Evaluation of Comment Sentiments in Educational YouTube Videos

by

Elif Naz Özdamar

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of
Master of Science

Sabancı University

July 2023

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor, Prof. Dr. Yücel Saygın, for his mentorship throughout this journey. I feel grateful for the opportunity to work with him on my thesis.

I would like to express my appreciation to Assoc. Prof. Asta Zelenkauskaite for her valuable contributions and guidance during the final development of my thesis. Additionally, I extend my gratitude to Asst. Prof. Onur Varol for taking the time to attend my presentation and valuable feedback.

I am immensely thankful to my family, Ezgi Özdamar, Şengül Güven, and Ali Özdamar, for their emotional support and encouragement during this academic pursuit.

I am also deeply appreciative of my friends, Fatmanur Sever, İrem Anter, Gözde Kırcı, Arın Zeyneloğlu, Göksu Üreten, Emre Yeşil, Melisa Gündüz, Yiğit Can Karaköylü, Nursel Dere and Kaan Aytekin for their endless understanding and being there for me when I needed them. Their belief in me and their presence in my life have made this academic journey meaningful and fulfilling. Their constant encouragement has meant the world to me.

Thank you all for being an essential part of my journey and for making this thesis possible.

Evaluation of Comment Sentiments in Educational YouTube Videos

Elif Naz Özdamar

Data Science, Master's Thesis, 2023

Thesis Supervisor: Yücel SAYGIN

Keywords: sentiment analysis, gender bias, educational videos, online education

Abstract

In the digital era, YouTube has become one of the most popular resource for online education. A prior study has reported a bias towards male narrators in YouTube search results for educational videos. Videos with male narrators are exposed more on the platform for both in STEM(Science, Technology, Engineering, and Mathematics) and NON-STEM subjects. To further investigate the bias, the user generated comments are used in this thesis. To better understand viewers' interactions with YouTube videos, it is crucial to explore the sentiment patterns in the comments. Moreover, user engagements differ between genres. This thesis aims to explore the sentiment expressed in comments on educational YouTube videos. The research has two objectives. Firstly, understanding the sentiment behavior of users towards male and female narrators. Secondly, investigating the sentiment pattern between different subjects: STEM and NON-STEM. In the first part of the study, video's ranking on YouTube platform is taken into consideration to further investigate the behavioral change between the videos with different rankings. In the second part, the comment's ranking is taken in to account to understand the comment ranking behavior. For both of the parts the comment sentiment patterns are examined and the behavior is compared between perceived genders of the narrators and subjects. By addressing these objectives, this thesis aims to understand the underlying sentiment behavior behind comments and the differences between the perceived genders of video narrators and subjects in the context of educational content on YouTube.

Eğitim İçerikli YouTube Videolarında Yorum Duygu Analizinin Değerlendirilmesi

Elif Naz Özdamar

Veri Bilimi, Yüksek Lisans Tezi, 2023

Tez danışmanı: Yücel SAYGIN

Anahtar Kelimeler: duygu analizi, cinsiyetçi ön yargı, eğitim içerikli videolar, çevrimiçi eğitim

Özet

Günümüzde YouTube, çevrimiçi eğitim için en polüler kaynaklarden bir ihaline geldi. Platformda eğitim içerikli video'lar göz önünde bulundurulduğunda, YouTube arama sonuçlarında erkek video anlatıcılarına yönelik bir cinsiyetçi ön yargı olduğu saptandı. Bu ön yargı hem STEM(Bilim, Teknoloji, Mühendislik ve Matematik) hem de NON-STEM ile ilgili içeriklerde mevcut. Erkek anlatıcılar kadın anlatıcılara göre platformda daha fazla temsil ediyor. Bu yanlılığı daha fazla araştırmak için bu tezde kullanıcıların video'lara yaptığı yorumlar kullanılmıştır. İzleyicilerin etkileşimlerini daha iyi anlamak için yorumlardaki duygu analizi davranışlarını incelemek önemlidir. Ek olarak, kullanıcı etkileşimleri video türleri arasında farklılık da gösterebilir. Bu tez, eğitici YouTube videolarındaki yorumlarda ifade edilen duyguyu davranışını keşfetmeyi amaçlamaktadır. Araştırmanın iki amacı vardır. İlk olarak, kullanıcıların erkek ve kadın anlatıcılara karşı duygu analizi davranışlarını anlamak. İkincisi, farklı eğitim konuları arasındaki duygu davranışının araştırılması. Çalışmanın ilk bölümünde, farklı sıralamalara sahip videolar arasındaki, ikinci bölümünde ise farklı sıralamalara sahip yorumlar arasındaki yorum duygu analizi incelenmektedir. Her iki bölüm için de platformdaki yorum duygu analizleri incelenir ve anlatıcıların algılanan cinsiyetleri ile video konuları arasındaki davranış karşılaştırılır. Bu tez, yorumların altında yatan duygu davranışını video anlatıcılarının algılanan cinsiyetleri ve video'nun konusu bağlamında incelemeyi amaçlamaktadır.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The objective of this thesis is to understand the relation between comment sentiment patterns of educational YouTube videos and explore the difference in user behavior among videos with different genders and subjects while considering the rankings of videos and comments. In Section 1.1 the motivation behind this study is explained by highlighting the increasing popularity of YouTube platform for online education and the observed bias towards male narrators in educational video search results within different educational subjects. Section 1.2 provides how this thesis addresses this concept with research questions and overview of the important outcomes.

1.1 Motivation

In recent years, YouTube become one of the most preferred online education field [1]. With the vast amount of content creators, thousands of videos are uploaded every day providing diversity in context and ease of use [2]. Recent studies show that the YouTube displays gender bias in education related search results [3]. There is an inequality in the context of gender representability in which the videos with perceived gender female narrators are displayed less than male narrators. Female narrators tend to receive less visibility than males especially in STEM (Science, Technology, Engineering and Mathematics) related search results. This positive bias towards male narrators stems from both the data, uploaded videos in the platform, and the video sorting algorithm on the platform. Not only in online platforms, STEM is male dominant field in general [4]. With this finding in mind, we wanted

to further analyze if this bias exists in user generated content as well.

In [5] and [6], it is stated that women YouTuber's receive more aggressive and hateful comments in the platform. These studies covered different genders in YouTube such as comedy, gaming, sports, how to, etc. and conclude on the fact that the user generated YouTube comments display more negative sentiments towards women.

In this thesis we used this findings as a base and formed our research questions in order to further address the biased behavior in the YouTube platform. The objective of this study is to further address the gender bias in educational YouTube videos by considering user generated comments. We aim to provide a comprehensive study to understand user engagements by analyzing comment sentiment behavior in the context of educational videos.

1.2 Overview of the Methodology and Contributions

In this thesis we investigated the relationship between YouTube comment sentiments and following variables: perceived gender of the video narrator (male and female), subject of the video (STEM and NONSTEM), video ranking in the search result and comment ranking. We formed research questions in order to employ this study systematically. The research questions begin with a broader perspective and subsequently we incorporate the mentioned variables in order to identify the behavioral differences in a more detailed way. A statistical test is conducted for each of the research question to conclude our hypotheses.

We studied research questions in two parts: video rank related experiments and comment ranking related experiments. With video ranking based analyses, we focused on how perceived gender and the video subject affect the comment sentiment behavior in educational context by also considering the video's ranking on the platform. The main findings in this theses are summarized below:

1. Within the same subject (STEM or NON-STEM), different video rankings do not change sentiment scores significantly. Within the same video ranking

group there is no significant difference between STEM and NON-STEM related videos.

2. Within the same subject and video ranking, there is no comment sentiment difference between genders.

3. Male narrators receive more positive videos in NON-STEM field than STEM.

The second part of the studies focus on the similar variables, gender of the narrator and video field by also considering the comments' rankings in the platform. The research questions tries to comprehend the different patterns caused by these variables. he key findings are summarized below:

1. Comments with upper ranks display more positive sentiments compared to the ones below. This behavior is same for different educational contexts STEM & NON-STEM and perceived genders of narrators and their combinations.

In Section 2 we discuss related works that forms our research questions. This section provides an overview of studies and research relevant to our topic. Section 3 covers the research questions and gives detailed explanations on how the data is collected, what data cleaning and pre-processing steps are applied, which method is used to predict the sentiments of comments and how statistical testing is conducted. Then in Section 4 the hypothesis testing results of mentioned research questions are presented. Finally, Section 5 sums up the research and addresses the limitations and possible improvement areas.

Chapter 2

Related Work

In this section we covered prior studies that inspired and guided our research. Section 2.1 covers the gender biased towards male video narrators in educational YouTube videos and how YouTube search algorithm amplifies gender bias. Section 2.2 specifically covers the academic sources that focuses on gender bias in user generated contents.

2.1 Gender Bias in Educational YouTube Search Results

With the COVID-19 lockdown periods, the popularity of online learning materials accelerated quickly [2]. Students are relying on online resources to learn new subjects or use it as an additional source to complement their education. In [3], the authors investigated the gender bias in online education by focusing on YouTube search results. Their study examined if a certain gender is represented more in the platform. The authors used the equality of representation as a bias measure. In their study they investigated the behavioral difference of two educational different subjects: STEM and NON-STEM. The results show that although YouTube stated on platform policy [7] that they design their algorithms to avoid any bias towards genders and there is no bias present in the platform that discriminates videos based on their gender, the findings in [3] states that male videos are presented more on YouTube in educational context. According to the results, YouTube search results

in educational context are biased towards videos with male narrators concluding that male videos are represented more. Although, this bias exists in both STEM and NON-STEM related queries, they concluded that STEM is even more biased than NON-STEM. Similarly, in their study [8] the authors stated that YouTube does not provide gender diversity in their platform.

In their study [3], the authors also measure the source of this bias and conclude that both data end the search algorithm is biased towards male narrators. These findings provide starting point for further exploring gender bias on YouTube in this thesis.

2.2 User Generated Gender Bias in Online Platforms

In this thesis we aim to comprehend the sentiment pattern changes between perceived genders of video narrators in educational YouTube context. The focus is on STEM and NON-STEM related videos to understand the potential difference in how male and female narrators' comment sentiments behave. Women are not as represented as males in STEM related occupations [9]. They win less awards, gain lower salaries etc. than their male counterparts [8].

In their work [4] the authors investigated the attitude towards female in STEM. They argued that STEM is a male dominant field and women are not as equally represented as males. In recent studies, it is stated that in social media women are exposed hateful and sexist behaviors more than males and they are harassed more in online platforms [10]. And this behavior affects females more than males as they spent more time in social media. In their paper, Fouad and Alkooheji [4] investigated the tweets in Twitter towards women in STEM. They applied BERT based NLP methods to identify the sentiments of tweets and they concluded that the attitude towards women in STEM in Twitter is mainly positive and didn't specifically see gender bias.

Besides other social media platforms, we also cover the academic studies that focus on YouTube specifically. In their study, Wotanis and McMillan investigated the

gender bias towards female narrators on YouTube [5]. They focused on popular female and male YouTuber's Jenna Marbles and Ryan Higa. They gathered the top 10 videos from their channel and retrieved approximately 1000 comments each. Then, they label the comments as either supportive or hostile and observe the differences in sentiment patterns. The results show that female YouTuber gets significantly more hostile comments than the male YouTuber.

In their research, Döring and Mohseni also investigated the attitude towards male and female YouTuber's and tested if gendered hate speech in YouTube [6] exists. They collected more extensive dataset on German speaking YouTube videos. They collected most popular videos from different genres: comedy, gaming, how to and fitness and compare the level of negativity in genders. They concluded that, women tend to receive more aggressive and hostile comments on YouTube and the positive comments mostly in context of physical appearance [6].

Amarasekara and Grant's article [8] also supports this finding, stating that most of the STEM related content creators in YouTube are male and out of the top 50 most subscribed channels in STEM only 2 of them are female. They also compared the comment sentiment between male and female creators by manual labeling. The authors concluded that females receive more hostile, sexual, critical comments in STEM related videos.

All these prior works lead us to investigate the user generated gendered bias in YouTube specifically in educational context. The majority of the studies covered in this section that investigates the comment sentiment behavior in YouTube did not utilize the state of the art machine learning techniques in their research. To address this gap in the literature, we applied advanced machine learning methods to gain deeper insights regarding the sentiment patterns of different genders of video narrators in educational YouTube videos, focusing on both STEM and NON-STEM related subjects.

