chapter 5 of this research paper.

Please find code used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/code/Sentiment_Analysis_of_Usernames_and_Titles_on_YouTube_and_Twitch_Exploratory_Data_Analysis.ipynb

Please find data used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/data/twitch_data.csv.zip

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/data/youtube_data.csv

CHAPTER 5: SENTIMENT ANALYSIS OF USERNAMES AND TITLES

This research project aims to investigate the impact of username and title sentiment on viewership performance on YouTube and Twitch. This chapter investigates the first component of the research question. It provides a survey of sentiment used in usernames and titles on YouTube and Twitch. Furthermore, it gives a detailed understanding of the sentiment in titles through emotion detection.

5.1 Sentiment Analysis and Emotion Detection: Methodology

The YouTube and Twitch datasets were created through the ingestion of trending videos and trending streams. Ingestion was performed from the 28th of May till the 28th of July. This meant that a longitudinal study was performed of the observations of each video or steam on the trending tab for that period. Multiple observations of the same video or stream would appear in the dataset, each with different viewership figures.

The YouTube dataset was filtered from 11,783 observations to 1,910 unique observations of videos. The inclusion criterion into the dataset were observations with the same publish time, same channel username and same titles. The same methodology was employed to reduce the size of the Twitch dataset from 139,500 observations to 23,578 unique video stream observations. The entry date and exit date into the data set was computed for all streams and videos.

Username	NEUTRAL DROP
Pre-Processed Username	['ut', 'ra', 'ne', 'neutral', 'drop', 'ro', 'al', 'tra', 'eu']
Video/ Stream Title	Camerman's Friends Love This Incredible Gift From A Jeep #Shorts
Title Sentiment	valence polarity score = 0.8807
Username Sentiment	valence polarity score = -1.0
Title Emotion	{'fear': 0.0, 'anger': 0.0, 'trust': 0.0, 'surprise': 0.25, 'sadness': 0.0, 'disgust': 0.0, 'joy': 0.5, 'anticipation': 0.25}

Table 5 – Example of methodology

5.1.1 Sentiment Analysis of YouTube and Twitch Usernames

The usernames had been pre-processed in section 3.4 in order to extract nominal realism. The format of the pre-processed usernames was a list of tokenised English words found in the NLTK lexicon. Sentiment analysis on YouTube and Twitch usernames was performed using a lexicon-based approach resulting in a valence polarity score. This valence polarity score is a measure of positive, negative or neutral sentiment.

Sentiment analysis was performed by applying a model on the pre-processed usernames. AFINN was chosen as a model to extract nominal realism. AFINN is the model that Kelly used to extract sentiment and nominal realism from business names (Kelly, 2017). AFINN is a lexicon-based model that maps words to valence polarity scores. Hence why it was chosen over Vader which was trained on social media sentences. A valence polarity score was calculated for each word extracted out of each username.

Kelly analysed one-word English surnames (e.g., surname 'Blind'). They hence equated the valence polarity score to the score given by the model. Usernames on social media often consist of multiple words (e.g., username 'LoserFruit'). The assumption is made that most of nominal realism is derived from the word with the highest absolute value valence polarity score (i.e., 'Loser' in 'LoserFruit'). The maximum absolute valence polarity score of all words extracted from one username is set to be the valence polarity score of that username.

5.1.2 Sentiment Analysis of YouTube and Twitch Titles

Vader sentiment analysis was applied to YouTube titles. Vader was chosen as model because of its state-of-the-art performance on data from social media. This is discussed in the literature review of this research project (Chapter 2, Table 1). Additionally, Vader was trained on traditional news media and social media words (Hutto and Gilbert, 2014). Hence, it is expected that Vader will have a good performance on YouTube title data from traditional news media videos. Vader returns a compound valence polarity score for each YouTube title.

Vader was applied to Twitch titles in combination with the emote-controlled lexicon-based model (Chapter 2, Table 1). The emote-controlled model was developed by Kobs et al to compensate for twitch-speak. Kobs et al employ a sequential approach to the calculation of valence polarity scores. The emote-controlled model is applied to the data ahead of Vader and a valence polarity score is produced. The score is given by Vader in the absence of a score given by the emote-controlled model. Kobs et al give precedence to the emote-controlled model due to its specialisation on Twitch data (Kobs et al, 2020).

5.1.3 Detection of Emotion in YouTube and Twitch Titles

Italian researchers used the NRC word-emotion association lexicon (NRCLex) on YouTube comments (Chapter 2, Table 1). They were aiming to understand the sentiment surrounding the Italian vaccination efforts in 2017 and 2018 (Porreca, Scozzari and Di Nicola, 2020). Earlier Gebeyaw used NRCLex to extend sentiment analysis work done on Warren Buffett's annual shareholder letters (Gebeyaw, 2017). NRCLex was enriched through the usage of synonyms from WordNet in 2019 (Bailey, 2019). NRCLex is also used in this research to further the understanding of sentiment of social media titles.

NRCLex requires a tokenised sentence as input. NRCLex returns a key-value pair of Plutchik eight primary emotion as keys and emotion valence polarity score as value. The eight primary emotions defined by Plutchik are: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger and Anticipation. The detected emotions are then programmatically visualised using Plutchik's Flower of Emotion, also known as Plutchik's Wheel of Emotion. Each flower petal depicts how much an emotion is detected. Proximity of emotions is considered. Opposing emotions are depicted in opposite directions (Semeraro, Vilella and Ruffo, 2021).

5.2 Sentiment Analysis and Emotion Detection: Findings

5.2.1 Sentiment Analysis on Data from YouTube: Findings

Out of a total of 1,454 unique videos, 60% (867) of YouTube title observations were classified as neutral by Vader. 22% (335) of title observations were classified as positive. 17% (252) of YouTube title observations were classified as having a negative valence polarity score. Data from YouTube titles had a mean and median title valence polarity score of mean = 0.03 and median = 0.

75% (1,090) of YouTube username observations were classified as neutral by Afinn. 16% (228) of username observations were classified as positive. 9% (136) of YouTube username observations were classified as negative. Data from YouTube usernames had a mean and median name valence polarity score of mean = 0.20 and median = 0.

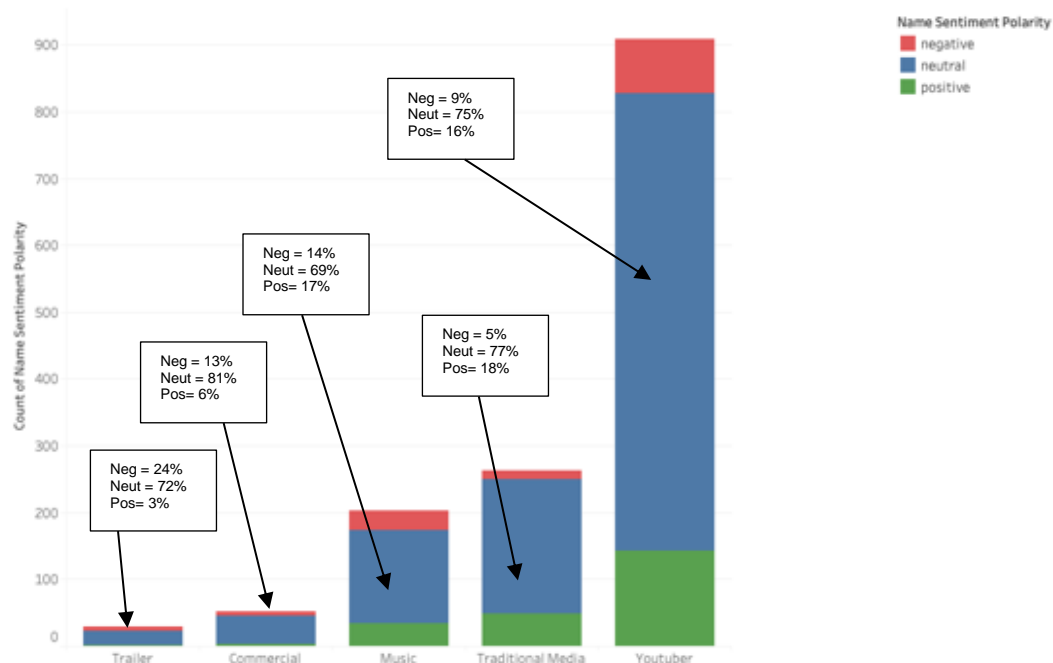


Figure 12 - Username Sentiment Polarity per Genre on YouTube

It is observed that both positive usernames and positive titles are more prevalent than negative usernames and negative titles on YouTube. Kelly states that positive usernames being more prevalent than negative usernames is an indication of nominal realism (Kelly, 2021). This is observed across most YouTube genres.

