MEASURING BIAS AMONG TEXT DATA USING NLP METHODS

by
Egemen Uğur Dalgıç

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
June 2024

ABSTRACT

MEASURING BIAS AMONG TEXT DATA USING NLP METHODS

EGEMEN UĞUR DALGIÇ

DATA SCIENCE MSC. THESIS, JULY 2024

Thesis Advisor: Yücel Saygın

Keywords: sentiment analysis, gender bias, educational videos

YouTube became one of the main digital education mediums during the pandemic. Previously, it was found that there exists a positive bias towards males in face-to-face education. For instance, in one study, it was found that the attitude of the teachers while grading changed when they received the exam papers together with the student names. Teachers graded male students more generously compared to their females i.e. exhibited positive discrimination. In another study, the responsiveness of the professors to their emails was measured. The researchers concluded that the professors are the most responsive when the email belongs to a white male student. Gender bias found in face-to-face education is also reflected in digital education platforms. In a study, the experimenters created a set of videos for science, technology, engineering mathematics (STEM), and the remaining fields of education (non-STEM). They gauged the gender bias and suggested that there is a bias towards males in both STEM and non-STEM videos although the the degree between the two differs from each other. The goal of this study is to explore the reason behind this situation as well as understand the impact of the COVID-19 pandemic on the video and comment characteristics shared on the digital platforms. Our data includes 19867 educational video details collected from YouTube as well as their top-ranked comments. These videos were made by different narrators from January 2007 to March 2021, and they were grouped based on STEM, and non-STEM queries. We focus on finding important evidence related to gender bias by working on the differences in the video and comment details such as the number of likes or views they get, the polarity of the comments, and the rank of the most common words. In this regard, we used a large variety of data preprocessing, statistical analyses, and sentiment analysis techniques.

ÖZET

DOĞAL DIL İŞLEME METOTLARINI KULLANARAK YAZILI VERI ÜZERINDEKI YANLILIĞI ÖLÇME

EGEMEN UĞUR DALGIÇ

VERI BILIMI YÜKSEK LİSANS TEZİ, TEMMUZ 2024

Tez Danışmanı: Prof. Dr. Yücel SAYGIN

Anahtar Kelimeler: duygu analizi, cinsiyet önyargısı, eğitim videoları

YouTube, pandemi sırasında ana dijital eğitim ortamlarından biri haline geldi. Daha önce, yüz yüze eğitimde erkeklere karşı pozitif önyargının var olduğu bulunmuştu. Örneğin, bir çalışmada, öğretmenlerin sınav kağıtlarına öğrenci isimleriyle aynı anda eriştiklerinde notlandırma tutumlarının değiştiği saptandı. Öğretmenler, pozitif ayrımcılık sergileyerek erkek öğrencileri kadın öğrencilere göre daha cömert bir şekilde notlandırdılar. Başka bir çalışmada, profesörlerin e-postalarına karşı duyarlılığı ölçüldü. Araştırmacılar, profesörlerin beyaz erkekler öğrencilerden gelen e-postalara karşı en fazla duyarlı oldukları sonucuna vardı. Yüz yüze eğitimde bulunan cinsiyet önyargısı, dijital eğitim platformlarına da yansıtılmaktadır. Bir çalışmada, deneyciler bilim, teknoloji, mühendislik, matematik (STEM) ve eğitimin geri kalan alanları (STEM dışı) için bir dizi video topladılar. Cinsiyet önyargısını ölçtüler ve STEM ve STEM dışı videolarda, her iki durumda da erkeklere yönelik bir önyargı olduğunu, ancak önyargı düzeyinin iki grup arasında birbirinden farklılık gösterdiğini öne sürdüler. Bu çalışmanın amacı, bu durumun arkasındaki nedeni keşfetmek ayrıca COVID-19 pandemisinin dijital platformlarda paylaşılan video ve yorumların karakteristikleri üzerindeki etkisini anlamaktır. Verilerimiz, YouTube'dan toplanan 19867 eğitim videosuna ait detayları ve bu videoların en üst sıradaki yorumlarını içermektedir. Toparlanan videolar, Ocak 2007'den Mart 2021'e kadar farklı video anlatıcıları tarafından yapılmış ve STEM ve STEM dışı sorgulara göre gruplandırılmıştır. Topladığımız video ve yorum detaylarındaki farklılıkları inceleyerek cinsiyet önyargısına dair kanıtlar bulmaya çalıştık. Bu bağlamda, geniş bir veri ön işleme, istatistiksel analiz ve duygu analizi teknikleri kullandık.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor, Prof. Dr. Yücel Saygın, for his invaluable guidance, support, and mentorship throughout this academic journey. His open-heartedness and tolerance created a learning environment that was both challenging and supportive, allowing me to grow without feeling overwhelmed by pressure. Working with him was a privilege for me.