Chapter 3

Problem Definition and Methodology

In this chapter we first define the problem that is studied in this thesis in Section 3.1. Section 3.2 covers the methodology applied in our study in detail.

3.1 Problem Definition and Research Questions

In this section first we define the problem that this study aims to address in Section 3.1.1. Then in Section 3.1.2 we clarify the the research questions that guide our research. These questions will explore how comment sentiment behavior in YouTube changes depending on different variables such as gender, query field and rankings of videos and comments.

3.1.1 Problem Definition

The objective of this thesis is to examine the behavior of user generated comments' sentiments in educational YouTube videos. The study aims to understand how the comment sentiments differ based on certain variables. These variables include the perceived gender of the video narrator (male/female), the context of video (STEM/NON-STEM) by considering the video rankings and comment rankings. The study starts with the general questions and then we introduce experiments in more granular level in order to understand the differences in comment sentiment

behavior. By analyzing the sentiment behavior under different conditions, we aim to understand how certain variables effect the sentiment patterns.

3.1.2 Research Questions

In this research, the objective is to analyze the comment sentiment behavior in educational YouTube videos. To understand the behavioral change in sentiment, we introduced subsequent variables: video's rankings in the search result, comment's rankings on the video, perceived gender of the video's narrator and the subject of the search query from which the video is retrieved.

We divided research questions and respective experiments into two according to their context: video ranking based and comment ranking based. Video ranking based experiments takes corresponding video's ranking in the YouTube search results. On the other hand, comment ranking based research questions include video comment's ranking to the experimental setup. In each study, we started with the most general question and subsequently added supplementary variables to identify the comment sentiment behavior in detail for educational YouTube videos. Research questions studied in this thesis are listed below.

Video ranking based research questions:

- **RQ1:** Do the positive sentiment scores of the videos vary across different rankings of a search result?
- **RQ2:** Do the positive sentiment scores of the videos within the same ranking differ when comparing STEM and NON-STEM queries?
- **RQ3:** Do the positive sentiment scores of the videos within the same ranking and subject vary across perceived gender of the videos?
- **RQ4:** Do the positive sentiment scores of the videos within the same ranking and gender vary across different subjects?

Comment ranking based research questions:

- **RQ1:** Is there a relation between the positiveness of the comments in YouTube videos and their rankings?

- **RQ2:** Is the comment sentiment more positive in higher ranks for each query field STEM and NON-STEM?
- **RQ3:** Is the comment sentiment more positive in higher ranks for each gender male, female and other?
- **RQ4:** Is the comment sentiment more positive in higher ranks for each gender and query field pairs individually?

3.2 Methodology

In this section the overall methodology applied in our study is covered end-to-end. First in Section 3.2.1 the obtained dataset from previous study will be explained. Section 3.2.2 covers how YouTube comment data is collected by using the YouTube Data API. Before conducting our experiments we applied some data cleaning processes and Sectiom 3.2.3 focuses on these steps. In Section 3.2.4 the NLP model used to predict sentiments is presented and the scoring methodology is explained in detail. In Section 3.2.6 we present descriptive analysis to further understand the data. Section 3.2.7 clarifies the pre-processing steps that prepares the data to test research questions and finally in Section 3.2.8 we cover the experimental setup on how the research questions are tested.

3.2.1 Dataset

In this thesis, we used the search queries provided in PhD thesis "Bias in search: Evaluating search results through rank and relevance based measures" [3]. In that study, gender bias in online education is analysed by using YouTube search results. For detailed information on previous study please refer to Section 2.1.

The dataset consists of educational queries related to STEM and NON-STEM domains. STEM denotes the academic fields of science, technology, engineering and mathematics, whereas NON-STEM refers to courses outside of STEM's scope. Examples of NON-STEM subjects include art, literature, humanities and management, among others [3]. A total of 5 topics were selected for both STEM and NON-STEM domains by the authors. For each topic, 10 queries are provided. The query topics

and search texts are obtained from TheUniGuide website which is an online platform that provides information for students to guide their university options [11]. For the selected queries a set of 200 videos are collected by using YouTube Data API by the author. In this thesis, we focused on top 20 videos retrieved from these queries. The topics are presented in Table 3.1.

Table 3.1: Subjects and sub-fields of collected YouTube videos

Query Field	**Query Sub Field**
NON-STEM	English Language and Literature
	Politics
	Psychology
	Public Relations
	Sociology
STEM	Biology
	CS
	Chemistry
	Maths
	Physics

In the previous work [3], the audio components of the videos were acquired alongside below elements. The audio data is used to predict the narrators' perceived gender. In order to assign a gender to videos, first the audio information is retrieved. The downloaded audio information is then used to annotate genders by using speech recognition model. The model provided in article [12] allows to assign perceived gender based on an audio information. The segmenter is trained based on the dataset provided in by [13]. Provided INA's Speaker Dictionary contains 32000 samples that consists of of 94 hours of male and 27 hours of female speakers. First step of gender detection is separating audio data into parts in order to segment the data. Separated audio file contains three files: music, speech and noise. Then, the speech part is used to detect the perceived gender of the video. As a result, inaSpeechSegmentor provides scores for each gender male and female. The dataset provided by [3] contains gender scores for each video. In this thesis, we used these

scores to assign perceived genders to the collected YouTube videos.

Furthermore, the subsequent list provides explanations of data acquired from a prior study. Retrieved data from previous research [3] are used in this thesis to identify the comment sentiment behaviour.

- **query_field:** General subject of a query, STEM or NON-STEM
- **query_sub_field:** Subject of a query
- **query_text:** Search query for YouTube
- **video_id:** Unique YouTube video id's
- **video_rank:** Search result ranking of a video for respective query
- **audio_downloadable:** If the video's audio can be downloadable, binary field
- **sampling_speech:** Amount of speech in the audio
- **language:** Speech language in the audio
- **sampling_biased_audio_male_ratio:** The likelihood of a narrator being male
- **sampling_biased_audio_female_ratio:** The likelihood of a narrator being female

3.2.2 Data Collection

To conduct the sentiment analysis of comments on YouTube search results for videos, YouTube Data API is used as a data collection tool. YouTube Data API enables executing YouTube commands outside of the YouTube platform programmatically and retrieve the results [14]. The API enables collecting various types of data: YouTube video search result (video title, video id, channel id, etc.), video details (title, description, view count, like count, etc.), channel details (title, subscriber count, video count etc.).

We used CommentThreads resource to obtain the comments of the retrieved videos. List method provides collecting comments written in videos or channels from the

YouTube platform. Below parameters are used to retrieve comments for each video collected.

- **video_id:** The video's unique identifier from which comments are to be collected
- **order:** Sorting option for comments

The YouTube platform provides two comment and video sorting options: relevance and time. Relevance is the default sorting behavior on the platform fed by an algorithm. Time is another sorting option to retrieve the comments/videos ordered by its created time. In this research, comments and videos are obtained based on their relevance, which aligns with the default behavior of the YouTube platform. Per video, a maximum of 500 comments and their rankings are gathered.

3.2.3 Data Cleaning

Prior to conducting experiments, data cleaning methods are performed. The dataset contains predictions pertaining to the perceived genders of both male and female, which are subsequently utilized to label the perceived genders of the videos. However, some videos in the dataset could not be assigned gender labels because of audio related constraints. We did not assign gender labels to the videos displays these constraints.

One of the contributing factors for this inability to predict the video narrator's gender is audio unavailability. The audio file could not be downloaded because either the video has been made private or deleted. As a result, the *audio_downloadable* information for these videos is annotated as 0. Out of 2000 videos, audio of 23 of them could not be downloaded.

Another contributing factor is the absence of speech elements in certain videos. They only contain visual content. These videos may either lack any auditory components or a speech, thereby the gender of the video narrator cannot be determined through audio analysis methods. Out of 2000 videos, audio component is not available for a subset of 68 videos. *sampling_speech* value of those videos are 0 as the audio does not contain speech.

Furthermore, some portion of the retrieved videos are not in English. Consequently, these videos have been excluded from the dataset as the comments obtained would not be in English. The dataset comprises 215 videos with non-English speaker narrators.

The videos with unavailable audio files and with non-English speakers are excluded from the data and the videos without a speech part are labeled as "no-narrator" and examined differently in Section 3.2.6.

The comments of the videos are collected by using YouTube Data API as aforementioned in Section 3.2.2. Some videos do not contain any comments. These videos are excluded from the dataset since comment sentiment scores cannot be generated for them.

3.2.4 Sentiment Analysis Model

In this thesis we used sentiment analysis task to predict comment sentiment behavior. We used a Robustly Optimized BERT Pretraining Approach (RoBERTa) based model in our predictions. In this section the sentiment analysis method is briefly explained along with its importance in NLP tasks. Then in Section 3.2.4 the RoBERTa model is discussed and how we ensure its effectiveness in YouTube data is elaborated.

Natural Language Processing and Sentiment Analysis Overview

Natural Language Processing (NLP) is a field in machine learning that focuses on understanding and generating human language texts. In recent years, understanding human generated texts become even more significant as the Internet becomes primary source of information and users frequently engage with contents [15]. The NLP tasks can be classified in two parts: Natural Language Understanding (NLU) and Natural Language Generation (NLG) [16].

NLU tasks aim to understand human generated texts and use it for further analyses. Some of the commonly used NLU tasks are as follows: part of speech tagging, sentiment analysis, named entity recognition, text classification, etc. The shared goal of each NLU task is to enable computers to interpret human language and use

this information in various applications [16]. On the other hand, the aim of NLG tasks is to produce human-like natural languages. Some of the NLG tasks are as follows: text summarization, chatbots, machine translation, etc. In this thesis, our aim is to comprehend the attitude in user generated YouTube comments. Therefore, we will be interested in NLU part of the field with a specific focus on sentiment analysis. Our aim is to gain insights from the user attitudes by using the textual data presented in YouTube.

The aim of sentiment analysis (opinion mining) is to understand the main attitude expressed in a text [15]. The primary aim of sentiment analysis is to classify the sentiment of a given text. This task is not only used in science related domains but also in various other fields such as: marketing, management, customer surveys, etc. Although the most popular sentiment analysis task is to determine if a test is negative, neutral or positive, there are other tasks that extract different types of sentiment cues from a text as well. Some of them are determining sarcasm, anger, irony, emotion, hate, offensive, etc. In this thesis we focused on extracting positivity or negativity of user generated YouTube comments in educational context.

RoBERTa Sentiment Analysis Model

In this thesis, we used a Robustly Optimized BERT Pretraining Approach (RoBERTa) based sentiment analysis model to predict YouTube comment sentiments. Before introducing RoBERTa, it is important to know basics of Bidirectional Encoder Representations from Transformers (BERT) as RoBERTa is based on BERT. BERT is an NLP model proposed by Google AI researchers [17]. BERT outperformed 11 state of the art NLP tasks with considerable performance improvements on various NLP evaluation benchmarks such as GLUE [18], MultiNLI [19] and SQuAD [20] and become the new state of the art. Different from previous state of the art methods, BERT is trained in bidirectional fashion which allows to comprehend text in more holistic way.

RoBERTa is a BERT based model. In [21] the authors stated that they realized BERT is undertrained and has lots of potential for improvement. Therefore, they trained BERT based model with improvements. The difference between BERT and

RoBERTa is that RoBERTa uses more dynamic way to mask the inputs that results more robust model. Moreover, the authors addressed the problem of undertraining in BERT by training the model longer, using larger batches with more data. By this way, RoBERTa become new state of the art method by outperforming in well known NLP evaluation benchmarks such as GLUE [18] and SQuAD [20]. Therefore, in this thesis we selected RoBERTa to predict sentiment scores of comments.

Specifically, we used a RoBERTa based model finetuned on Twitter dataset with sentiment analysis objective. The model "Twitter RoBERTa Base for Sentiment Analysis" is published in [22]. In their paper, the authors provided a benchmark for tweet classification tasks. The tasks are as follows: sentiment analysis, emotion recognition, offensive language detection, hate speech detection, stance prediction, emoji prediction, and irony detection. In recent years, transformers based language models dominated the state of the art methods in most of the NLP tasks and sentiment analysis is one of them. In their paper, authors tried different modelling techniques to provide a compatible benchmark for Twitter data. The model is trained on Twitter data proposed in [23] that consists of 50k train and 12k test samples. The model output gives three scores for following class labels: negative, positive and neutral. Table 3.2 shows three comments and their respective sentiment scores for each class. The detailed information on how this sentiment scores are used in our study will be explained in Section 3.2.8.