NRCLEX detected emotions in 64% (931) of YouTube titles out of a total of 1,454 unique titles. 'Trust' was the most frequently detected emotion in YouTube titles (mean score = 0.22). This was followed by 'anticipation' (mean score = 0.08) and 'fear' (mean score = 0.07).

Depicted below is Plutchik's Flower of Emotion with mean emotion values on the left. The depiction on the right has been filtered for data that had emotions to facilitate comparison. Each emotion has been scaled by its unique scale factor. This scale factor is defined as the total number of observations divided by the number of times a unique emotion is detected. It is a representation of the mean emotion valence polarity score, given the presence of that emotion.

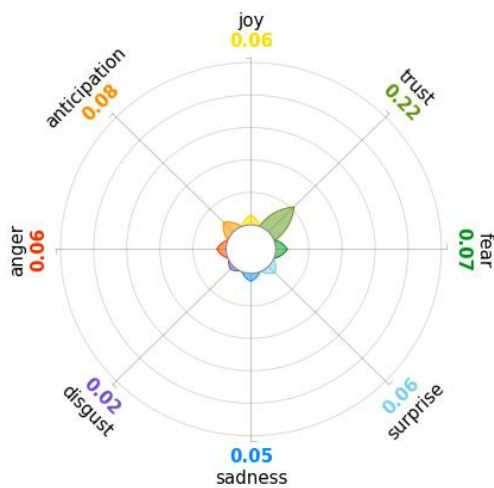


Figure 13 – Plutchik's Flower of Emotion depicting mean emotion valence polarity score detected in YouTube titles.

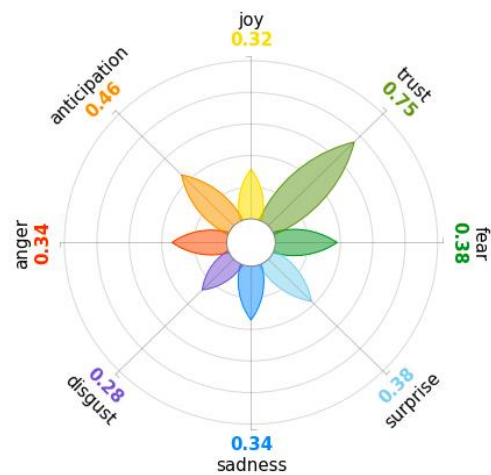


Figure 14 – Plutchik's Flower of Emotion depicting scaled mean emotion valence polarity score detected in YouTube titles.

It is observed in figure 13 that opposing emotions 'anticipation' and 'surprise' have similar emotion valence polarity score (respectively 0.08 and 0.06). Viewers will get bored of content that is too familiar. Simultaneously, a channel might alienate their recurrent viewers if their videos become too different. YouTube monitors the trending tab closely. 'Trust' being the most dominant emotion on the trending tab is an indication of their efforts in combating misleading information. CNBC reports that YouTube allocated \$300 Million to tackle misleading information in 2018 (Castillo, 2018).

A scaled Plutchik's Flower of emotion is produced for YouTube videos labelled as 'music'. The high detection of the emotion 'trust' in the music genre is because musicians use the word 'official' in their titles at high rates. This was previously observed in the co-occurrence network of most frequent words in YouTube titles (chapter 4, figure 10). An example of this is the title 'Anne-Marie & Niall Horan - Our Song [Official Video]'. Multiple versions of the music will be uploaded to YouTube e.g., a live version, an acoustic version etc. Using the word 'official' distinguishes a version of that music video. Consequently, the 'official' version of the music is likely to trend.

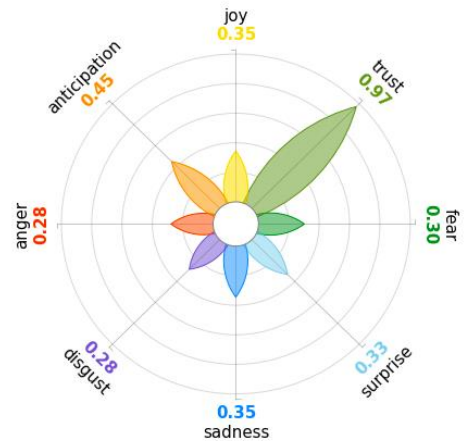


Figure 15 - Plutchik's Flower of Emotion depicting scaled mean emotion valence polarity score detected in YouTube titles for the genre 'music'.

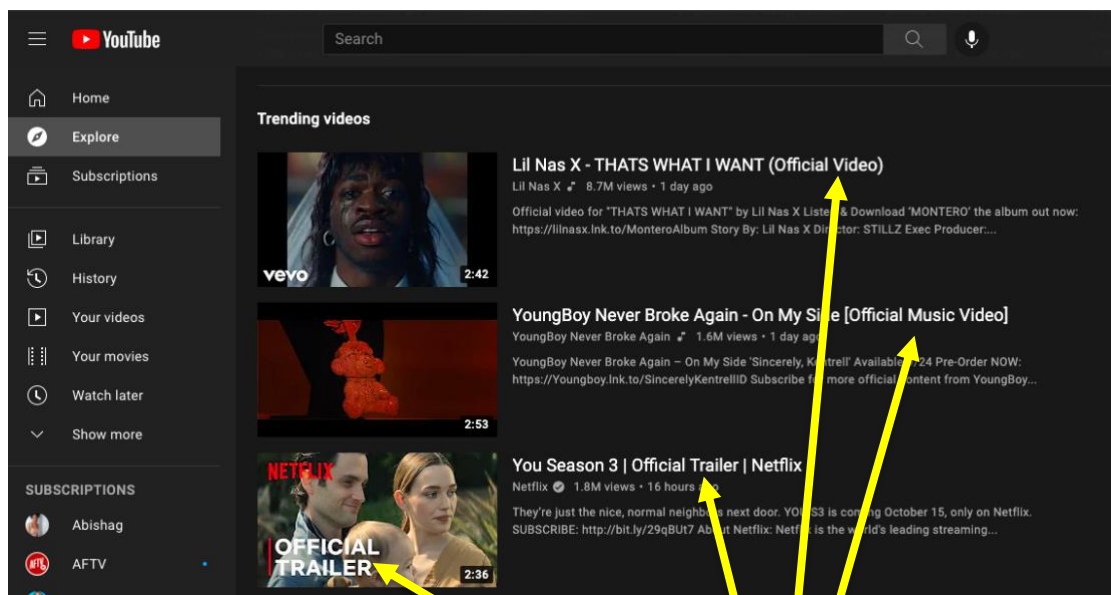


Figure 16 – The YouTube Trending Tab

The number 1, 2 and 3 trending videos on the YouTube trending tab contain the word 'official' that conveys the emotion 'trust'. Netflix goes further and includes the word 'official' in their thumbnails.

5.2.2 Sentiment Analysis on Data from Twitch: Findings

Out of a total of 23,578 unique titles, 56% (13,179) of Twitch title observations were classified as neutral by Emote-Controlled and Vader. 29% (6,872) of title observations were classified as positive. 15% (3,527) of Twitch title observations were classified as negative. Data from Twitch titles had a mean and median title valence polarity score of mean = 0.09 and median = 0.

Afinn classified 78% (18,347) of Twitch username observations as neutral out of a total 23,578 usernames. 16% (2,911) of username observations were classified as negative. 12% (2,240) of Twitch username observations were classified as positive. Data from Twitch usernames had a mean and median name valence polarity score of mean = -0.06 and median = 0.