I am also grateful to Asst. Prof. Sinan Yıldırım and Assoc. Prof. Asta Zelenkauskaite for their contributions and guidance during my thesis process.

I would like to thank to my father Cengiz Dalgıç, my mother Sabiha Dalgıç, my sister Deniz Dalgıç Damar, my brother Koray Damar and my lovely nephew Çınar for their emotional support. Their support was really important for me to surpass this challenging period.

I also appreciate my friends Hasan Kemik, Mekan Myadrov, Semih Gülüm, Zeynep Işık, and Seçilay Kutal, who have been with me throughout my educational journey, providing unlimited support and encouragement.

Finally, I want to thank everyone who played a role in the completion of this thesis. Your support has been indispensable in making this achievement possible.

To my family

Contents

List of Figures

List of Tables

1 Introduction

People are born either men or women. Depending on their sex, not surprisingly, they differ from each other. But how much or on which aspects? One should understand that the difference we are trying to quantify is largely affected by our expectations. In other words, the distinct characteristics of the men and women that we are aware of are not only sourced by nature but also from our *opinions* [12]. Gender stereotype is the term that is widely used for describing the *accepted* sexual characteristics.

Gender stereotypes are one of the core concepts in day-to-day life and it is hard to change for a couple of reasons. First, our attention mechanism is largely affected by our conscious and unconscious beliefs. There are so many stimuli at any moment so our attention mechanism searches the information in a way that the perceived information is compatible with our prior beliefs. Secondly, when we have no prior knowledge about something, one of the first things we generally do is to take the opinions of others. Thirdly, people who behave differently from the stereotypical expectations are devalued by the community and lastly, as people have the stereotypical information in their minds, they tend to behave accordingly.

Our biased view of men and women tends to cause inequalities in daily life. Despite the efforts towards equality, the reality often falls short [27]. In other words, in general, men are one step ahead of women just because of their gender. Even in professional science, gender bias remains, though it is often unintentional. A charming example is the workforce composition in the UK, where only 23% are women, indicating a significant gender gap. Similarly, in institutions like the National Institutes of Health (NIH), women make up just 31% of the workforce. These examples highlight the existing problem where, despite having equivalent skills and qualifications, women are less likely to be hired compared to their male counterparts [23]. Furthermore, gender bias also affects the performance evaluation. As researchers from Harvard University claim, the quality and impact of work done by women in science are frequently undervalued when compared to that of men [29]. For instance, in the field of economics, women receive lower credit compared to men from the papers they publish.

Previously, it was found that similar inequalities towards women also appear both in face-to-face education and in digital educational platforms [15]. The related work and further information are provided in Chapter 2. This thesis project aims to contribute to the investigations related to gender bias by exploring the presence of gender bias on digital platforms, focusing specifically on educational content. Additionally, it examines the impact of COVID-19 on the video and comment characteristics. To achieve this, we analyzed various aspects of a large collection of educational YouTube videos. These aspects include video-related information such as the ranking of the videos, the number of videos received, and the view counts, as well as the details of the top-ranked comments like comment post date and comment like count. Further details of the features we used in our analyses are provided in the next section as well.

In our study we mostly focused on the differences between these dimensions:

- **Gender**: Since our topic is to find the presence of a gender bias, it is important to analyze the difference between the genders of the narrators. To denote the videos having female narrators, we used the word “Female” and for the videos having male narrators we used the word “Male”.

- **Query Type**: Previously, it was found that gender bias is more severe in science, technology, engineering, and mathematics (STEM) related videos compared to non-STEM videos [15]. Therefore we also wanted to include this dimension.

- **Time period**: The COVID-19 pandemic changed day-to-day life significantly. While working on gender bias, we also wanted to see whether the video and comment characteristics changed with this recent event. We named the time frame after the COVID-19 announcement date as “Postcovid” and the time frame before the COVID-19 announcement date as “Precovid”.