Table 3.2: Sample Sentiment Scores of Comments

Comment	**Positive Score**	**Negative Score**	**Neutral Score**
I'm studying for Sociology and you explained this so well. Thank you!	0.98	0	0.02
my teacher didn't send me here I want to learn this on my own :)	0.47	0.11	0.42
Why keep saying it is difficult. Your creating a negative towards the subject. Perhaps if you give a positive approach people won't feel defeated before they even start.	0.12	0.41	0.47

Model Validation

As explained in previous section, Twitter RoBERTa Base Sentiment model is fine tuned on tweets [22]. However, in our case we predict sentiments of YouTube comments. Although, RoBERTa is pre-trained on huge text corpus which allows the model to perform well on different contexts, we wanted to cross validate the model performance in order to ensure the reliability of its results.

To evaluate RoBERTa's performance, we used a labeled dataset. This dataset helped us to understand how Twitter RoBERTa performs on YouTube comments. The dataset to validate the model is obtained from an article [24]. In this work, the authors built their own YouTube comment dataset by using Youtube Data API. They collected 10000 comments from YouTube tutorial videos. Then they labeled the dataset by themselves. The classification task had 6 labels: corrective, imperative, interrogative, miscellanceous, negative and positive. Different from the model used in this thesis, this dataset contains additional labels. In order to be consistent with our study, we only used negative and positive labeled comments from this dataset.

To conduct model validation we predict comment sentiment scores by using Twitter RoBERTa Base Sentiment model. In order to compare the model performance we

assigned labels to RoBERTa sentiment predictions by labeling the comment with highest sentiment score: negative, neutral or positive. Furthermore, we examined the model performances by observing classification metrics: precision, recall and F1 score.

Table 3.3 shows the precision, recall, and F1-score metrics for two classes, negative and positive for the validation dataset. For the negative labels the RoBERTa model achieved 84% F1 score, that combines the two metrics precision and recall. Similarly, for positive labels the model get 95%. This results show that RoBERTa model that is fine tuned on Twitter data for sentiment analysis performs good in YouTube comment classification task as well. With these results, we cross validated the performance of RoBERTa model that is fine tuned on Twitter data for sentiment analysis task and concluded that it is robust enough to conduct our analysis.

Table 3.3: Precision, Recall, F1-Score for Validation Dataset

Label	**Precision**	**Recall**	**F1-Score**
Negative	0.87	0.80	0.84
Positive	0.99	0.91	0.95

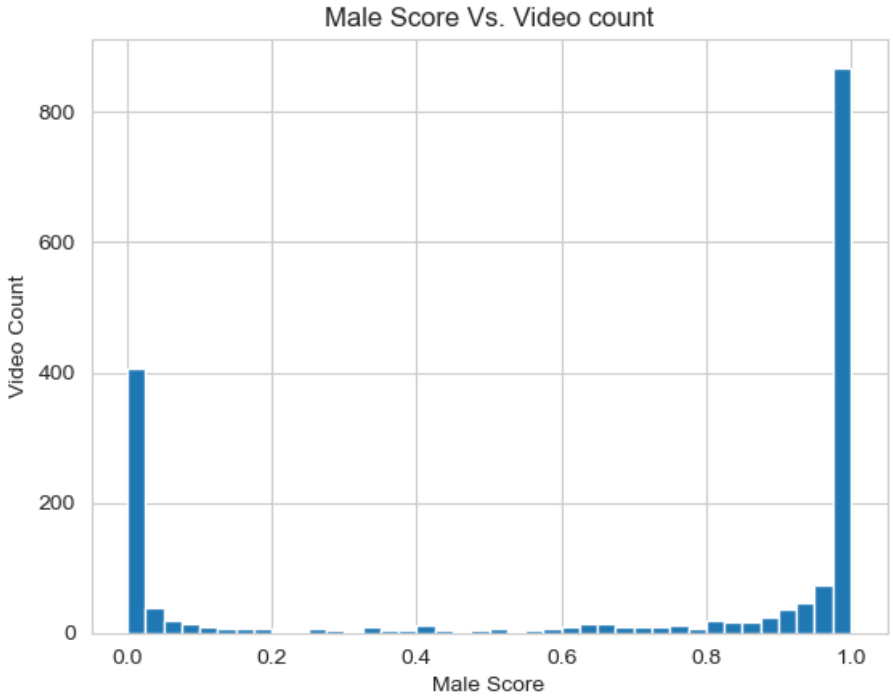
3.2.5 Gender Assignment

The videos are assigned perceived genders based on audio gender scores. The gender scores are calculated by using the speech audio information of the videos. Detailed explanation of how audio gender scores are calculated is provided in Section 3.2.1. In order to validate the confidence to provided gender scores, we examined the gender score distributions.

Figure 3.1 provides the video male gender score distribution obtained from the audio data. As the scores are probabilities, the male and female gender scores sum up to one and the histogram of male scores is the mirror image of female scores. We can infer from Figure 3.1 that the audio gender scores of the videos are mostly fully polarized. Each audio gender score represent the probability of the video having male/female narrators. Out of 1768 videos with available audio files and containing English speech, 1077 of them have audio gender scores of either 1 or 0, the model

to predict audio's gender gives fully polarized and confident scores. Furthermore, the video count increases when the audio male scores approach to the edges (0 and 1), inferring that there are less samples where the gender of the narrator is less confident.

Figure 3.1: Audio Gender Score Histogram



To further validate the gender scores, we took random video samples from the most confident and the least confident gender scores and manually checked if the gender attributions are correct. We observed that when the difference between gender scores get closer the confidence decreases significantly leading to incorrect assignments. We noticed that high pitched male voices, low pitched female voices and videos with multiple narrators with mixed genders are tend to be misclassified. Therefore, in order to make more confident gender assignments, we introduced another gender class named other and assigned this label to the videos with gender scores between 0.05 and 0.95. In other words, if the gender score indicates a label with 0.95 probability, we assigned the videos as to that gender. Otherwise, we classified them as other. With this approach, our primary goal was to ensure high confidence in gender attribution.

3.2.6 Descriptive Analysis

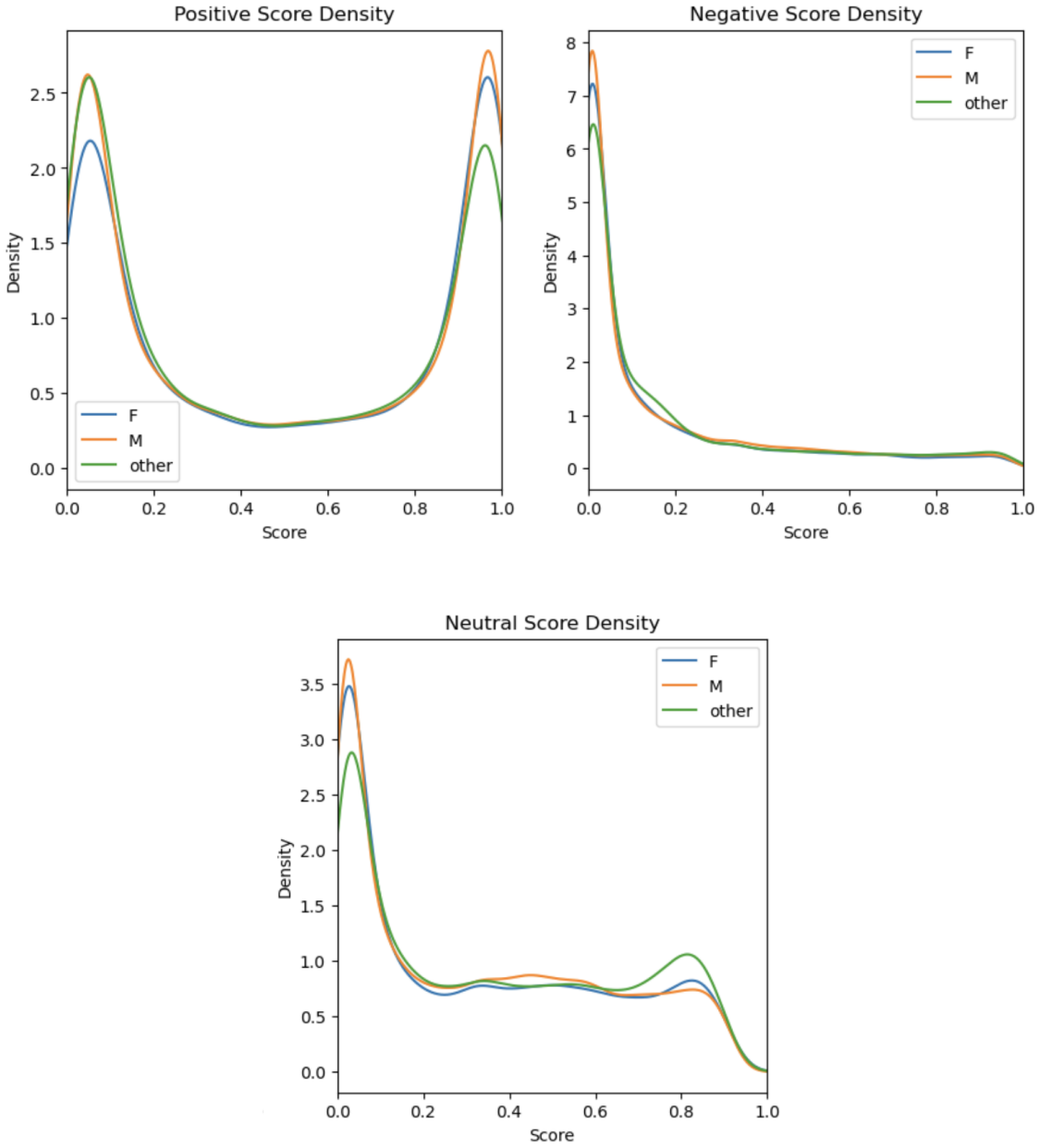
Prior to conducting statistical analysis on research questions, we employed descriptive analysis on the data. This method is applied to provide an overview of the data and identify the fundamental tendencies and distributions. By conducting descriptive analysis, a comprehensive understanding of the dataset's general characteristics are obtained and used in further steps to better interpret the data.

In some of the analysis, we conducted hypothesis testing to identify patterns between groups. We used Welch's t-test with bootstrapping, and the confidence level was set at 0.95 for all the tests. The column labeled "result" in the hypothesis test tables refers to the hypothesis testing result. If the column value is "Fail To Reject" it means we didn't have enough evidence to reject the null hypothesis and there is no statistical difference between mean values.

Comment Sentiment Scores

The comment sentiment scores are obtained by Twitter RoBERTa model as previously discussed in Section 3.2.4. The model provides three sentiment scores for each class positive, negative and neutral for a given text. For each comment the class scores are sum up to one. In Figure 3.2, the probability distributions for each gender's class scores are illustrated. The graphs demonstrate that the score distributions are generally similar between gender labels female, male and other. However, on the edges, where the scores are closer to 0 and 1, the scores display different patterns. Positive score distribution has two peeks around extreme score values, showing that the positive score tendency is mainly low or high. In other words, the positive score distributions shows bimodality for each gender. Conversely, negative score distribution is right skewed and the majority of the comments have low negative scores. Similarly, comment neutral scores are mostly concentrated on lower values.

Figure 3.2: Sentiment Score Distribution by Gender



Video Count

The dataset consists of educational videos that are taken from YouTube platform search results. Collected videos are examined in two subjects: STEM and NON-STEM. Detailed explanations of STEM and NON-STEM fields are presented in Section 3.2.1 and in this thesis we denoted those subjects as query fields. Additionally, there are 5 topics for each query field and for each topic there are 10 queries. For each query top 20 videos are collected. Altogether, we have 100 queries, 50 for STEM and 50 for NON-STEM. Consequently, comprehensive analysis is conducted

on a dataset consisting 2000 videos. Table 3.4 shows the distribution of video counts in the dataset. Out of these 2000 videos unavailable audio cases are eliminated and as mentioned in Section 3.2.3 and the descriptive analyses are conducted out of 1813 videos.

Table 3.4: Video Counts by Case

Case	**Count**
no-narrator	45
unavailable audio	187
female	445
male	942
other	381

Comment Count Statistics

User entered comments on a YouTube video are obtained by using YouTube Data API as explained in Section 3.2.1. For our study, we have obtained top 500 comments for each video by using the API. If a video contains more than 500 comments, the first 500 of them are retrieved. Therefore, the analyses done in this section are based on first 500 comments for each video. In total 446 videos do not contain comments.