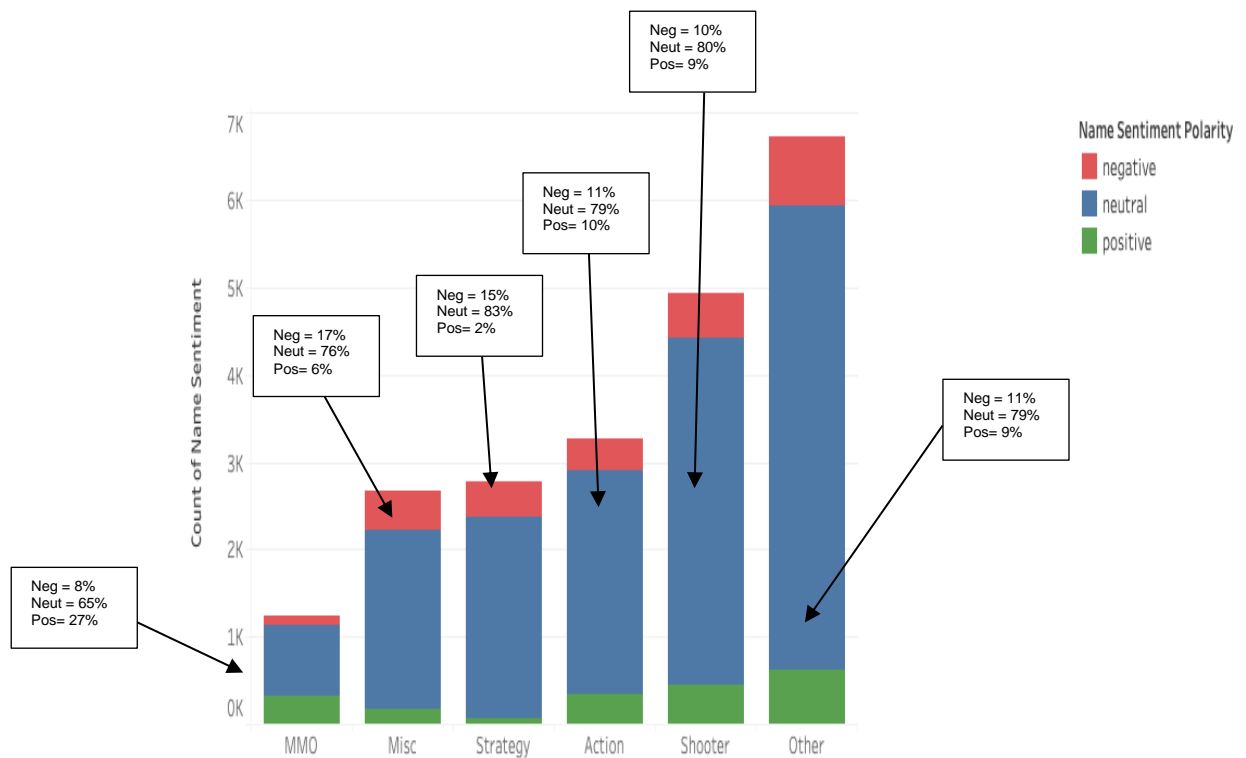


Figure 17 – Username Sentiment per Genre on Twitch

Twitch streamers prefer words with positive valence polarity scores in their titles. It is observed that negative usernames are more prevalent than positive usernames on Twitch. This trend is observed throughout the most popular genres on Twitch. This is not an argument against the presence of nominal realism. The Twitch streamer usernames ‘LoserFruit’ and ‘TrainwrecksTV’ are both classified as negative. Yet, the argument can be made that these Twitch usernames were chosen with nominal realism in mind.

NRCLex detected emotions in 58% (13,752) of Twitch titles out of total of 23,578 unique titles. 'Anticipation' was the most frequently detected emotion in Twitch titles (mean score = 0.12). This was followed by 'trust' (mean score = 0.08). Depicted below is Plutchik's Flower of emotion with mean emotion values left. The depiction on the right has been filtered for data that had emotions as explained in section 5.2.1.

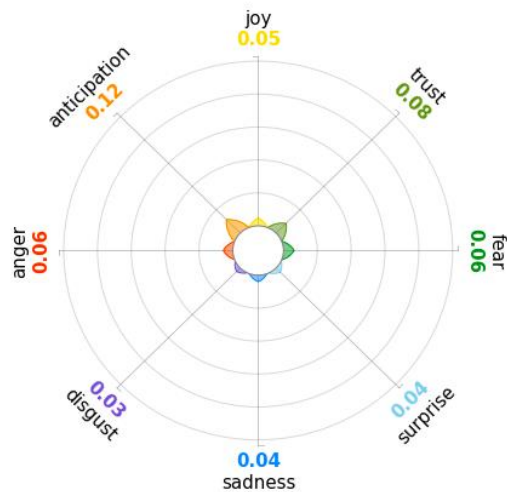


Figure 18 - Plutchik's Flower of Emotion depicting mean emotion valence polarity score detected in Twitch titles.

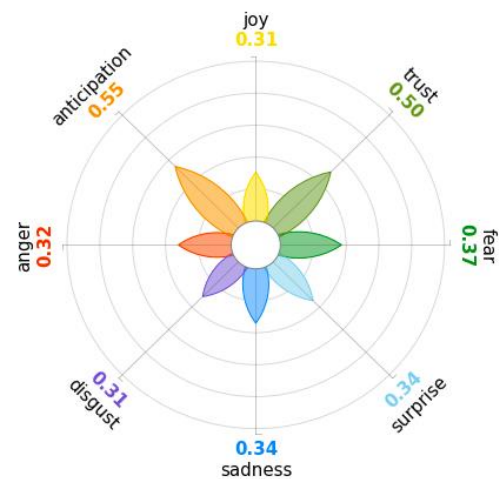


Figure 19 - Plutchik's Flower of Emotion depicting scaled mean emotion valence polarity score detected in Twitch titles.

'Anticipation' being the dominant emotion reflects Twitch subscription-for-view culture that encourages repeat viewership. Viewers can support Twitch creators through their paid Amazon Prime Subscription for an advertisement free channel experience (Stat, 2016). An Amazon Prime subscription is valid for either a month or a year.

5.3 Sentiment Analysis and Emotion Detection: Limitations

Valence polarity on Twitch titles was computed by combining the emote-controlled model with Vader. Precedence was given to the domain specific emote-controlled model. Emote-controlled employs a valence polarity scale from -1 to +1. Vader employs a scale from -4 to +4 (Kobs et al, 2020). Non-standardised valence polarity scores are used from both models. The literature review failed to find a clarification for the difference in scale in valence weighting. The implicit assumption of the difference in scale in valence is that the text comments have more valence than emotes. An alternative method would have been to transform between linear scales. The assumption is then made that emotes and text comments carry equal valence.

Analysis revealed that only one percent of Twitch titles contained emotes. This concurs with the findings of lack of emotes in the most frequent words in Twitch titles as observed in section 4.3. Hence, the effect of the usage of models with different scales in valence weighting is limited.

Lexicon based approaches are limited in their capability of finding sentiment. Words that are not part of the lexicon will be marked as neutral. E.g. Bella Poarch is marked as neutral, although 'belle' is easily detectable from 'bella'. An argument for lexicon-based approaches is that they are easily extensible.

5.4 Sentiment Analysis and Emotion Detection: Summary

A detailed explanation of the methodology used to extract sentiment and emotion from usernames and titles is provided. The results of the sentiment analysis and emotion detection are critically analysed.

This chapter found that positive usernames and positive titles are more prevalent than negative usernames and negative titles on YouTube. The observation is made that the emotion 'trust' was the most dominant emotion on the YouTube trending page.

It is appreciated that negative usernames are more prevalent than positive usernames on Twitch. This chapter also found that Twitch streamers prefer words with positive valence polarity scores in their titles. 'Anticipation' was the most dominant emotion on Twitch, followed by the emotion 'trust'.