1.1 Dataset and Data Collection

For our study, we used the dataset provided in [15]. This dataset contains information about a selection of educational YouTube videos. It focuses on five different topics each in STEM and non-STEM queries, making up 10 different types of query sub-fields in total. For each query sub-field, there are approximately 2000 videos. The query sub-field and the respective video counts were summarized in Table 1.1.

Instead of the full name, some query sub-fields are denoted with their acronyms. The term *EngLangLit* stands for English Language Literature, *CS* abbreviates Computer Sciences, and *PublicRel* represents Public Relations.

Table 1.1: Query Sub-fields and Video Counts

query__field	**query__sub__field**	**video__count**
non-STEM	EngLangLit	2000
	Psychology	2000
	PublicRel	2000
	Sociology	2000
	Politics	1867
STEM	Biology	2000
	CS	2000
	Chemistry	2000
	Maths	2000
	Physics	2000

For each video, we have some extra information such as video post date, video like count, and video view count. The features available in the original dataset were not comprehensive enough to answer the research questions of this work therefore, we utilized YouTube API, a tool for gathering a variety of video information, to retrieve additional information. Table 1.2 shows the features about videos that we worked with during our analyses:

Table 1.2: The Features of Video Details

Feature	**Explanation**
query__field	Indicates whether the query falls under STEM or non-STEM categories.
query__text	The actual search query used on YouTube.
video__id	Unique identifier for each video on YouTube.
video__rank	The search rank of the video within the given topic.
audio__downloadable	A binary indicator showing whether the video's audio is downloadable.
sampling__speech	The degree of measure of speech present in the video.
language	The primary language spoken in the video.
sampling__biased__audio__male__ratio	The probability that the narrator of the video is male.
sampling__biased__audio__female__ratio	The probability that the narrator of the video is female.
view__count	The total number of views for each video.
like__count	The total number of likes for each video.
channel__id	The unique identifier for the channel that posted the video.
channel__creation__date	The date when the channel was created.
published__at	The date and time when the video was published.
tags	The tags or keywords of the video

As mentioned before, our analyses involve not only the details of educational videos but also the comments posted under them. In this regard, the features we gathered about comments were shown in Table 1.3:

Table 1.3: The Features of Video Comments

Feature	Explanation
comment	The comment posted under the video.
comment_author	The author of the comment.
comment_date	The date when the comment was posted.
comment_like_count	The number of likes the comment received.
comment_rank	The rank of the comment based on relevance or comment post date.

While retrieving comments, we prioritized the *Relevancy* option and therefore used the most relevant comments to the videos in our analyses. However in some analyses, to gauge the degree of relevancy of the recently posted comments, we also gathered the most recent comments on the videos and used those comments together with the most relevant ones.

There are a couple of considerations regarding the comment retrieval process. First, not all the videos got comments, second many videos disabled the comments, third some videos are forbidden to pull the comments and lastly, YouTube API does not allow a developer to pull all the comments. This situation led the comment dataset to be imbalanced. Table 1.4 shows the numerical representation of this issue:

Table 1.4: Comment Availability

Category	Number of Videos
Forbidden to pull comments	2550
0 comments/Disabled comments/Video is not available	5848
Duplicated	1689
Videos with available comments	9780
Sum Total	**19867**

Some videos also appear more than once in different video rankings. This caused duplicate occurrences. Table 1.5 shows an illustration for that:

Table 1.5: Duplicate Videos

video_id	query_field	query_sub_field	query_text	video_rank
-B-lFjzHXgU	STEM	Biology	Human physiology	42
-B-lFjzHXgU	STEM	Biology	Human physiology	57
YYSn4vZn1Sc	STEM	Biology	Human physiology	79
YYSn4vZn1Sc	STEM	Biology	Human physiology	108

1.2 Contributions

The goal of this thesis project is to analyze the presence of gender bias in the educational context as well as understand the effect of COVID-19 pandemic on video and comment characteristics. On this subject, we collected various features about a set of educational videos and their respective comments. After that, we formed our research questions and lastly, we tested our findings statistically. The main findings of our analyses can be summarized as follows:

- Videos posted during the COVID-19 period are significantly longer than the ones posted before. Not only an average video tend to be longer but also longer lecture videos are posted on the platform.
- The video release rates increase over time over the channels in general but we can't attribute this increase to the COVID-19 pandemic.
- Female narrators tend to get more likes per view compared to their male counterparts. Similarly, the videos posted after the COVID-19 announcement date get more likes per view compared to the videos posted beforehand. Also with a small difference, non-STEM videos outperform STEM videos in the same domain.
- The average length of the comments posted under the videos is significantly different between the periods.
- Emoji ranking similarities of the videos belonging to opposite dimensions are similar indicating neither of the dimensions have a meaningful impact on emoji ranking similarities.
- The narrator's gender has no meaningful impact on the comments section in terms of polarity and emotions.
- Title and keyword similarities vary significantly between dimensions.
- The comments that were recently posted under the videos having female narrators are more relevant compared to the ones having male narrators. Similarly, the comments released under non-STEM videos surpass the ones posted under STEM videos. Lastly, the comments posted under the videos belonging to the Postcovid period overcome the ones written under the videos associated with the Precovid period.

In Chapter 2 the related work that has been done so far was discussed. In Chapter 3 the preliminary information that was used in this thesis work was detailed. In Chapter 4 the problem was defined and the methodology of the experiments was shared. Chapter 5 includes the descriptive analysis and the test results. Finally, in Chapter 6, all the work that we have done so far was summarized and discussed about the potential improvement areas.

2
Related Work

In this section, we overview the existing literature on gender bias. Section 2.1 explains the previous work about gender bias in the educational context. Section 2.2 examines the findings of gender bias in the YouTube platform, and lastly Section 2.3 shows the outcomes of sentiment analysis in the online platforms related to women in STEM.

2.1 Gender Bias In Education

As briefly introduced in Chapter 1, gender bias is a common problem in various fields including education. Our behavioral tendency towards women is likely to be negative.

Previous research [26] indicates the existence of a gender bias in education favoring males. Globally, teachers tend to spend more time, energy, and attention on male students on average. For example, a study found that professors are the most responsive to white males compared to all categories of students [21]. Moreover, this bias extends beyond the classroom, influences social dynamics, and leads to some unfortunate misconceptions such as girls do not require as much education. Also since gender bias is sunk to social life, its effect is reflected in various educational materials, including lesson plans, textbooks, and language usage.

Due to the COVID-19 pandemic, there was a significant shift towards online education [24]. YouTube, in particular, gained popularity as an educational medium, largely due to its accessibility and free usage. This shift raised an important question: Does the gender bias favoring males also persist in online educational platforms?

2.2 Gender Bias on YouTube

Numerous studies investigated the existence of gender bias in online platforms including YouTube. For example, [9] worked on online hate speeches on the YouTube platform by examining these kinds of speeches in the comments posted on the videos of popular German-speaking channels. They discovered that female YouTubers are subjected more to hate speeches that are sexually aggressive, sexist, and racist compared to their male counterparts. Moreover, female YouTubers receive fewer compliments on their personalities and the content of their videos. The researchers concluded that the reason for this situation is that maintaining a successful YouTube channel is a sign of dominance and being dominant goes against the conventional female stereotype.

In the context of education, similar results have been found. [2] indicated only 32 of 391 most popular STEM channels have female hosts and searched for the reason behind this by using 450 videos gathered from 90 most popular STEM-related channels. To explore gender-related effects, they classified the videos distinctively such as *Continuous Female Host*, *Teams of Hosts*, and *Female Voice Over*. They sampled 15 videos randomly for each class and compared the comments of these videos to each other. They discovered that channels with female hosts receive more comments for each view. However, the proportion of hostile, critical, and sexist comments through all the comments they get is significantly higher on average.

In their research Gezici et. al. made efforts on the same issue by measuring bias in ranked search results related to STEM and non-STEM educational videos [15]. They introduced two novel bias measures and used these to gauge gender bias in a quantified fashion for different query sub-fields such as *sociology*, *psychology*, and *maths*. They found that there exists a bias towards videos having male narrators in STEM and non-STEM contexts though the magnitude of the bias differs from each other.

2.3 Discourse on Women in STEM-related Fields

So far we explained the existence of gender bias in both face-to-face interactions and on the YouTube platform. This could have more severe effects than expected. In their study, [14] discusses how the discrepancy of gender increases in STEM-related fields and how girls are affected by male dominance. They highlight the fact that girls use social media more than boys and the increasingly male-dominant environment affects their careers. For this reason, they suggested a framework having 96% accuracy for classifying tweets about women in STEM on social media based on sentiment analysis. In addition, contrary to previous findings, they found that the attitude of the people is mostly positive towards women in STEM on the Twitter platform, though this positivity is highly related to dates honoring women's accomplishments.