The initial analysis focuses on the video comment counts. We first calculated the total comment counts in each video and observed its statistics. Table 3.5 shows the overview descriptive analysis of video comment counts. The mean, percentile 50, 75, 80 and 90 values are 98.38, 10, 82, 149 and 500 respectively. The comment counts in the data are not distributed evenly. Half of the videos have comment counts smaller than 10 and most of the videos have comment counts around 82 or less. This pattern shows that the data is skewed, it has more values on smaller side. The histogram of comment counts for retrieved YouTube videos are presented in Appendix A.

Table 3.5: Summary Statistics of Video Comment Counts

Sum	**Mean**	**Min**	**Median**	**P75**	**P80**	**P90**	**Max**
173,933	98.38	0	10	82	149	500	500

In this thesis, the objective is to understand the comment sentiment behavior based on some variable factors. Independent variables to be evaluated in hypothesis testings are ranking information of the videos, ranking information of the comments, gender of the video narrator and the subject of the video (STEM or NON-STEM). We observed the mean and median comment counts per variable in order to determine if the dataset has preliminary biases.

First variable to analyze is the query field of the video. In Table 3.6, we can see that although the gathered video counts are same for STEM & NON-STEM fields, there are more comments presented in STEM related videos. The mean and median values for STEM and NON-STEM videos also show that this difference is not caused by the outliers but the STEM videos tend to have more comments.

Table 3.6: Statistics of Comment and Video Counts by Query Field

Query Field	**Video Count**	**Mean Comment Count**	**Median Comment Count**
NON-STEM	895	73.09	5
STEM	873	124.29	20

The other variable tested in our experiments is the perceived gender of the video narrator. The statistical summary of comment count of the videos are presented in Table 3.7. Out of 1768 videos there are 445 videos whose narrators' are female and 942 of the videos have male narrators. The number of videos labeled as male is approximately twice that of videos labeled as female. This bias is expected because of the study we discussed in Section 2 related works [3]. However, having gender and query field bias in the dataset does not affect our experimental setup because we aggregate the data in video level and make the data independent from the comment count of videos. The details of the pre-processing step is explained in Section 3.2.7.

Table 3.7: Statistics of Comment and Video Counts by Gender

Gender	**Video Count**	**Mean Comment Count**	**Median Comment Count**
Female	445	81.64	8
Male	942	114.59	15
Other	381	77.823	16

Table 3.8 shows the statistical summary of the combined version of gender and educational subject variables. This analysis includes mean and median values of the comments by both query field and gender together. Both STEM and NON-STEM related videos male video count is more than female and other labeled videos. Additionally, mean comment count for male videos are higher than female and other labeled videos as well. This is a notable finding to discuss. For both of the query fields male narrators get more comments than female narrators and other class.

Table 3.8: Summary Statistics of Comment Counts by Gender

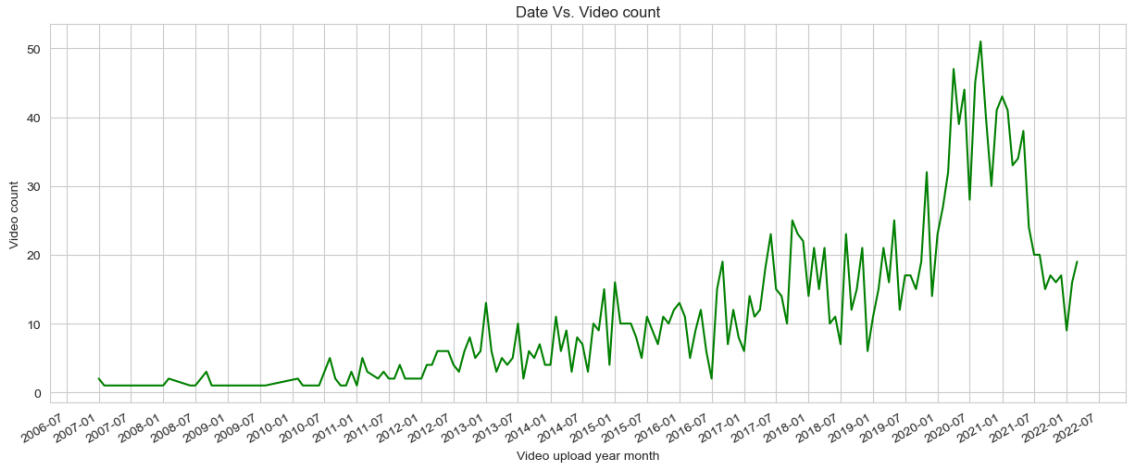
Query Field	**Gender**	**Video Count**	**Mean Comment Count**	**Median Comment Count**
NON-STEM	F	249	51.52	4
NON-STEM	M	446	88.82	6
NON-STEM	Other	200	64.94	3
STEM	F	196	119.89	21
STEM	M	496	137.79	28
STEM	Other	181	92.08	10

Video Publish Date Analysis

The videos were collected via YouTube API on 2022 by Gizem Gezici [3]. To provide a comprehensive understanding of the data, a graph presenting the relationship between video publish dates and video count is created. Figure 3.3 is the visual

representation of how the number of educational videos on YouTube has evolved over time. The graph allows us to observe the trends in the volume of educational content uploaded to the platform. During 2020-03 there is a spike in video upload count. The period where the video counts has a steep increase overlaps with the COVID-19 pandemic period showing that lockdown drove content creators to upload more YouTube videos.

Figure 3.3: Video Publish Date Vs. Video Count



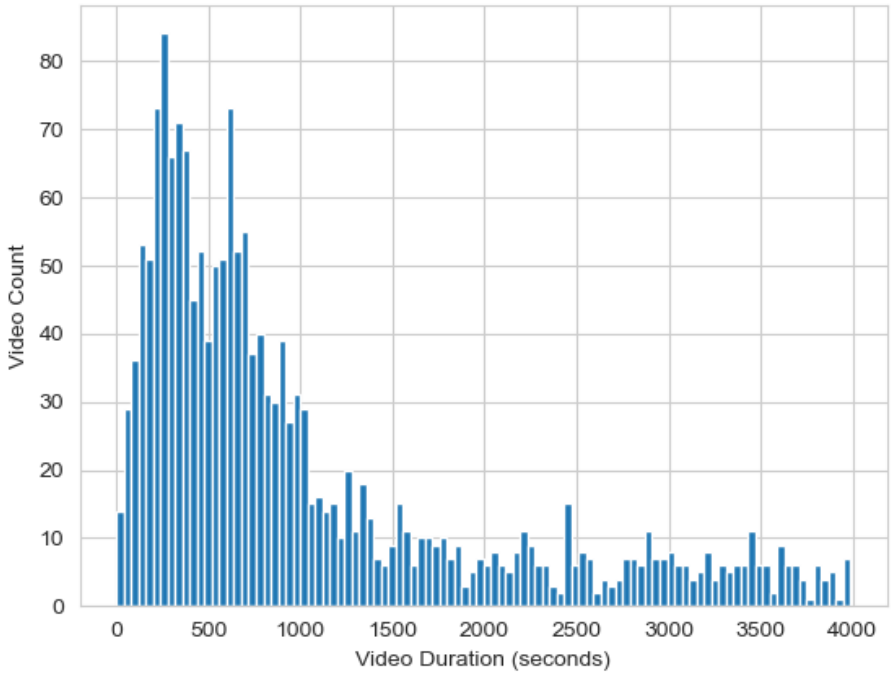
In this descriptive analysis, a comprehensive exploration of video durations in the dataset was conducted by creating a histogram. By plotting the histogram of video durations, we obtain a visual representation of the distribution of video lengths across the dataset. The histogram graphically depicts the frequency of videos falling within specific duration ranges, enabling an assessment of the most common and least common video lengths. This investigation serves as a foundational step in understanding the time dependency of the videos, which can subsequently contribute to further insights into viewer preferences, engagement patterns, and content creator strategies. By examining the distribution of video durations through this histogram, the analysis aims to offer valuable information about the characteristics of educational YouTube videos in terms of their length.

Video Duration Analysis

Figure 3.4 displays the histogram of video durations within the dataset. The duration values in the figures are in seconds and the average video duration is 25.8

minutes.

Figure 3.4: Video Duration Histogram



We conducted hypothesis testing to observe the difference between content creators behavior on video durations. First, we test the hypothesis that the mean video durations are same for narrators with perceived gender male, female and other. Table 3.9 provides the statistical test results. The results show that male narrators tend to upload longer videos than female narrators. However, when we compare female and male videos with other, it is statistically significant that other class uploads longer videos than both male and females. This result could reflect differences in video uploading patterns between different genres. When there are multiple narrators present in the video, the duration tends to be longer. This could be because more discussion happens when there are multiple narrators.

Similar hypothesis testing is done for video query fileds and the video duration is compared between STEM and NON-STEM. Table 3.10 shows the test results and the null hypothesis different subjects having same mean video duration is failed to reject. There is no significant difference between the video durations of STEM and NON-STEM videos.

Table 3.9: Video Duration Hypothesis Test by Gender

Experiment	**Mean**	**Result**
Female vs. Male	(1112.59, 1447.12)	Male > Female
Female vs. Other	(1112.59, 1990.80)	Other > Female
Male vs. Other	(1447.12, 1990.80)	Other > Male

Table 3.10: Video Duration Hypothesis Test by Query Field

Experiment	**Mean**	**Result**
STEM vs. NON-STEM	(1392.30, 1510.79)	Fail To Reject

In addition to above analysis, we also test the effect of educational subject within the same perceived gender of the narrator. Similarly the mean duration difference is examined for gender classes male, female and other within the same query field. Table 3.10 displays the test results, for all of the gender classes the mean duration does not significantly change between different fields STEM and NON-STEM. However, when two genders are compared, for STEM related queries both other and male genders have longer videos than females. In NON-STEM, other class uploads significantly higher videos than both male and female narrators.

Table 3.11: Video Duration Hypothesis Test by Query Field and Gender

Fixed Variable	**Experiment**	**Mean**	**Result**
Female	STEM vs. NON-STEM	(1018.90, 1186.23)	Fail To Reject
Male	STEM vs. NON-STEM	(1474.04, 1417.02)	Fail To Reject
Other	STEM vs. NON-STEM	(1747.06, 2212.26)	Fail To Reject
STEM	Female vs. Male	(1018.90, 1474.04)	Male > Female
STEM	Female vs. Other	(1018.90, 1747.06)	Other > Female
STEM	Male vs. Other	(1474.04, 1747.06)	Fail To Reject
NON-STEM	Female vs. Male	(1186.23, 1417.02)	Fail To Reject
NON-STEM	Female vs. Other	(1186.23, 2212.26)	Other > Female
NON-STEM	Male vs. Other	(1417.02, 2212.26)	Other > Male

In order to observe the video duration patterns of videos with no narrator, we examined the mean value and percentile statistics. For narrator availability case we didn't apply hypothesis testing because the sample size was small. Table 3.12 demonstrates that all statistical values are smaller for the videos with no narrator. This led us to the conclusion that videos with narrators and those without narrators behave differently. Therefore, we excluded videos without narrators in order not to introduce bias into our experiments.

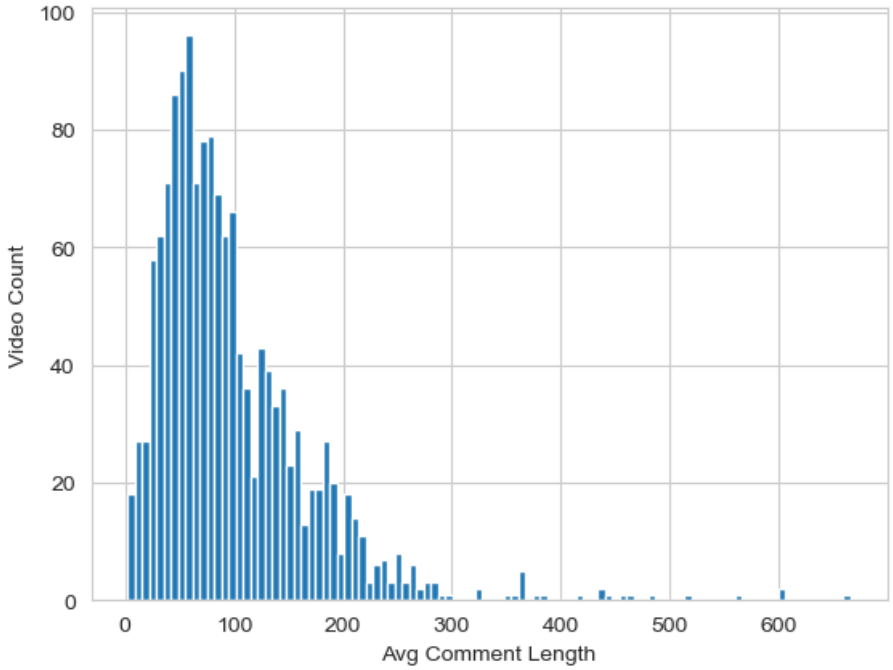
Table 3.12: Video Duration Statistics by Narrator Presence

Has Narrator	**Mean**	**P25**	**P50**	**P75**
False	257.53	41	109	231
True	1546.67	373	728	1855

Comment Length Analysis

We used video comment sentiments as the primary factor to observe user engagement. Besides the sentiment we analyzed comment length behavior of users. The length of the comments represents the character count of comments. The overall average comment length is 102.62 characters.