Please find code used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/code/Sentiment_Analysis_of_Usernames_and_Titles_on_YouTube_and_Twitch_Twitch_Analysis.ipynb

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/code/Sentiment_Analysis_of_Usernames_and_Titles_on_YouTube_and_Twitch_YouTube_Analysis.ipynb

Please find data used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/data/twitch_data.csv.zip

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/data/youtube_data.csv

CHAPTER 6: ANALYSIS OF SENTIMENT IMPACT ON VIEWERSHIP PERFORMANCE

This chapter investigates the second component of the research question. It investigates the impact on viewership, given a sentiment or emotion. Hence, it introduces the performance metric ‘percentage increase in viewership over lifetime on the trending tab’. A comparison is performed between the different sentiment and emotions and their influence on the performance metric.

6.1 Analysis of Sentiment Impact on Viewership Performance: Methodology

Kelly analysed the impact of sentiment of business names on their revenue (Kelly, 2017). This section performs similar analysis by investigating the impact of sentiment on viewership performance. ‘Time on the trending tab’ is a metric introduced by Nitire. It focussed on the lifetime of a video on the trending tab in terms of days (Nitire, 2021). Analogously, this research introduces the performance metric ‘percentage viewership increase over lifetime on the trending tab’. The different trending criteria discussed in section 3.1 result in videos with vastly different absolute viewership numbers and lifetime on the trending tab. Using percentages instead of absolute numbers allows for objective performance comparison.

Special attention was paid to the quality of the data through the data ingestion and data cleaning process. However, producing the performance metric revealed data quality issues. Data quality issues were observed with the validity of the metric. Validity is defined as meeting defined business rules or constraints (Shepperd, 2020). Theoretically, computing the performance metric yields a positive figure. This is set to be the referential integrity constraint of the performance metric. This is because the performance metric is a cumulative count of the views over the lifetime on the trending tab. Negative and zero values of the performance metric were found in both data originating from YouTube and Twitch.

Some observations on the YouTube trending tab decreased in viewership over time. This is a result of YouTube deleting artificial views. Artificial views are views obtained through programmed scripts or real people in viewing farms. Validity issues with data originating from Twitch were observed through the appearance of data with zero as viewership number. The Twitch API occasionally reversed the parameter responsible for ingesting the streams with the highest viewership figures. Instead, it ingested the streams with the lowest viewership figures. The validity concerns with both datasets were resolved by deletion of the observations that failed to meet the referential integrity constraint.

Boxplots were created with categorical sentiment polarity and emotion variables on the x-axis. The numerical variable 'percentage increase in viewership' was depicted on the y-axis. A logarithmic scale was used on the y-axis. This was to facilitate the depiction of extreme viewership values from viral videos on the trending tab. Additionally, the median statistic was chosen for comparison of effects. This is due to its insensitivity to extreme values. Bootstrapped 95% confidence intervals were produced for point estimates of median statistics to further enhance effect comparison.

A non-parametric ranked analysis of the variance of the performance metric was performed. The Kruskal-Wallis test provides an indication of whether the ranked medians are different (McDonald, 2009). The decision was made to perform the analysis with Kruskal-Wallis ANOVA because the shape of the non-parametric right-skewed distribution of the analysed samples were the same.

A table was produced containing the values of the medians, confidence intervals and F-ratio of the ranked medians to facilitate comparison of small differences. Finally, a pairwise comparison of the ranked effects was performed using Tukey Honest Significance Difference (HSD). Tukey HSD comparison was performed where sufficient evidence was available to suggest the null-hypothesis of the Kruskal-Wallis test as incorrect. The Tukey HSD test was performed with a confidence coefficient of 95%.

6.2 Analysis of Sentiment Impact on Viewership Performance: Findings

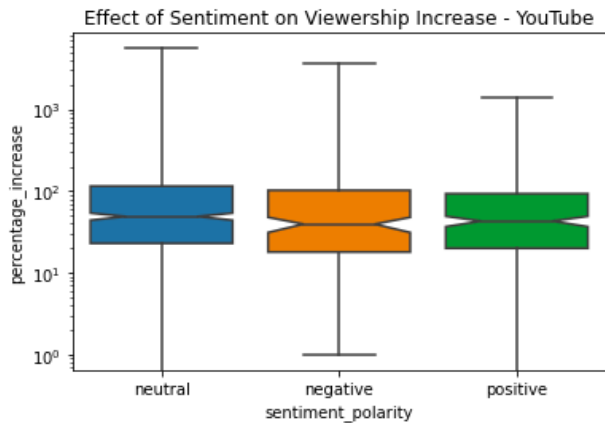


Figure 20 – Effect of Title Sentiment on Viewership Increase on YouTube. Sentiment Polarity produced by Vader. Details found in table 6.

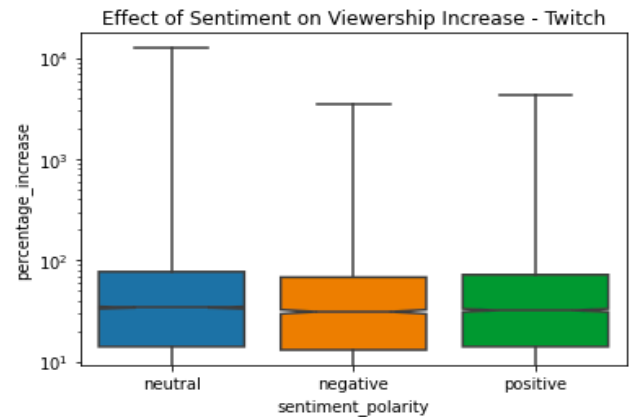


Figure 21 – Effect of Title Sentiment on Viewership Increase on Twitch. Sentiment Polarity produced by Emote-Controlled and Vader. Details found in table 7.

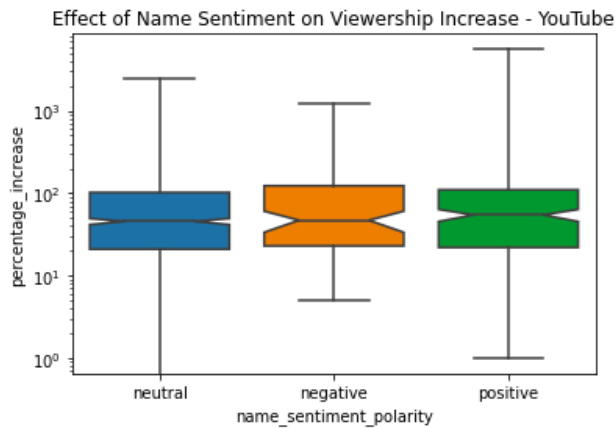


Figure 22 – Effect of Name Sentiment on Viewership Increase on YouTube. Sentiment Polarity produced by AFINN. Details found in table 6.

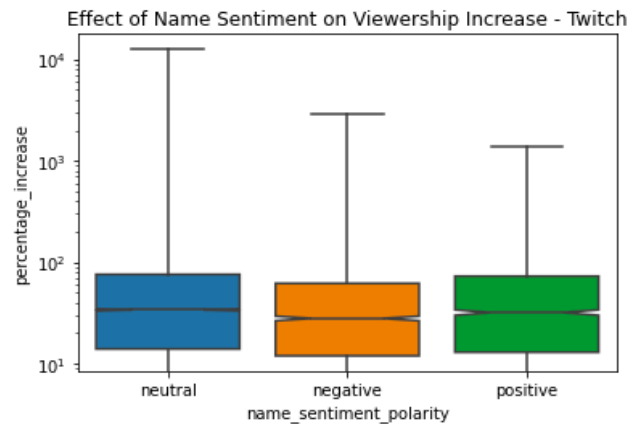


Figure 23 – Effect of Name Sentiment on Viewership Increase on Twitch. Sentiment Polarity produced by AFINN. Details found in table 7.

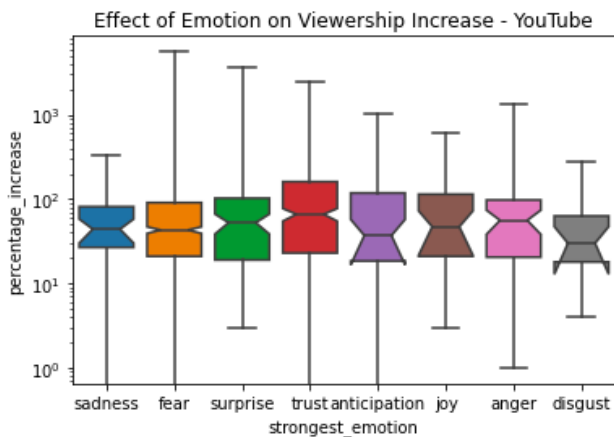


Figure 24 – Effect of Title Emotion on Viewership Increase on YouTube. Emotion valence produced by NRCLEX. Details found in table 6.