In her research, Alkhammash discovered the discourse of women in STEM in the Twitter platform [1]. For this purpose, she gathered numerous tweets from 31/10/2017 to 01/11/2017 using popular hashtags like *#womeninTech*,*#GirlsWhoCode*, and *#womenTechTalk*. She found that, on average, Twitter users encourage the women workforce in STEM fields and express their gratitude by using positive adjectives such as *amazing*, *inspirational*, and *great*. She concluded that the discourse of women in STEM supports females in STEM-related fields on online platforms.

All the studies we listed until now, approached the same issue from different perspectives. The results together are not completely compatible with each other. For instance, the first few studies we explained found negative outcomes for women in daily life, education, or in the workforce. The last few, on the other hand, found a supportive attitude towards women in the same contexts. The contradiction in results motivated us to conduct our research and make contributions to the community. To do that, we analyzed the details of 19867 educational videos gathered from various STEM and non-STEM fields as well as their top 500 comments. The time frame of the data we have also allowed us to discover the impact of COVID-19 on video and comment characteristics.

3
Preliminaries

3.1 Statistical Distributions and Tests

In our research, we used the Student t-test, Welch's t-test, and Chi-Square Test of Independence, and Wilcoxon Signed Rank Test to determine the statistical significance of our observations. In this section, we will briefly establish the key statistical concepts behind these tests and then explain them one by one.

3.1.1 Normal Distribution

Normal distribution -or Gaussian distribution- is the most frequently used distribution in statistics [18] due to its properties for using it for statistical inferences. The distribution is not constant; in other words, the shape and location of the distribution change with its parameters σ and μ.

The fundamental properties of Gaussian distribution are:

- Normal distribution is symmetric around the mean.[18]
- A perfect normal distribution has the same mean, median, and mode.
- Since it shows the probability density, the area under the curve equals 1.
- Normal distribution can be defined by using only two parameters. These parameters are mean (μ) and standard deviation (σ).
- 68% and 95% of the data are one and two standard deviations away from the sample mean respectively.

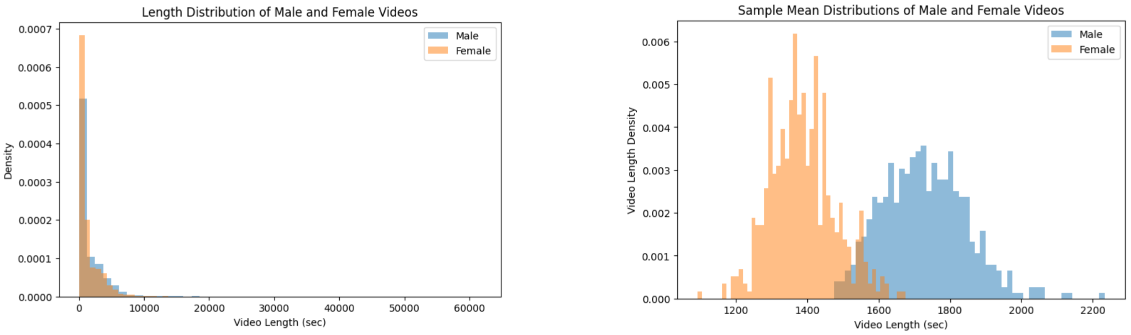
Given μ an σ, a probability density of x can be calculated by using equation 3.1:

$$\mathbb{P}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{3.1}$$

3.1.2 Central Limit Theorem

Central Limit Theorem (CLT) suggests that the combined statistics of randomly selected samples with size n constitute a normal distribution around the population statistic [17]. The selection of n is crucial for conducting a statistical test successfully. When n is *too* small, the ability to reject the null hypothesis (H_0) of a test becomes *too* low which makes the test unreliable. Conversely, as sample size n increases excessively, the probability of rejecting H_0 becomes *too* large. In such cases, a test can reject H_0 even if the difference is negligible [13].