Figure 3.5: Comment Length Histogram



In order to observe if the comment lengths differ between query fields and perceived genders of narrators, we applied statistical tests. Table 3.13 provides the results of testing null hypothesis female, male and other labeled narrators receive an equal length of comments. Comparison of female and male & other and male showed the male narrators receive longer comments than female and other narrators. Similarly, the test for comparison between mean length of the comments is conducted for educational subjects STEM and NON-STEM. Table 3.14 shows that the mean length of the comments written in STEM and NON-STEM results are significantly different and NON-STEM related videos receive longer comments.

Table 3.13: Comment Length Hypothesis Test by Gender

Experiment	**Mean**	**Result**
Female vs. Male	(88.56, 110.46)	Male > Female
Female vs. Other	(88.56, 98.56)	Fail to Reject
Male vs. Other	(110.46, 98.56)	Male > Other

Table 3.14: Comment Length Hypothesis Test by Query Field

Experiment	**Mean**	**Result**
STEM vs. NON-STEM	(94.52, 111.57)	NON-STEM

Additionally to understand the comment length behavior in detail, we tested each gender within different query fields and each query fields with different genders separately. According to hypothesis testing results shown in Table 3.15, male narrators receive longer comments than females in both of the educational subjects STEM and NON-STEM. Additionally, within each gender the length of the comments are significantly higher in NON-STEM related queries.

Table 3.15: Comment Length Hypothesis Test by Query Field and Gender

Fixed Variable	**Experiment**	**Mean**	**Result**
Female	STEM vs. NON-STEM	(77.68, 99.12)	NON-STEM
Male	STEM vs. NON-STEM	(103.0, 119.21)	NON-STEM
Other	STEM vs. NON-STEM	(89.87, 107.94)	NON-STEM
STEM	Female vs. Male	(77.68, 103.0)	Male > Female
STEM	Female vs. Other	(77.68, 89.87)	Fail to Reject
STEM	Male vs. Other	(103.0, 89.87)	Fail to Reject
NON-STEM	Female vs. Male	(99.12, 119.21)	Male > Female
NON-STEM	Female vs. Other	(99.12, 107.94)	Fail to Reject
NON-STEM	Male vs. Other	(119.21, 107.94)	Fail to Reject

In Table 3.16 the mean and percentile values for comment length are displayed based on the presence of a narrator in the video. Although the mean comment length is higher for the videos without a narrator, all other percentile values are lower and this indicates that there are outliers in the data affecting the average values. We observe a pattern difference for comment length statistics between videos with and without a narrator.

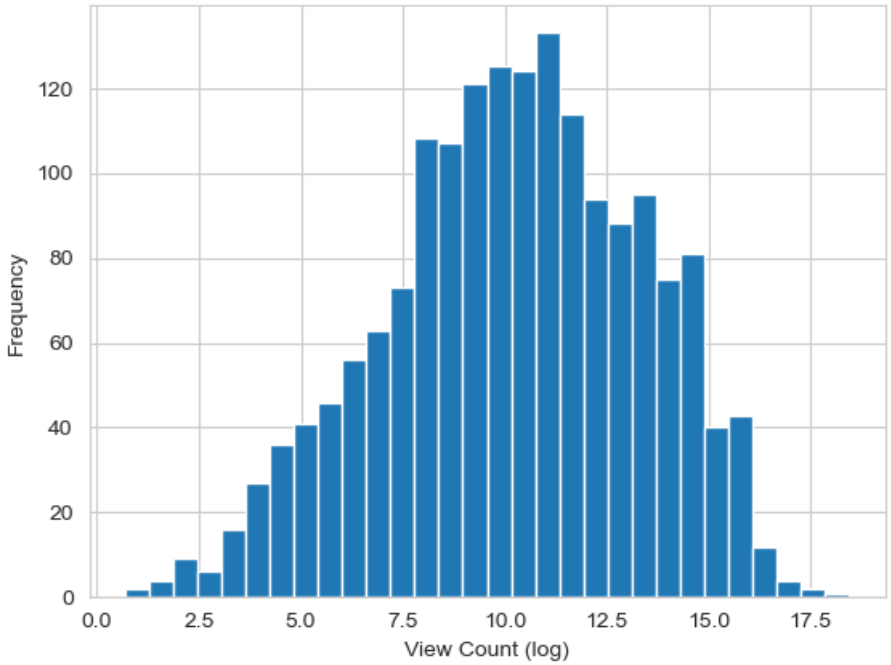
Table 3.16: Comment Length Statistics by Narrator Availability

Has Narrator	**Mean**	**P25**	**P50**	**P75**
False	230.39	24.2 0	46.0	75.46
True	102.62	52.51	85.6	134.79

Video View Count Analysis

The video view counts exhibit wide range of values spanning from around 2 to approximately 100 million. Consequently, to better visualize the distribution of video view counts, we have applied logarithmic scaling to the values. Figure 3.6 displays the distribution of the data and the log scaled plot follows a normal like distribution. This suggests that the original view count data is skewed and the log transformation reduced the skewness.

Figure 3.6: View Count Histogram



In order to explore the user engagement of different genders and educational fields we tested the hypothesis of these variables having equal average view counts per video. The results of statistical test that compares the mean video view counts between narrator genders are presented in Table 3.17. The results show that the male narrators significantly receive more views than both females and others.

There could be several implications for this. Firstly, since the male narrators represented more in the YouTube search results, they might get more views. Higher exposure of male videos on the YouTube platform suggested in prior studies [3]. Secondly, people might prefer male narrators over females. This could be because male content creators make their videos more engaging and consequently get more views. Previous analysis of video duration also supports this conclusion, indicating that male narrators upload longer videos. This observation suggests that male narrators strategize and invest more effort into their content and making it more preferable.

Similarly, the mean difference between STEM and NON-STEM fields are statistically tested. Results in Table 3.18 show that videos uploaded in STEM field gets more views than NON-STEM videos suggesting that STEM has more user engage-

ment.

Table 3.17: View Count Hypothesis Test by Gender

Experiment	**Mean**	**Result**
Female vs. Male	(509111.71, 854323.57)	Male > Female
Female vs. Other	(509111.71, 494797.67)	Fail To Reject
Male vs. Other	(854323.57, 494797.67)	Male > Other

Table 3.18: View Count Hypothesis Test by Query Field

Experiment	**Mean**	**Result**
STEM vs. NON-STEM	(864273.07, 519499.48)	STEM > NON-STEM

Furthermore, we conducted hypothesis testing for each gender and query field pair to have a deep dive understanding of how view count pattern changes between different genders and educational subjects. The results of these statistical tests are shown in Table 3.19. According to the test results, for both female and other labeled videos, the mean video view count does not significantly change between different fields STEM and NON-STEM. However, male videos receive more views in STEM related videos. When we compare the genders within the same query field, we observe that for STEM, males receive more views than both females and others. There is no significant difference for NON-STEM related queries. This statistical test concludes that there is no view count difference between different gender groups within NON-STEM related videos. However, for STEM videos males receive significantly more views than both female and other gendered videos.

Table 3.19: View Count Hypothesis Test by Query Field and Gender

Fixed Variable	**Experiment**	**Mean**	**Result**
Female	S vs. NS	(643045, 402736)	Fail To Reject
Male	S vs. NS	(1050455, 634567)	STEM > NON-STEM
Other	S vs. NS	(591546, 405946)	Fail To Reject
STEM	F vs. M	(643045, 1050455)	Male > Female
STEM	F vs. O	(643045, 591546)	Fail To Reject
STEM	M vs. O	(1050455, 591546)	Male > Other
NON-STEM	F vs. M	(402736, 634567)	Fail To Reject
NON-STEM	F vs. O	(402736, 405946)	Fail To Reject
NON-STEM	M vs. O	(634567, 405946)	Fail To Reject

When we compare the mean and percentile values for videos with and without narrators in Table 3.20, we notice that videos without narrators receive fewer view counts for all statistical measures. This indicates that the presence of a narrator affects behavioral patterns.

Table 3.20: View Count Statistics by Narrator Availability

Has Narrator	**Mean**	**P25**	**P50**	**P75**
False	40382	202.0	556	27004
True	748601	3268.25	28899	281214

Comment Emoji Occurrence Analysis

Another user engagement factor that can be derived from comments is usage of emojis and emoticons. We extracted emojis from user comments and visualize the top 10 most used emojis for the related variable. Table 3.21 and Table 3.22 shows the most used emojis in STEM and NON-STEM subjects respectively. Occurrence ratio represents the share of that emoji among all used emojis. The frequently used emojis in both query fields are relatively similar to each other. The thing that grabs our attention is that NON-STEM queries have negative emotions in top 10 such as "😭" and ":/" while there is no negative emojis in STEM top 10.

Emoji	Occurrence Ratio
:)	9.2%
❤️	8.0%
👍	6.4%
😂	5.4%
🙏	5.3%
:3	4.1%
😊	4.0%
:D	2.9%
😍	2.5%
👌	2.1%

Table 3.21: Top 10 Emoji Occurrence Ratio in STEM Queries

Emoji	Occurrence Ratio
:)	11.6%
❤️	6.7%
😂	5.7%
👍	4.6%
🙏	3.8%
:3	3.8%
:D	3.0%
😊	2.7%
:/	2.1%
😭	1.9%

Table 3.22: Top 10 Emoji Occurrence Ratio in NON-STEM Queries

Similar to above visualizations, top 10 most used emojis for male and female narrators are calculated. Tables 3.23, 3.24 and 3.25 represents the emoji shares in male, female and other gendered videos respectively and the distributions seem similar for different genders.

Emoji	Occurrence Ratio
:)	12.0%
❤️	6.6%
👍	5.9%
😂	5.3%
:3	4.6%
🙏	4.3%
:D	3.5%
😊	3.2%
:/	2.1%
👏	2.0%

Table 3.23: Top 10 Emoji Occurrence Ratio in Male Videos

Emoji	Occurrence Ratio
❤️	9.7%
:)	8.6%
👍	5.9%
🙏	5.5%
😊	5.2%
😂	4.1%
😍	3.6%
:3	2.8%
:D	2.4%
👌	2.1%

Table 3.24: Top 10 Emoji Occurrence Ratio in Female Videos

Emoji	Occurrence Ratio
😂	7.8%
❤️	7.6%
:)	5.9%
👍	5.6%
🙏	5.5%
:3	3.9%
😊	2.7%
😭	2.4%
😍	2.3%
👌	2.3%

Table 3.25: Top 10 Emoji Occurrence Ratio in Other Labeled Videos

Besides most used emoji distribution, we also examined the ratio of emoji usage in overall. Table 3.26 shows that male videos have lower emoji usage ratios compared to female & others and female & other labels are relatively similar. This observation could imply that either the videos created by males engage less emojis or

the commenters that prefer male narrators are less expressive with emojis. Different emoji usage patterns between different genders might indicate the variation between engagement styles within groups.

In Table 3.27 we observe that there is nearly 4% difference between STEM and NON-STEM related videos and STEM has more emoji usage rate.

Table 3.26: Emoji Usage by Gender

Gender	**Comment W/ Emoji Count**	**Comment Count**	**Emoji Usage**
Female	8331	36328	22.9%
Male	17828	107952	16.5%
Other	5953	29653	20.1%

Table 3.27: Emoji Usage by Query Field

Query Field	**Comment W/ Emoji Count**	**Comment Count**	**Emoji Usage**
NON-STEM	10515	65424	16.1%
STEM	21597	108509	19.9%

In order to understand the root source of emoji usage rate differences we observed the percentages for each gender and query field pair. From Table 3.28 we see that for NON-STEM queries emoji usage ratios are similar. However, in STEM queries, female videos outnumber other gender labels in terms of percentages. Females display nearly 9% higher emoji usage rate compared to males.