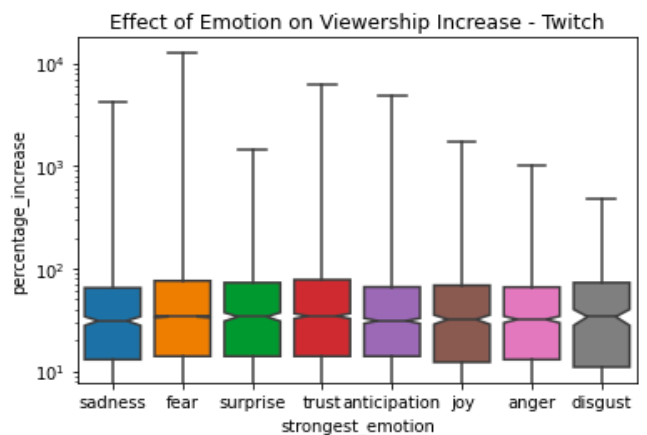


Figure 25 – Effect of Title Emotion on Viewership Increase on Twitch. Emotion valence produced by NRCLEX. Details found in table 7.

SENTIMENT ANALYSIS OF SOCIAL MEDIA USERNAMES AND TITLES ON YouTube AND Twitch

YouTube	median viewership % increase	bootstrapped confidence interval (95%)	Kruskal-Wallis Ranked ANOVA	Depiction figure number
negative title	39%	(32, 54)	F ratio = 6.76 p-value = 0.03	Figure 20
neutral title	49%	(44, 57)		
positive title	43%	(38, 53)		
negative name	47%	(35, 79)	F ratio = 2.28 p-value = 0.32	Figure 22
neutral name	46%	(42, 50)		
positive name	55%	(44, 66)		
sadness in title	44%	(33, 66)	F ratio = 22.44 p-value = 9.52×10^{-4}	Figure 24
fear in title	43%	(40, 48)		
surprise in title	53%	(36, 66)		
trust in title	66%	(53, 80)		
anticipation in title	37%	(29, 57)		
joy in title	47%	(33, 92)		
anger in title	56%	(32, 81)		
disgust in title	30%	(18, 63)		

Table 6 – Results of Analysis of Sentiment Impact on Viewership Performance on YouTube

Twitch	median viewership % increase	bootstrapped confidence interval (95%)	Kruskal-Wallis Ranked ANOVA	Depiction figure number
negative title	31%	(29, 32)	F ratio = 21.02 p-value = 2.72×10^{-5}	figure 21
neutral title	34%	(33, 35)		
positive title	32%	(32, 34)		
negative name	28%	(27, 30)	F ratio = 48.56 p-value = 2.85×10^{-11}	figure 23
neutral name	34%	(33, 35)		
positive name	32%	(30, 35)		
sadness in title	31%	(29, 34)	F ratio = 20.16 p-value = 5.24×10^{-3}	figure 25
fear in title	34%	(33, 34)		
surprise in title	34%	(30, 37)		
trust in title	35%	(32, 37)		
anticipation in title	31%	(29, 34)		
joy in title	32%	(28, 34)		
anger in title	32%	(29, 34)		
disgust in title	34%	(28, 45)		

Table 7 - Results of Analysis of Sentiment Impact on Viewership Performance on Twitch

1,454 unique YouTube videos and 23,578 unique Twitch streams were analysed. In a consultancy setting where the goal is to increase viewership, the actionable advice would be to consistently optimise for the emotion 'trust' in titles. It is appreciated that the emotion 'trust' had the highest median increase in viewership over the lifetime of the observations on the trending tabs (figure 24 and figure 25). The emotion 'trust' resulted in a median percentage increase in viewership of 66% on YouTube and 35% on Twitch.

On both platforms, out of all effects, 'trust' resulted in the highest median increase in viewership. This finding is consistent with the earlier observations in the exploratory data analysis section and sentiment analysis section of this report (chapter 4, figure 10; chapter 5, figure 14 and 19). Comparison of the ranked effects including the emotion 'trust' revealed a Kruskal-Wallis ranked ANOVA p-value < 0.05 for both YouTube and Twitch.

Interesting observations surrounding the emotions 'fear' and 'disgust' are made by looking at the variances. The emotion 'fear' had the most extreme outliers in terms of viral viewership performance for both YouTube and Twitch (figure 24 and figure 25). Comparatively, videos evoking the 'disgust' emotion in their title were less likely to obtain extreme viral viewership on both platforms (figure 24 and figure 25). In a risk-tolerant consultancy setting where the goal is to increase viewership, the actionable advice would be to occasionally optimise for the emotion 'fear' in titles. The emotion 'disgust' is to be avoided in titles.

The variances of the effects show significant differences. However, the median values of the effects are in proximity of each other on a logarithmic scale. This is consistent for all studies on both YouTube (table 6) and Twitch (table 7). Furthermore, additional uncertainty is revealed through further investigation into the confidence intervals of the median point estimates. The point estimates for the median viewership statistics on YouTube are located within wide confidence intervals (table 6: figure 20, table 6: figure 22, table 6: figure 24). There is more confidence in stating the median values of the effects, for all Twitch studies (table 7: figure 21, table 7: figure 23, table 7: figure 25).

Finally, it is observed that the ranked medians of the different effects have p-values < 0.05 . The only study for which there was not enough evidence to conclude that the effect exist, is the study on the impact of username sentiment on viewership on YouTube (p-value = 0.32). Given the proximity of the point estimates of the median values and failing to provide sufficient evidence through the Kruskal-Wallis test that the null hypothesis is false; there is not enough evidence to suggest that username sentiment impacts viewership performance on YouTube. Although sufficient evidence was provided to suggest that usernames impact viewership on Twitch, usernames with a neutral sentiment polarity outperformed positive and negative usernames.

Evidence of the impact of nominal realism on viewership performance is inconclusive. This is observed through median viewership increase of 47%, 46% and 55% for respectively negative, neutral and positive YouTube usernames (table 6: figure 22). Those figures are 28%, 34% and 32% percent for respectively negative, neutral and positive Twitch usernames (table 7: figure 23).

Pairwise comparison of the effects confirmed the observations made through the confidence intervals. The p-value of Tukey HSD was < 0.05 where the confidence intervals of the median point estimates of the effects did not significantly overlap. For all the other pairwise combinations, there was not enough evidence to suggest that the ranked medians were not equal (Tukey HSD confidence coefficient = 95%).

Tukey HSD pairwise combination	p-value	bootstrapped confidence intervals
Twitch negative-neutral titles	0.001	(29, 32); (33, 35)
Twitch neutral-positive titles	0.0318	(33, 35); (32, 34)
Twitch negative-neutral usernames	0.001	(27, 30); (33, 35)
Twitch negative-positive usernames	0.001	(27, 30); (30, 35)
YouTube fear-trust titles	0.001	(40, 48); (53, 80)

Table 8 – Tukey Honest Significance Difference Pairwise Combination

The inconclusive results are in contrast with research performed by Kelly. Kelly demonstrated that businesses with positive sentiment outperformed businesses with a negative sentiment in terms of revenue (Kelly, 2017). The findings in this analysis are consistent with the rise of clickbait titles (Alexander, 2018). These clickbait titles often evoke strong emotions.

6.3 Analysis of Sentiment Impact on Viewership Performance: Limitations

Kelly analyses name sentiment influence on performance of businesses by comparing the absolute values of the revenues of those businesses (Kelly, 2017). Comparison of the absolute values of the performance metric in this research will yield a different conclusion for data originating from YouTube. YouTube videos that originated from channels with a positive username had a higher percentage of viewership increase than videos with a negative username (figure 22).

Kobs et al classified all sentences with a valence polarity score between -0.3 and +0.3 as neutral (Kobs et al, 2020). This is different from this research which only classified titles with a valence polarity score of 0 as neutral.