Table 3.28: Emoji Usage by Gender

Query Field	**Gender**	**Comment W/ Emoji Count**	**Comment Count**	**Emoji Usage**
NON-STEM	Female	2149	12828	16.8%
NON-STEM	Male	5937	39609	15.0%
NON-STEM	Other	2429	12987	18.7%
STEM	Female	6182	23500	26.3%
STEM	Male	11891	68343	17.4%
STEM	Other	3524	16666	21.1%

Similarly Table 3.29 shows the emoji usage percentages by with and without narrator videos and without narrator cases receive less comments with emoji and emoticons.

Table 3.29: Emoji Usage by Narrator Availability

Has Narrator	**Comment W/ Emoji Count**	**Comment Count**	**Emoji Usage**
False	99	667	14.8%
True	32112	173933	18.5%

As a result of all the descriptive analyses, we decided to exclude no-narrator videos from the sentiment analysis experiments. For all of the metrics mentioned in previous parts (video duration, comment length, view count and emoji usage) the statistics are different for the videos with no-narrators. Therefore, we concluded that no-narrator videos exhibit different behavioral patterns and should be excluded from experiments in order not to introduce noise to the data.

Comment Language Analysis

The model used to predict comment sentiment is for English text only as explained in Section 3.2.4. Although non English narrators are excluded from the dataset, we further analyze the language spoken in user comments. XLM-RoBERTa (Cross-Lingual RoBERTa model) based language detection model is used to assign language

to comments. XLM-RoBERTa is multilingual version build on original RoBERTa model [25]. It is designed to work on different languages efficiently. The model was trained with large dataset in an unsupervised fashion. It learns the relationships and patterns between languages. Huggingface platform provides a language model [26] fine tuned over XLM-RoBERTa that provides detecting following languages: Arabic, Bulgarian, German, modern Greek, English, Spanish, French, Hindi, Italian, Japanese, Dutch, Polish, Portuguese, Russian, Swahili, Thai, Turkish, Urdu, Vietnamese and Chinese.

By using "xlm-roberta-base-language-detection" model we labeled the user comments and observe the language distribution in the comments. Table 3.30 shows the percentage ratios of languages used in comments. 94.8% of the comments are in English and Swahili follows with 1.9% showing that high majority of the comments are in English.

Table 3.30: Comment Language Distributions

Predicted Language	Comment Count	Comment Percentage
English	161688	94.885 %
Swahili	3370	1.978 %
Hindi	1745	1.024 %
Urdu	894	0.525 %
Italian	540	0.317%
Portuguese	424	0.249%
Spanish	351	0.206%
German	304	0.178%
Dutch	252	0.148%
Turkish	229	0.134%
Arabic	185	0.109%
Polish	141	0.083%
Russian	117	0.069%
French	76	0.045%
Vietnamese	40	0.023%
Bulgarian	33	0.019%
Greek	12	0.007%
Thai	3	0.002%

3.2.7 Data Pre-Processing

Prior to conducting hypothesis testing for the research questions, we pre-processed the data to prepare it for statistical analysis. The pre-processing for the video ranking based research questions and comment based questions differs in average comment sentiment score calculation. Table 3.31 lists the set of symbols utilized in this section to illustrate the pre-processing methods to make the data suitable for hypothesis testing.

Table 3.31: Symbol Definitions

Symbol	Definition
K	Video ranking set, K = {3, 5, 10, 20}
T	Comment ranking set, T = {10, 25, 50, 75}
V	Set of videos
C_i	Set of comments in video i, $i \in V$
$C_i@n$	Set of top n comments in video i, $i \in V$
S_{ij}	Positive sentiment score of video i, and comment j, $i \in V, j \in C_i$
SV_i	Positive sentiment score of video i, $i \in V$
$SV@n$	The set of positive sentiment scores for top n videos, $n \in K$
$SVC_i@n$	Positive sentiment score of a video i for top n comments, $i \in V, n \in T$
$SVC@n$	The set of positive sentiment scores considering top n comments for each video, $n \in T$

Pre-processing for Video Ranking Based Experiments

The initial step in the pre-processing of video ranking based experiments is aggregating the comment positive sentiment scores based on query video pairs. For each video, top 500 comments are taken into consideration. For each query video pair the average of comment positive sentiment scores are calculated. Throughout this thesis the term "average video sentiment score" will be used to refer to the calculated mean comment positive sentiment score associated with a video and it is denoted by SV_{ij}. Equation 3.1 shows the mathematical notation of how average video sentiment score is calculated. The detailed explanations of the symbols used in the equations are in Table 3.31.

$$SV_i = \frac{1}{|C_i|} \sum_{j \in Ci} S_{ij} \qquad i \in V \tag{3.1}$$

For video ranking based experiments we specifically focused on the videos ranked in the top 3, top 5, top 10, and top 20 positions separately. Top n refers to the videos that are ranked in the first n ranks in the search result for that query. In the

experimental design we use this different ranking groups in order to compare the comment sentiment patterns of videos placed in different positions in the YouTube search results. The aim of having different groups is to understand the relationship between video's ranking and its comment sentiment. For each ranking group the set $SV@n, n \in K$ is considered in the experimental design. $SV@n$ represents the set of average positive scores for each video ranked in top n.

Some of the research questions involve analysis on the effect of gender of the video narrator on the comment sentiments. For those experiments, only the set of scores for a specific gender is considered. Three sets of average sentiment scores are calculated for perceived genders male, female and other. Another variable in the research questions is educational subject: STEM and NON-STEM. For the experiments where the query field is considered as a variable, we calculated two different scores for both of the subjects.

Pre-processing for Comment Ranking Based Experiments

The pre-processing part for comment ranking based research questions is mostly similar to the video ranking based experiments explained in Section 3.2.7. The main difference is that the objective of video ranking based experiments is to investigate the effect of video rankings on comment sentiments, whereas the aim of comment ranking based experiments is to find the relationship between comment's ranking and its sentiment.

Initial step is to take average over each query video pair. Differently from video ranking based pre-processing, we include the comment's ranking position while calculating the average video sentiment score. In the previous part, mean video sentiment was calculated by taking the average sentiment score of all 500 comments.

In this part, instead of averaging all of the comments, the mean sentiment score is calculated for different ranking group subsets. The comment ranking subsets are top 10, 25, 50 and 75. Top n represent the highest ranked n comments of the video. Detailed explanations of comment sorting options and how the data is collected can be found in Data Collection Section 3.2.2. Video's positive sentiment score ($SVC_i@n$) is calculated for each of the ranking subset. Detailed calculation of how

sentiment score of comments are aggregated over videos is shown in Equation 3.2. Each query video pair has four top n comment average sentiment score. These ranking group scores will be used in research questions where the objective is to identify the relationship between comment's rank and its sentiment.

$$SV_i@n = \frac{1}{|C_i@n|} \sum_{j \in Ci@n} S_{ij} \qquad i \in V_i, n \in T \tag{3.2}$$

The set of average sentiment scores for top n comments are represented as $SVC@n$ and these set of sentiment scores are used in the comment ranking based experiments.

Similar to video ranking based research questions, comment ranking experiments have experimental setups that takes gender into account as well. In those experiments, sentiment scores of queries are calculated for both male, female and other labeled videos. Similarly, average sentiment scores are calculated for educational subject related experiments separately as well.

3.2.8 Experiment Methodology

The objective of this thesis is to understand how the comment sentiments are affected by different perceived genders of the video narrators and educational subjects: STEM & NON-STEM. We aimed to observe the relations by asking some research questions explained in Section 3.1.2. The research questions are formulated to address whether there are differences in comment sentiments between the two groups or not.

In summary, the Central Limit Theorem provides a basis for understanding the behavior of sample means, allowing for powerful inferential statistics like the t-test to be applied in a wide range of scenarios, even when working with small sample sizes or non-normally distributed data.

Our experimental design requires to use independent two-sample t-test method to test the hypothesis. According to central limit theorem, samples from a large enough sample will approximate to normal distribution [27] enabling us to apply t-tests on our data. Additionally, equal sample sizes and variances were not ensured in our

samples. According to [28], authors state that when sample sizes are not big enough the test's sensitivity to the equal variance assumption increases and becomes less robust. It suggests that instead of using Student's t-test, using Welch's t-test gives more reliable results. Therefore, we used Welch's t-test instead of Student's t-test.

There are certain biases in our dataset as mentioned in Section 3.2.6. For example, there are more videos with perceived gender male, the duration of male videos is longer than females, male videos receive longer comments than females, etc. In order to avoid these biases to affect our tests, we used bootstrapping technique. Bootstrapping provides more robust confidence intervals in hypothesis tests when the variance is not equal and sample size is smaller [29]. Therefore , we used Welch's t test with bootstrapping.

Every research question is designed to compare two different sample means. For each of the defined research questions we conducted hypothesis testing to conclude if our hypothesis is statistically significant or not. Equation 3.3 represents the definition of null and alternative hypothesis for all research questions.

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned} \tag{3.3}$$

These experiments are constructed in two groups. First part of the research questions focuses on how comment sentiment behavior differs for different video ranking groups. To test those hypotheses, the samples are formed by considering the video ranking information along with perceived gender of the video narrator and video subject. The second part focuses on how these comment sentiment behaviors differ in comment sorting.

The RoBERTa sentiment prediction model explained in Section 3.2.4 is used in this thesis to predict the sentiment of user generated comments. The model output consists of three components: negative sentiment score, neutral sentiment score and positive sentiment score. In general, the comment is assigned the label that has the maximum score: negative, neutral or positive. However, in this thesis instead of labeling the comments, we used class probabilities of the sentiments. This approach allows us to capture the uncertainty in the sentiment classification and provides

more comprehensive understanding of the sentiment distribution in the data. In order to use the continuous model scores we applied softmax activation to the score set. This is because the output range of scores may differ between samples. In order to conduct an analysis based on these scores, they need to be scaled. Therefore scores are then converted into class probabilities by the softmax activation function. Softmax function gives outputs for each predicted class between 0 and 1, and the scores are interpreted as class probabilities [30]. As the model class predictions are complementary, they sum up to one, to test the comment sentiments we selected positive sentiment score output of the NLP model after applying softmax. This score is referred as "positive sentiment score" throughout this thesis.

The tested parameters are same for all experimental setups, positive comment sentiment score. In other words, μ_1 and μ_2 corresponds to the average positive sentiments of a given sample 1 and 2. The objective of the experiments is to test if two groups' mean positive sentiment score is equal under given confidence interval. For all of the research questions the hypothesis testing is done under 95% confidence level.

This means that we are 95% confident in the validity of the test results and there is 5% chance that the conclusions are due to the noise. By defining the confidence level at 95%, we aim to ensure reliable analysis with strong evidence. Additionally, using Welch's t-test with bootstrapping method gives further credibility to our findings that results more robust statistical conclusions. The combination of Welch's t-test along with bootstrapping contributes to achieve more accurate analysis and this setup reduces the risk of drawing wrong conclusions caused by variance.

Chapter 4

Evaluation

In this Chapter, previously introduced research questions and the experimental setup will be explained in detail. The experiments are divided into two sections: video ranking and comment ranking based. The objective of the video ranking based experiments is to investigate the relation between the search result ranking of a video and its comment sentiment behavior, while comment ranking based studies are focused on the association between video comment's rank and its sentiment. Besides using comment positive sentiment score and ranking data as variables, we introduced two other variables: gender of the video narrator and subject of search query STEM & NON-STEM. Hypothesis testing is conducted for each research question to identify the behavior differences between mentioned variables.

4.1 Experiments

Conducted experiments can be examined in two parts: video ranking based and comment ranking based setups. The objective of video ranking based experiments is to investigate the behavior of comment sentiments in YouTube videos by analyzing the impact of different video rankings within query search results. The aim of comment ranking based experiments is to identify the relation between comment's ranking and its sentiment.

For each experiment the mean positive sentiment score is used as the dependent variable that is being tested. Objective of each hypothesis test is to identify the

effect of independent variables on mean positive sentiment score of the videos.