Observations that exited the trending tab on the first date of data collection have an inaccurate performance metric. Similarly, observations that entered the trending tab on the last date of data collection also have an inaccurate performance metric. This is because the cumulative viewership values of the videos are truncated, resulting in a smaller performance metric.

Although this research is confident in its findings and actionable advice through the volume and vastness of data analysed, the confidence intervals of the effects and Tukey HSD point to their limitations. Additionally, the findings are based on analysis of videos on social media trending tabs. This quantitative non-probabilistic sampling approach of the best performing videos results in survivorship bias. The findings are applicable to actors that seek to perform well on trending tabs. They might not generalise well to people just commencing their video or stream journey on YouTube and Twitch.

6.4 Analysis of Sentiment Impact on Viewership Performance: Summary

There is not sufficient evidence to suggest that username sentiment has an influence on the performance of a video in terms of viewership. There is enough evidence to suggest that title sentiment and title emotions affect video viewership performance.

This chapter finds that the emotion ‘trust’ had the highest median increase in viewership over the lifetime of the observations on the trending tabs. It also finds that the emotion ‘fear’ had the most extreme outliers in terms of viral viewership. Videos evoking the emotion ‘disgust’ were less likely to obtain extreme viral viewership. The Kruskal-Wallis 1-way ranked ANOVA test found that there were differences in the median values of the title sentiment and emotion effects.

Most pairwise comparison of the ranked effects through Tukey HSD showed that there was not enough evidence to suggest that the ranked medians of the effects were not equal. This concurred with the observations made through the overlapping confidence intervals of the median point estimates of the effects. Although this research would give the actionable advice to optimise for the emotion ‘trust’ in a consultancy setting, the confidence intervals and Tukey HSD point to caution.

Please find code used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/code/Sentiment_Analysis_of_Usernames_and_Titles_on_YouTube_and_Twitch_Twitch_Analysis.ipynb

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/code/Sentiment_Analysis_of_Usernames_and_Titles_on_YouTube_and_Twitch_YouTube_Analysis.ipynb

Please find data used for this chapter at:

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/data/twitch_views.csv

https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/blob/main/data/youtube_views.csv

CHAPTER 7: PROTOTYPE

The web app was deployed on Firebase but has been taken down due to cloud computing bills. The app takes in a username or title and returns its sentiment classification.

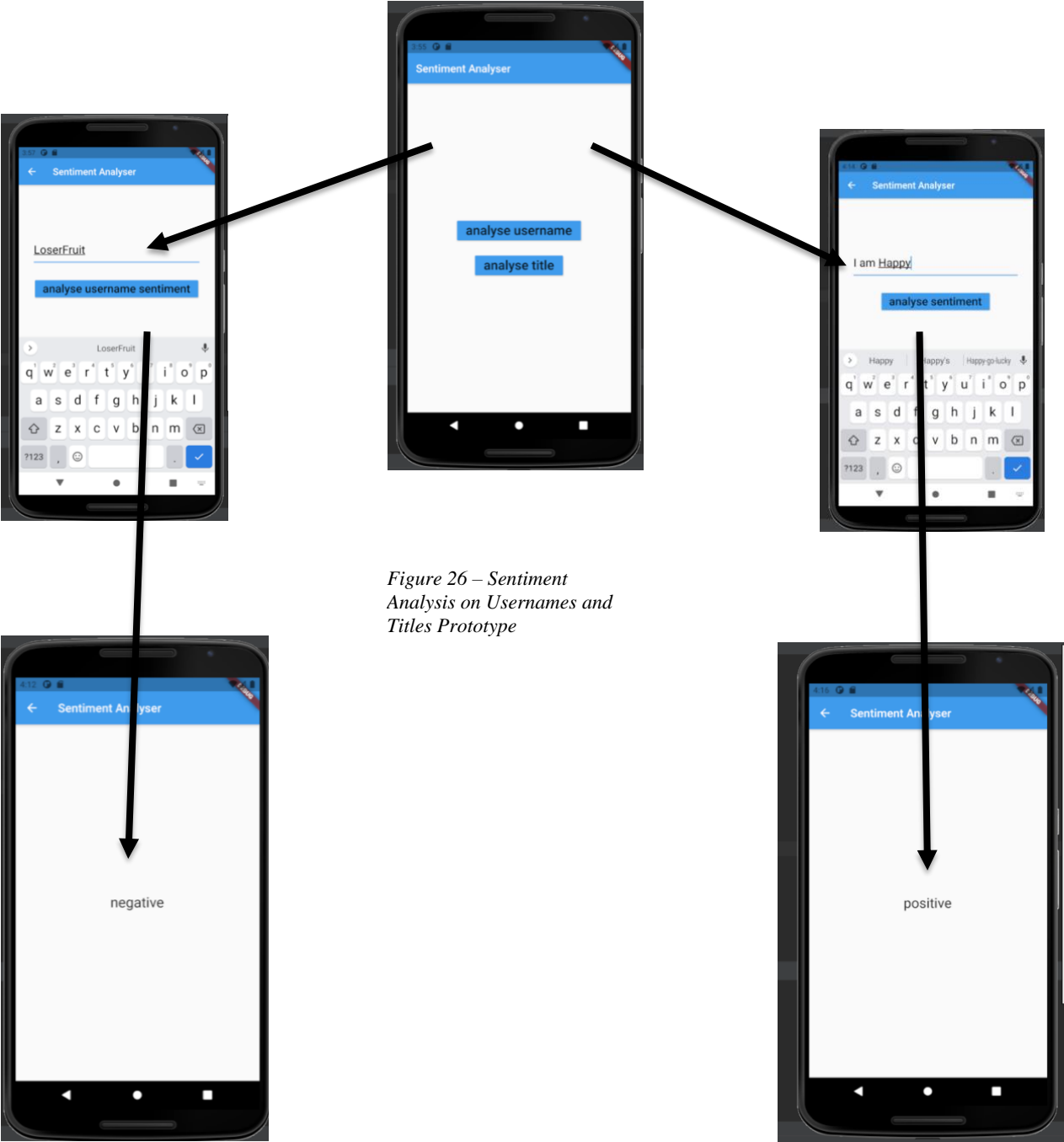


Figure 26 – Sentiment Analysis on Usernames and Titles Prototype

The aim behind the prototype was to expose the sentiment analysis methods from chapter 5 and chapter 6 to a user interface. The initial plan was to build a Python app using the Python web user interface library Flask. Using Python to build the user interface would result in programming language consistency throughout the whole project.

This Flask app would be deployed on the Google Cloud Platform using the Google App Engine. Unfortunately, (py)linting source code errors, due to the underlying technology of the App Engine, impeded importing third party libraries. These third-party libraries included the sentiment analysis models. These environment errors were replicated on different Google Cloud accounts in different countries. These errors shortly resolved themselves, libraries were able to be imported, but the errors returned. Hence, the decision was made to explore other technology to create the user interface to expose the sentiment analysis methods.

The cross-platform Flutter SDK was chosen as the alternative user interface technology to expose the underlying sentiment analysis methods. Flutter was precisely chosen because of its code reusability for different platforms. Flutter allows one codebase to be deployed to different clients such as MacOS, Linux, Windows, android, iOS, chrome extensions, web, wearables etc. Flutter apps are written using the Dart programming language. The back-end was built by deploying the Python sentiment analysis models as Google Cloud Functions APIs. This allowed the reliable Python code used in the analysis sections to be reused.

The first Flutter iteration of the user interface was developed as an Android mobile app. This was to permit rapid testing and iteration of the user interface in the local android emulator. The cross-platform Flutter app was then compiled as a web app and deployed on Google Firebase. Testing the web app revealed Cross Origin Resource Sharing (CORS) errors. The user interface would take the input and send it to the back-end. The back-end would no longer send a response to the front-end user interface.

This was specific to the web app, but the same code would run perfectly as a mobile app. This was resolved by including HTTP headers. These HTTP headers add information to the user input in the front-end. These headers also add information to the server reply in the back-end. They explicitly permit the front-end to receive information from the back-end i.e. in this project the sentiment classification of the username or title.

The usage of Google Cloud Functions to wrap the tried and tested sentiment analysis methods, the usage of the Flutter SDK developed by Google and the usage of Google Firebase to deploy the model resulted in the reliability of the application.

Please find code used for this chapter at:

<https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch/tree/main/prototype>

CHAPTER 8: TECHNOLOGY STACK AND DATA MANAGEMENT

This section discusses the technology stack used throughout this open research project. The data management strategy is also discussed.

8.1 Discussion of Technology Stack

The data analysis in this research project was performed using Python as its main programming language. The decision was made to use Python because of the researcher's experience with the language. The usage of Python would allow for reusability and consistency of code across the data ingestion, data analysis and prototype phases of this project. Google Collaboratory was used as the primary IDE for script development. Google Collaboratory provides automated version history and was chosen for its reliability. Most of the data visualisation was performed using Python for its data visualisation versatility. Tableau was used where customisability of the visualisations was not deemed relevant.

The Google Cloud Platform and Google Cloud Functions were used to automate the data ingestion process. They were also used to deploy the sentiment analysis methods that were embedded in the application prototype of the sentiment analysis tool. The Google Cloud Platform was chosen ahead of AWS and Microsoft Azure for its user interface.

Finally, the decision was made to develop the sentiment analysis software prototype using the cross-platform SDK Flutter. Flutter provides deployment to iOS, Android and Web using one codebase. Dart is the programming language used for Flutter applications. The software prototype used Python Google Cloud Functions in the back-end to analyse the sentiment in usernames and titles.

8.2 Discussion of Code and Data Management

This research project follows an open research approach to code and data management. It aims to share the data, code and findings on ResearchGate and GitHub. It also aims to publish a copy of the findings on Arxiv.org, given an invitation to the platform. Finally, the data visualisation and sentiment analysis section of the research will be rewritten to be published on Medium, LinkedIn and Kaggle. Hence, the code and data used for this project will be available in various places to ensure replicability, reproducibility and extensibility of the findings.

The Google Collaboratory IDE provides built-in code versioning. GitHub was also used to maintain copies of all the programming files produced throughout this project. Copies of the data files were stored on both Google Drive and GitHub.

Please find all the Code and Data files produced as part of this research for your appreciation at

<https://github.com/JefNtungila/Sentiment-Analysis-of-Usernames-and-Titles-on-YouTube-and-Twitch>

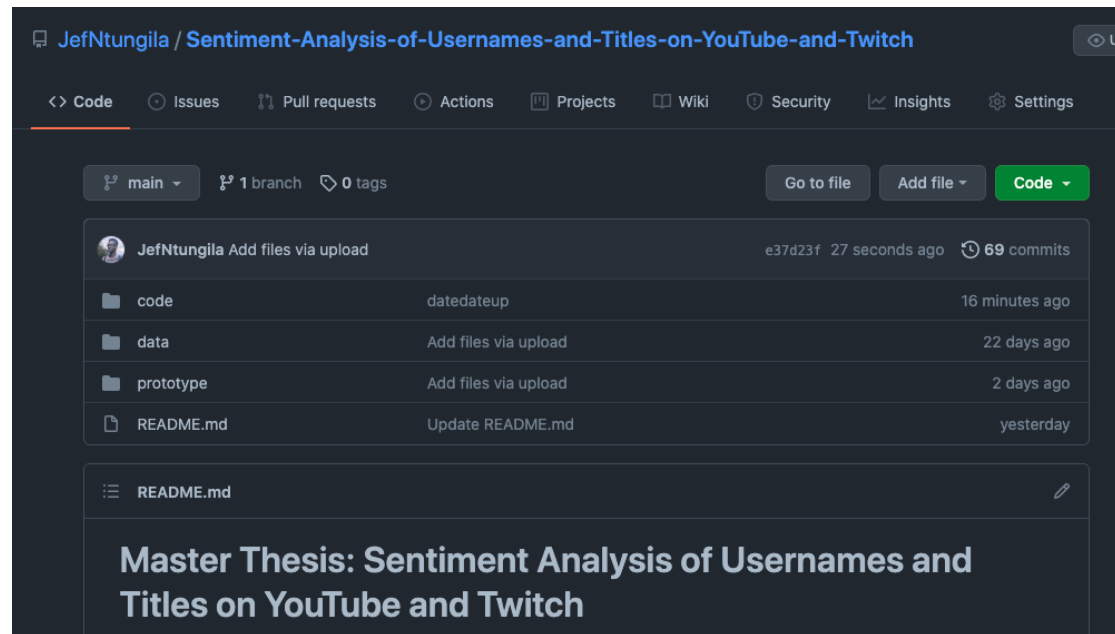


Figure 27 – Research Code and Data

CHAPTER 9: CONCLUDING DISCUSSION AND FURTHER WORK

9.1 Summary of the Dissertation

The following is a summary of work done in this report:

- Appraisal of the literature surrounding sentiment analysis and emotion detection on social media and titles was provided. This resulted in an overview of the state-of-the-art of sentiment analysis techniques.
- The data ingestion, data preparation and data cleaning process were discussed.
- NLP data visualisation techniques were used to obtain an in-depth understanding of the data.
- Sentiment analysis and emotion detection was performed on usernames and titles.
- A web app that packaged the sentiment analysis models used in this research was deployed.

The following is a summary of the findings of this report:

- It is observed that rule-based methodologies are preferred to machine learning and deep learning alternatives for sentiment analysis on social media. An explanation for this is their state-of-the-art performances e.g., Vader; F1 score=96% on tweet polarity classification (Hutto and Gilbert, 2014).
- The data granularity refinement process found that YouTube had kept their promise of allocating half of the trending tab to YouTube creators (Alexander, 2019).
- Exploratory Data Analysis showed that the numerical features such as like count and dislike count were strongly correlated with viewership. The observation is made that viewers have strong comment engagement with viral traditional media videos.
- Analysis of sentiment showed that both positive usernames and positive titles are more prevalent than negative usernames and negative titles on YouTube. It is observed that negative usernames are more prevalent than positive usernames on Twitch. This research finds that positive titles are more prevalent than negative titles on Twitch.
- Sufficient evidence was provided to suggest that title sentiment and title emotion impacts viewership on YouTube and Twitch. Although nominal realism is observed in YouTube and Twitch usernames, insufficient evidence was provided to suggest that username sentiment impacts viewership on both platforms.
- In a consultancy setting where the goal is to increase viewership, the actionable advice would be to consistently optimise for the emotion 'trust' in titles.

9.2 Future Research and Development

Shivhare and Khethawat quotes Sebea et al in stating that '[...] work has been done regarding speech and facial emotion recognition' (Shivhare and Khethawat, 2012). It is widely believed that titles and thumbnails influence video viewership performance. The YouTube Creator Academy states that 'thumbnails and titles act like billboards to help viewers decide to watch your videos' (YouTube Creator Academy, n.d.). Further research would investigate the influence of thumbnail facial emotion on video performance in terms of viewership. This is analogous to this research that investigates the influence of username and title sentiment on video performance in terms of viewership.

9.3 Personal Reflections

I truly enjoyed the researching and documenting process. A lot of personal growth was made throughout the extensive data analysis. I often found myself at the very extreme limits of my comfort zone, but not too far into discomfort. This allowed me to drastically improve in my research abilities, in my engineering abilities, in my programming and statistical abilities. I intend to deepen my knowledge in biological statistics.

The web app is currently powered by sentiment analysis models running server-side as Google Cloud Functions on the Google Cloud Platform. This is expensive. If I were to do this project again with the benefit of hindsight, I would run the artificial intelligence models client-side on the users' devices. This would remove the cloud-computing inference costs, whilst increasing the speed of inference.