4.1.1 Video Ranking Based Experiments

In this thesis, we designed different experiment setups to investigate the sentiment behavior of educational videos with respect to the query field, video ranking and gender of the narrator. In each experiment one variable changed and others are kept same in order to understand the effect of independent variables. Table 4.1 shows how different set of variables are tried in each research question to identify the effect of query field, ranking and gender to the comment sentiments of the videos.

Table 4.1: Research questions and experimental setup

RQ No	**Query Field**	**Video Ranking**	**Gender**
1	constant	varying	-
2	varying	constant	-
3	constant	constant	varying
4	varying	constant	constant

RQ1: Do the positive sentiment scores of the videos vary across different rankings of a search result?

The null hypothesis in this question is that the mean sentiment scores of videos retrieved through YouTube search results are unequal when taking into account videos with different rankings. In the test setup query subjects of the videos were kept same and different rank groups are compared. Each query filed STEM and NON-STEM are statistically tested within their field and the effect of the mean positive sentiment score between different rank groups are observed.

When different ranked videos' mean positive sentiment scores are compared, there were no evidence that there is difference between different video rankings as seen in Table 4.2. Although the mean values of upper ranking groups are higher than lower ones, there is not enough evidence to reject the null hypothesis that the different ranking levels have different mean comment sentiment scores for each educational subject.

Table 4.2: RQ1 Hypothesis Testing Results

Query Field	Experiment	Mean	Result
STEM	top3 vs. top5	(0.537, 0.529)	Fail To Reject
STEM	top5 vs. top10	(0.529, 0.528)	Fail To Reject
STEM	top10 vs. top20	(0.528, 0.516)	Fail To Reject
NON-STEM	top3 vs. top5	(0.543, 0.539)	Fail To Reject
NON-STEM	top5 vs. top10	(0.539, 0.528)	Fail To Reject
NON-STEM	top10 vs. top20	(0.528, 0.535)	Fail To Reject

In Table 4.3 we excluded the gender label "other" from the experimental samples and replicated the hypothesis testing. Our aim was to observe the difference in test results to determine if the exclusion of the "other" category affects the outcomes significantly. However, the analysis shows that the results remained unchanged, leading to the conclusion that video rankings do not have a significant effect on comment sentiments.

Table 4.3: RQ1 Hypothesis Testing Results Without Gender Class "other"

Query Field	Experiment	Mean	Result
STEM	top3 vs. top5	(0.542, 0.528)	Fail To Reject
STEM	top5 vs. top10	(0.528, 0.534)	Fail To Reject
STEM	top10 vs. top20	(0.534, 0.518)	Fail To Reject
NON-STEM	top3 vs. top5	(0.544, 0.534)	Fail To Reject
NON-STEM	top5 vs. top10	(0.534, 0.538)	Fail To Reject
NON-STEM	top10 vs. top20	(0.538, 0.541)	Fail To Reject

RQ2: Do the positive sentiment scores of videos within the same ranking differ when comparing STEM and NON-STEM queries?

In this experiment different from RQ1, rankings of the videos were kept the same and query fields were changed. For each ranking group top 3, 5, 10 and 20, the mean positive sentiment scores of the videos were compared. Table 4.5 shows hypothesis testing results for each rank group. The results show that we fail to reject all null

hypotheses and conclude that for each ranking group the average positive comment sentiments between STEM and NON-STEM query fields are not statistically different. This concludes that there is no overall sentiment difference between educational subjects is YouTube.

Table 4.4: RQ2 Hypothesis Testing Results

Ranking	**Experiment**	**Mean**	**Result**
top3	STEM vs. NON-STEM	(0.537, 0.543)	Fail To Reject
top5	STEM vs. NON-STEM	(0.529, 0.539)	Fail To Reject
top10	STEM vs. NON-STEM	(0.528, 0.528)	Fail To Reject
top20	STEM vs. NON-STEM	(0.516, 0.535)	Fail To Reject

To investigate the potential bias introduced by including the ”other” category in our analysis, we replicated the test without including ”other.” Nevertheless, we obtained the same results leading us to conclude that the inclusion of the ”other” category in our analysis did not significantly affect the overall outcomes.

Table 4.5: RQ2 Hypothesis Testing Results Without Gender Class ”other”

Ranking	**Experiment**	**Mean**	**Result**
top3	STEM vs. NON-STEM	(0.542, 0.544)	Fail To Reject
top5	STEM vs. NON-STEM	(0.528, 0.534)	Fail To Reject
top10	STEM vs. NON-STEM	(0.534, 0.538)	Fail To Reject
top20	STEM vs. NON-STEM	(0.518, 0.541)	Fail To Reject

RQ3: Do the sentiment scores of the videos within the same ranking and subject vary across perceived gender of the videos?

This question explores the effect of gender to comment sentiment score. In this research question, for each query field STEM & NON-STEM and ranking group (top 3, 5, 10, 20) the mean positive sentiment scores of gender pairs are compared. Gender pairs are: female & male, male & other and female & other. For each experiment we got insignificant test results. Therefore we conclude that the sentiments between

different genders do not change within the same query field and ranking. The detailed hypothesis testing results can be found in Table 4.6.

Table 4.6: RQ3 Hypothesis Testing Results

Query Field	**Ranking**	**Experiment**	**Mean**	**Result**
STEM	top3	Female vs. Male	(0.559, 0.538)	Fail To Reject
STEM	top5	Female vs. Male	(0.528, 0.528)	Fail To Reject
STEM	top10	Female vs. Male	(0.552, 0.527)	Fail To Reject
STEM	top20	Female vs. Male	(0.535, 0.5109)	Fail To Reject
STEM	top3	Female vs. Other	(0.559, 0.518)	Fail To Reject
STEM	top5	Female vs. Other	(0.528, 0.534)	Fail To Reject
STEM	top10	Female vs. Other	(0.552, 0.507)	Fail To Reject
STEM	top20	Female vs. Other	(0.535, 0.511)	Fail To Reject
STEM	top3	Male vs. Other	(0.538, 0.518)	Fail To Reject
STEM	top5	Male vs. Other	(0.528, 0.534)	Fail To Reject
STEM	top10	Male vs. Other	(0.527, 0.507)	Fail To Reject
STEM	top20	Male vs. Other	(0.519, 0.511)	Fail To Reject
NON-STEM	top3	Female vs. Male	(0.513, 0.559)	Fail To Reject
NON-STEM	top5	Female vs. Male	(0.507, 0.547)	Fail To Reject
NON-STEM	top10	Female vs. Male	(0.521, 0.546)	Fail To Reject
NON-STEM	top20	Female vs. Male	(0.538, 0.543)	Fail To Reject
NON-STEM	top3	Female vs. Other	(0.513, 0.539)	Fail To Reject
NON-STEM	top5	Female vs. Other	(0.507, 0.554)	Fail To Reject
NON-STEM	top10	Female vs. Other	(0.521, 0.490)	Fail To Reject
NON-STEM	top20	Female vs. Other	(0.538, 0.51)	Fail To Reject
NON-STEM	top3	Male vs. Other	(0.559, 0.539)	Fail To Reject
NON-STEM	top5	Male vs. Other	(0.547, 0.554)	Fail To Reject
NON-STEM	top10	Male vs. Other	(0.546, 0.490)	Fail To Reject
NON-STEM	top20	Male vs. Other	(0.543, 0.510)	Fail To Reject

RQ4: Do the positive sentiment scores of the videos within the same ranking and gender vary across different subjects?

In this experiment the effect of query field on the comment positive sentiment score is observed while keeping the perceived gender of the narrator and ranking of the videos fixed. From the results gathered in Table 4.7 there is not enough evidence to conclude that the comment sentiment of videos with female and other genders differs between query fields.

However, for videos with male narrators NON-STEM related videos demonstrates more positive comment sentiments when compared to STEM videos for the videos ranked in top 20. This behavior is not the same for top 3, top 5 and top 10 videos. In other words, the comment positive sentiments for upper ranked videos with male narrators are not statistically different between query fields. Although mean values are different and NON-STEM is more positive than STEM in upper rankings as well, we cannot reject the null hypothesis on 95% confidence level.

Table 4.7: RQ4 Hypothesis Testing Results

Ranking	**Gender**	**Experiment**	**Mean**	**Result**
top3	Female	STEM vs. NON-STEM	(0.555, 0.513)	Fail To Reject
top5	Female	STEM vs. NON-STEM	(0.528, 0.507)	Fail To Reject
top10	Female	STEM vs. NON-STEM	(0.552, 0.521)	Fail To Reject
top20	Female	STEM vs. NON-STEM	(0.535, 0.538)	Fail To Reject
top3	Male	STEM vs. NON-STEM	(0.538, 0.559)	Fail To Reject
top5	Male	STEM vs. NON-STEM	(0.528, 0.547)	Fail To Reject
top10	Male	STEM vs. NON-STEM	(0.527, 0.546)	Fail To Reject
top20	Male	STEM vs. NON-STEM	(0.510, 0.543)	NON-STEM > STEM
top3	Other	STEM vs. NON-STEM	(0.518, 0.539)	Fail To Reject
top5	Other	STEM vs. NON-STEM	(0.534, 0.554)	Fail To Reject
top10	Other	STEM vs. NON-STEM	(0.507, 0.490)	Fail To Reject
top20	Other	STEM vs. NON-STEM	(0.511, 0.510)	Fail To Reject

4.1.2 Comment Ranking Based Experiments

In this experimental setup, we aim to identify the relation between comments' positive sentiment scores and their ranking. In each research question, we changed a

variable by keeping others constant to observe the changed variable's interaction with videos comment sentiment. Tested variables are similar with variables explained in Section 4.1.1. However, in this setup the behavior of comment sentiment scores are investigated based on comment rankings instead of video search result rankings. Table 4.8 shows how variables change for each research question.

Table 4.8: Research questions and experimental setup

RQ No	**Query Field**	**Comment Ranking**	**Gender**
1	-	varying	-
2	constant	varying	-
3	-	varying	constant
4	constant	varying	constant

RQ1: Is there a relation between the positiveness of the comments in YouTube videos and their rankings?

In this research question we aim to find the interaction between the comment positive sentiment score and the ranking of comments. To test the hypothesis that the sentiment changes withing different rankings we designed the test between consecutive ranking groups and compared the sentiments between top 10 & top 25, top 25 & top 50 and top 50 vs top 75. In this experiment we didn't include query field and gender as a variable but study the overall relation between ranking and comment sentiment.

From the hypothesis test results in Table 4.10, we infer that the sentiment is significantly more positive in higher ranks and this behavior is consistent within all experiments conducted in this research question.

Table 4.9: RQ1 Hypothesis Testing Results

Experiment	**Mean**	**Result**
top10 vs. top25	(0.655, 0.628)	top10 > top25
top25 vs. top50	(0.628, 0.604)	top25 > top50
top50 vs. top75	(0.604, 0.589)	top50 > top75

RQ2: Is the comment sentiment more positive in higher ranks for each query field STEM and NON-STEM?

Previous research question shows that the comments in higher rankings have more positive sentiments. That experimental setup was independent of the query field from which the video is retrieved from and the narrator's gender. To further identify the source of this behavior we added query field as an independent variable. Same experiment is replicated with query field aspect as well and the interaction between comment ranking and positive comment sentiment score is identified for STEM & NON-STEM queries separately.

The results in Table 4.10 show that for both query fields STEM and NON-STEM higher rank groups' comment sentiments are more positive than below ranking groups. Top 10 comment's positive sentiment score is higher than top 25's, top 25 comment's is higher than top 50's and top 50 comment's is higher than top 75's for both STEM and NON-STEM queries.

Table 4.10: RQ2 Hypothesis Testing Results

Query Field	**Experiment**	**Mean**	**Result**
STEM	top10 vs. top25	(0.667, 0.641)	top10 > top25
STEM	top25 vs. top50	(0.641, 0.613)	top25 > top50
STEM	top50 vs. top75	(0.613, 0.596)	top50 > top75
NON-STEM	top10 vs. top25	(0.642, 0.614)	top10 > top25
NON-STEM	top25 vs. top50	(0.614, 0.594)	top25 > top50
NON-STEM	top50 vs. top75	(0.594, 0.582)	top50 > top75

RQ3: Is the comment sentiment more positive in higher ranks for each gender male, female and other?