REFERENCES

- Acheampong, Francisca Adoma, Chen Wenyu, and Henry Nunoo-Mensah. "Text-based emotion detection: Advances, challenges, and opportunities." *Engineering Reports* 2, no. 7 (2020): e12189.
- Alexander, J. (2019) 'YouTube's Trending section puts creators at a huge disadvantage over big brands', *The Verge*, 29 May. Available at: <https://www.theverge.com/2019/5/29/18642833/youtube-trending-coffee-break-pewdiepie-late-night-sports-highlights> (Accessed: 10 August 2021).
- Asghar, M.Z., Ahmad, S., Marwat, A. and Kundi, F.M., 2015. Sentiment analysis on youtube: A brief survey. *arXiv preprint arXiv:1511.09142*.
- Bailey, M. NRClex. 2019. Available at: <https://pypi.org/project/NRClex/> (Accessed: 13 August 2021).
- Castillo, M. (2018) 'YouTube will use six popular YouTube stars to educate kids about fake news', *CNBC*, 9 July. Available at: <https://www.cnn.com/2018/07/09/youtubes-plan-to-fight-fake-news-includes-more-support-article-links.html> (Accessed: 11 August 2021).
- Chamlerwat, W., Bhattarakosol, P., Rungkasiri, T. and Haruechaiyasak, C., 2012. Discovering Consumer Insight from Twitter via Sentiment Analysis. *J. Univers. Comput. Sci.*, 18(8), pp.973-992.
- Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Coffee Break (2019) 'What 40,000 Videos Tell Us About The Trending Tab'. 21 May. Available at: <https://www.youtube.com/watch?v=fDqBeXJ8Zx8> (Accessed: 11 August 2021).
- Coffee Break (Coffezilla) (2019) 'Researching The Trending Tab (BTS)'. 21 May. Available at: <https://www.youtube.com/watch?v=sEvtpj-uChA> (Accessed: 11 August 2021).
- Delancey, J. 2020 'Pros and Cons of NLTK Sentiment Analysis with VADER', *CODE PROJECT*, 29 May 2020. Available at: <https://www.codeproject.com/Articles/5269447/Pros-and-Cons-of-NLTK-Sentiment-Analysis-with-VADE> (Accessed: 26 August 2021).
- Funk, M. 2020 'How Does YouTube Count Views? ', *Tubics*, 04 February. Available at: <https://www.tubics.com/blog/what-counts-as-a-view-on-youtube/> (Accessed: 10 August 2021).
- Gebeyaw, M. 2017 'Parsing Text for Emotion Terms: Analysis & Visualization Using ', *R-bloggers*, 11 May. Available at: <https://www.r-bloggers.com/2017/05/parsing-text-for-emotion-terms-analysis-visualization-using-r/> (Accessed: 10 August 2021).
- Hutto, C. and Gilbert, E., 2014, May. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).
- Iseli, M. 2020 'LULW: Here's What it Means (on Twitch) ', *LINGUABLOG*, 13 October. Available at: https://www.gamasutra.com/view/feature/1583/rethinking_the_mmo.php?page=3 (Accessed: 10 August 2021).
- Jianqiang, Z., Xiaolin, G. and Xuejun, Z., 2018. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6, pp.23253-23260.
- Kelly, M., 2000. Naming on the bright side of life. *Names*, 48(1), pp.3-26.
- Kelly, M. 2017 'What's in a Name – or Rather, What's the Cost of a Negative One?', *Viewpoints from Naxion*, May 2017. Available at: <https://www.naxionthinking.com/sites/naxthink/files/images/wysiwyg/NAXION%20VIEWPOINTS%20May%202017%20What%27s%20in%20a%20Name.pdf> (Accessed: 13 April 2021).
- Kelly, M. 2021 'Sentiment Analysis of Surnames', *R-bloggers*, 22 January. Available at: <https://www.r-bloggers.com/2021/01/sentiment-analysis-of-surnames/> (Accessed: 13 April 2021).
- Kim, J., Bae, K., Park, E. and del Pobil, A.P., 2019, November. Who will Subscribe to My Streaming Channel? The Case of Twitch. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (pp. 247-251).
- Kobs, K., Potthast, M., Wiegmann, M., Zehe, A., Stein, B. and Hotho, A., 2020. Towards Predicting the Subscription Status of Twitch. *tv Users*.

- Kobs, K., Zehe, A., Bernstetter, A., Chibane, J., Pfister, J., Tritscher, J. and Hotho, A., 2020. Emote-Controlled: Obtaining Implicit Viewer Feedback Through Emote-Based Sentiment Analysis on Comments of Popular Twitch. tv Channels. *ACM Transactions on Social Computing*, 3(2), pp.1-34.
- Loureiro, D., Marreiros, G. and Neves, J., 2011, October. Sentiment analysis of news titles. In *Portuguese Conference on Artificial Intelligence* (pp. 1-14). Springer, Berlin, Heidelberg.
- McDonald, J.H., 2009. *Handbook of biological statistics* (Vol. 2, pp. 6-59). Baltimore, MD: sparky house publishing.
- Mittal, A. and Goel, A., 2012. Stock prediction using twitter sentiment analysis. Stanford University, CS229 (2011 <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>), 15.
- Mohammad, S.M. and Turney, P.D., 2013. Nrc emotion lexicon. National Research Council, Canada, 2.
- Nielsen, F.Å., 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.
- Niture, A.A., 2021. Predictive analysis of YouTube trending videos using Machine Learning (Doctoral dissertation, Dublin Business School).
- Nuzzo, R., 2014. Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), pp.150-153.
- Porreca, A., Scozzari, F. and Di Nicola, M., 2020. Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination. *BMC public health*, 20(1), pp.1-9.
- Reagan, A.J., Danforth, C.M., Tivnan, B., Williams, J.R. and Dodds, P.S., 2017. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6, pp.1-21.
- Reis, J.M.G.B., 2020. Sentiment analysis: the case of twitch chat-Mining user feedback from livestream chats.
- Rogers, I., 2002. The Google Pagerank algorithm and how it works.
- Semeraro, A., Vilella, S. and Ruffo, G., 2021. PyPlutchik: visualising and comparing emotion-annotated corpora. *arXiv preprint arXiv:2105.04295*.
- Shepperd, M. (2020) *CS5702 Modern Data Book*. Available at: https://bookdown.org/martin_shepperd/ModernDataBook/licensing.html (Accessed: 12 August 2021).
- Shivhare, S.N. and Khethawat, S., 2012. Emotion detection from text. *arXiv preprint arXiv:1205.4944*.
- Sorens, N. 2007 'Rethinking the MMO', Gamasutra, 26 March. Available at: https://www.gamasutra.com/view/feature/1583/rethinking_the_mmo.php?page=3 (Accessed: 10 August 2021).
- Stat, N. (2016) 'Twitch will be ad-free for all Amazon Prime subscribers', *The Verge*, 30 September. Available at: <https://www.theverge.com/2016/9/30/13125824/twitch-prime-amazon-ad-free-game-discounts> (Accessed: 14 August 2021).
- Twitch, n.d. 'Understanding Viewer Count vs. Viewer List', Twitch. Available at: https://help.twitch.tv/s/article/understanding-viewer-count-vs-viewer-list?language=en_US (Accessed: 10 August 2021).
- Uryupina, O., Plank, B., Severyn, A., Rotondi, A. and Moschitti, A., 2014, May. SenTube: A Corpus for Sentiment Analysis on YouTube Social Media. In *LREC* (pp. 4244-4249).
- Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations* (pp. 115-120).
- YouTube Creator Academy, n.d. 'Make effective thumbnails and titles', Available at: <https://creatoracademy.youtube.com/page/lesson/thumbnails?hl=en-GB#:~:text=Thumbnails%20and%20titles%20act%20like,a%20broad%20range%20of%20advertisers>. (Accessed: 2 September 2021).

APPENDIX A: ETHICAL APPROVAL



College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom
www.brunel.ac.uk

23 June 2021

LETTER OF CONFIRMATION

Applicant: Mr Jephie Ntunga

Project Title: Sentiment Analysis of Social Media Usernames and Titles on YouTube and Twitch

Reference: 31181-NER-Jun2021- 32882-1

Dear Mr Jephie Ntunga,

The Research Ethics Committee has considered the above application recently submitted by you.

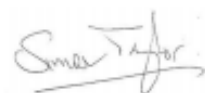
The Chair, acting under delegated authority has confirmed that, according to the information provided in your application, your project does not require ethical review.

Please note that:

- You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research, you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,



Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London