For RQ1, we observed that the comment sentiment scores are significantly more positive for higher comment rankings. In RQ2, query field parameter was included to experimental design and we found that this response does not change with the query field. In this research question we added the perceived gender of the video narrator to the hypothesis test in RQ1. For each gender male, female & other and

comment ranking group we compared the mean comment sentiment scores. For each gender, the experiment results show that higher ranks have significantly higher comment positivity. Adding gender aspect to the experiment does not change the overall relationship between comment ranking and its sentiment. Table 4.11 shows the mean values for the positive comments for each experiment.

Table 4.11: RQ3 Hypothesis Testing Results

Gender	**Experiment**	**Mean**	**Result**
Female	top10 vs. top25	(0.673, 0.638)	top10 > top25
Female	top25 vs. top50	(0.638, 0.609)	top25 > top50
Female	top50 vs. top75	(0.609, 0.595)	top50 > top75
Male	top10 vs. top25	(0.655, 0.635)	top10 > top25
Male	top25 vs. top50	(0.635, 0.615)	top25 > top50
Male	top50 vs. top75	(0.615, 0.598)	top50 > top75
Other	top10 vs. top25	(0.633, 0.597)	top10 > top25
Other	top25 vs. top50	(0.597, 0.570)	top25 > top50
Other	top50 vs. top75	(0.570, 0.558)	top50 > top75

RQ4: Is the comment sentiment more positive in higher ranks for each gender and query field pairs individually?

In the previous two research questions, we tested whether query field and gender of the video narrator differentiates the overall behavior of positive comments being ranked in higher positions. The results showed that both gender and query field variables do not change that conclusion. In this experiment we added those parameters together and conducted an hypothesis test for each gender and query field pair. The hypothesis testing results for each pair in Table 4.12 show that adding gender and query field variables does not change the sentiment behavior between different ranking groups.

Considering all the research questions above in Subsection 4.1.2 we can conclude that the positive sentiment scores of the comments and its respective ranking is correlated. Regardless of the gender of the narrator and the query field, positive comments are ranked in higher positions.

Table 4.12: RQ4 Hypothesis Testing Results

Query Field	**Gender**	**Experiment**	**Mean**	**Result**
STEM	Female	top10 vs. top25	(0.692, 0.662)	top10 > top25
STEM	Female	top25 vs. top50	(0.662, 0.631)	top25 > top50
STEM	Female	top50 vs. top75	(0.631, 0.616)	top50 > top75
NON-STEM	Female	top10 vs. top25	(0.655, 0.615)	top10 > top25
NON-STEM	Female	top25 vs. top50	(0.615, 0.587)	top25 > top50
NON-STEM	Female	top50 vs. top75	(0.587, 0.575)	top50 > top75
STEM	Male	top10 vs. top25	(0.662, 0.643)	top10 > top25
STEM	Male	top25 vs. top50	(0.643, 0.616)	top25 > top50
STEM	Male	top50 vs. top75	(0.616, 0.596)	top50 > top75
NON-STEM	Male	top10 vs. top25	(0.646, 0.625)	top10 > top25
NON-STEM	Male	top25 vs. top50	(0.625, 0.613)	top25 > top50
NON-STEM	Male	top50 vs. top75	(0.613, 0.601)	top50 > top75
STEM	Other	top10 vs. top25	(0.651, 0.609)	top10 > top25
STEM	Other	top25 vs. top50	(0.609, 0.584)	top25 > top50
STEM	Other	top50 vs. top75	(0.584, 0.572)	top50 > top75
NON-STEM	Other	top10 vs. top25	(0.614, 0.585)	top10 > top25
NON-STEM	Other	top25 vs. top50	(0.585, 0.556)	top25 > top50
NON-STEM	Other	top50 vs. top75	(0.556, 0.544)	top50 > top75

Chapter 5

Conclusion and Future Work

In this thesis we focused on the comment sentiment behavior in YouTube videos in educational context. We collected YouTube comments for both STEM and NON-STEM related videos and calculated their sentiment scores. In order to understand the behavioral change in comment sentiments for different variables we introduced research questions. The main variables we focus on are as follows: perceived gender of the video narrator (male, female or other), the subject of the video (STEM or NON-STEM). While observing the effects of these variables on comment sentiments, we also take video rankings and comment rankings into consideration. The aim of including ranking information in this research is to understand the behavioral change in comment sentiment patterns across different levels of user engagement and visibility. This multi-dimensional approach enables us to comprehend the patterns in a more holistic way.

We conducted hypothesis testing for each of the research questions in order to obtain reliable and statistically significant results. By applying statistical tests we examined the statistical significance of the observed differences in sentiment scores based on the introduced variables. We examined the research questions in two parts: video ranking based and comment ranking based experiments.

For video ranking based experiments we conclude that within same video subject (STEM or NON-STEM) video's ranking does not affect the overall sentiment pattern **(RQ1)**. Within the same rank cutoff group there is no significant difference between different query fields STEM and NON-STEM **(RQ2)**. When we introduced the

perceived gender of a video narrator into the experimental setup, the tests did not indicate a significant difference in average positive comment sentiments **(RQ3)**. We also tested the effect of query fields on comment sentiments within the same gender. The results showed that for male narrators NON-STEM related videos receive more positive comments compared to STEM **(RQ4)**. However, for female and other labels we did no observe any statistical difference **(RQ4)**.

In experiments based on comment ranking, we noted that as the comment's rank on the platform increases, its sentiment tends to become more positive **(RQ1)**. When we examine the results in more granular level, we observed that this behavior is same for different video subjects STEM & NON-STEM and perceived genders of video narrators male, female & other and their pairs **(RQ2, RQ3 & RQ4)**.

5.1 Limitations and Future Work

The dataset we have consists of 2000 videos in total and the genders male and female are not equally represented in the data. Additionally, this representativeness is less for female narrators in higher ranks. Therefore, the video counts of female narrators in higher ranks are considerably lower. In order not to drew any noisy conclusion because of this problem we ensured more reliable and robust hypothesis testing methods. However, due to the high variance in the data may not be seeing the results we should see. For future studies, we may address this issue by collecting more data that displays high variability.

Our data period covers COVID-19 period as well. In our study we did not use time as a variable. We observed from the descriptive analysis that YouTube video uploading pattern significantly changes during lockdown period. We observed this behavioral shift in content creators, but this period may have also affected viewers and their behavior. For future work, the comment sentiment behavior could be investigated using a time series approach to comprehend viewer patterns.

In this thesis, we only observed the comment sentiments in the context of negative, neutral or positive. However, sentiment analysis methods can provide more labels such as anger, hate, offensive, emotion, etc. For feature studies, we may consider

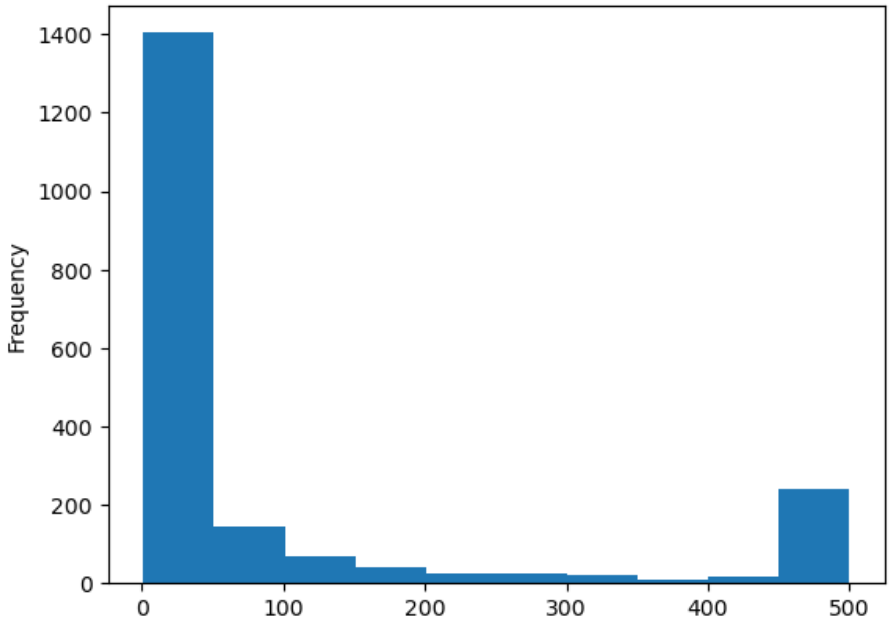
running experiments on those classification tasks as well. Furthermore, we used positive comment sentiment scores in our experiments as a sentiment indicator. However, for feature studies we may also consider scores from other sentiment classes to offer a more comprehensive sentiment analysis.

We used Twitter RoBERTa model to predict comment sentiments. Although this model performs well in labeled validation set, for further studies we can consider a model fine tuned on YouTube comment data. On the other hand, this NLP model is not multilingual, only trained on English tweets. For future studies we can use XLM's to achieve better performances on predicting cross-lingual comments.

Appendix A

Comment Count Distribution

Figure A.1: Video Comment Count Histogram



Bibliography

[1] D. Pattier, "Science on youtube: Successful edutubers," vol. 10, pp. 1–15, 02 2021.

[2] R. Rahmatika, M. Yusuf, and L. Agung, "The effectiveness of youtube as an online learning media," vol. 5, p. 152–158, Apr. 2021. [Online]. Available: https://ejournal.undiksha.ac.id/index.php/JET/article/view/33628

[3] G. Gezici, "Bias in search: Evaluating search results through rank and relevance based measures," Ph.D. dissertation, Sabancı University, 2022.

[4] S. Fouad and E. Alkooheji, "Sentiment analysis for women in stem using twitter and transfer learning models," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, 2023, pp. 227–234.

[5] L. Wotanis and L. McMillan, "Performing gender on youtube: How jenna marbles negotiates a hostile online environment," *Feminist Media Studies*, vol. 14, no. 6, pp. 912–928, 2014.

[6] N. Döring and M. Mohseni, "Gendered hate speech in youtube and younow comments: Results of two content analyses," *Studies in Communication and Media*, vol. 9, pp. 62–88, 03 2020.

[7] "Youtube - preventing bias," https://www.youtube.com/intl/ALL_in/howyoutubeworks/our-commitments/preventing-bias/.

[8] I. Amarasekara and W. J. Grant, "Exploring the youtube science communication gender gap: A sentiment analysis," *Public Understanding of Science*, vol. 28, no. 1, pp. 68–84, 2019.

[9] E. Smith, "Women into science and engineering? gendered participation in

higher education stem subjects," *British Educational Research Journal*, vol. 37, no. 6, pp. 993–1014, 2011. [Online]. Available: http://www.jstor.org/stable/23077020

[10] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, and L. Plaza, "Automatic classification of sexism in social networks: An empirical study on twitter data," *IEEE Access*, vol. 8, pp. 219 563–219 576, 2020.

[11] "The Uni Guide," https://www.theuniguide.co.uk/.

[12] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[13] F. Salmon and F. Vallet, "An effortless way to create large-scale datasets for famous speakers," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 348–352. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/32_Paper.pdf

[14] "Youtube data api reference," https://developers.google.com/youtube/v3/docs.

[15] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, pp. 5731–5780, 2022. [Online]. Available: https://doi.org/10.1007/s10462-022-10144-1

[16] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2022. [Online]. Available: https://doi.org/10.1007%2Fs11042-022-13428-4

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[18] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman,

"GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[19] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: http://aclweb.org/anthology/N18-1101

[20] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," 2018.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[22] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.148

[23] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 502–518. [Online]. Available: https://aclanthology.org/S17-2088

[24] R. Pokharel and D. Bhatta, "Classifying youtube comments based on sentiment and type of sentence," *arXiv preprint arXiv:2111.01908*, 2021.

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán,

E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020.

[26] papluca, "xlm-roberta-base-language-detection," Hugging Face Model Card, 2023. [Online]. Available: https://huggingface.co/papluca/xlm-roberta-base-language-detection

[27] S.-G. Kwak and J. H. Kim, "Central limit theorem: The cornerstone of modern statistics," *Korean J Anesthesiol*, vol. 70, no. 2, pp. 144–156, Apr 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5370305/

[28] M. Delacre, D. Lakens, and C. Leys, "Why psychologists should by default use welch's t-test instead of student's t-test," *International Review of Social Psychology*, vol. 30, no. 1, pp. 92–101, 2017.

[29] T. Hesterberg, "Bootstrap," *WIREs Computational Statistics*, vol. 3, no. 6, pp. 497–526, 2011. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.182

[30] M. A. Mercioni and S. Holban, "The most used activation functions: Classic versus current," in *2020 International Conference on Development and Application Systems (DAS)*, 2020, pp. 141–145.