One of the fundamental assumptions that t-tests have in general is that the groups in comparison are normally distributed (This will be explained in detail in Section 3.1.4). In such cases, one can combine multiple sample statistics from the relevant datasets, utilize the CLT, and perform a t-test over these distributions without a problem:



Whole Data Distribution

Sample Mean Distribution

Figure 3.1: Video Duration Distributions

3.1.3 Sampling Methodologies for Statistical Tests

As discussed recently, regardless of the population distribution, one can obtain a normal distribution around the population statistic by gathering many sample statistics together. It is known that in most cases, the bigger the n, the better although a *too* big sample size has downsides on its own. However since the sampling processes for large amounts of data can be impractical in many cases, it is generally not the main concern. In fact, collecting large amounts of data can be expensive, and may not be ethical [4]. In any case, to optimize the statistical results throughout the whole study, we followed the fundamental guidelines. We will mention the way we select the parameters as we explain the sampling processes. Let's start with the bootstrap sampling. The method can be summarized as [11]:

Consider a dataset D having size of N.

$$D = \{x_1, x_2, ..., x_N\} \tag{3.2}$$

A bootstrap sample x can be created if one sample from dataset D for N times with replacement.

$$X^{*1} = \{x_1^*, x_2^*, ..., x_N^*\} \tag{3.3}$$

For each bootstrap sample, the intended statistic is calculated. We used mean and median in this study. After creating many bootstrap samples, one obtains a normally distributed dataset:

$$Z = \{X^{*1}, X^{*2}, ...X^{*B}\} \sim \mathcal{N}(\mu, \sigma) \tag{3.4}$$

where B refers to the number of bootstrap samples. Since $Z \sim \mathcal{N}(\mu, \sigma)$, one can use Gaussian distribution properties (Section 3.1.1) for statistical inference without gathering new samples.

To successfully conduct a bootstrap procedure, the practitioner should select *the number of bootstrap*, and *the size of a bootstrap sample* wisely. The bootstrap method normally designed for small samples. Because of this, in his book, Bradley Efron -the founder of the bootstrap- recommends sample N data points from the dataset having N data points while creating a bootstrap sample [10]. He also mentions that although the optimal number of bootstrap sample can only be reached when $B \to \infty$, the returns become marginal after some point. Thus he recommends to create 1,000 - 2,000 bootstrap sample to get a reasonable estimate of CI and standard error.

Bootstrapping could work for large datasets, however due to increased the memory requirements, it may not be feasible for many cases. To mitigate this problem, various kind of bootstrap approaches have been recommended so far. The ”Subsampled Bootstrap” method is one of them [20]. The founders of the method suggest that, one can obtain a comparable result by only drawing $n \gtrsim \sqrt{N}$ samples from the whole dataset for $B \gtrsim N$ times.

To sum up, here is the parameter selection strategies we adopted during this study:

- **Small datasets:** When we have a small dataset ($N < 500$), we selected $B = 10,000$ since a large B gives a more reliable standard error estimation.
- **Large datasets:** As the size of given dataset increases, the memory constraints start to emerge. Therefore, for large enough datasets ($N > 500$), we

opt for $B = 1,000$.

- **Very large datasets:** When the previous parameter settings cause problems due to memory constraints, we used the Subsampled Bootstrap method ($n \approx \sqrt{N}$ & $B = N$).

Checking the Normality

Following the discussions in the previous section about sample means, we aimed to verify the normality of the distributions we obtained during the experiments. To achieve this, we compared the theoretical expectations with the actual data. For instance, 68% of data points should fall within one standard error from the sample mean, and 95% should be within two standard errors. If the data also verify the theory i.e. 68% & 95% of the data points are indeed one and two standard errors away from the mean respectively, we confirmed the normality of the data. Table 3.1 shows an example of this process:

Table 3.1: Normality Check

Emoji Rank Bootstaps	**68% Coverage**	**95% Coverage**
gender	0.686	0.948
query	0.683	0.95
temporal	0.677	0.95

3.1.4 Independent Student's t-test

The Independent Student's t-test is a statistical method used to determine if there is a significant difference between the means of the two samples [22]. Like many other statistical tests, Independent Student's t-test also has some assumptions:

- The sample groups in comparison are both normally distributed.
- The sample groups are coming from independent populations.
- The sample groups have equal variances.

The test statistic of the Student's t-test can be calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s}{\sqrt{n}}} \tag{3.5}$$

And the corresponding p-value can be found from statistical software or tables by using the test statistic and degree of freedom of the test.

In this thesis study, we determined critical p-value (p_{crit}) as 0.05. Since we conducted multiple hypothesis tests using the *same* dataset, we applied Bonferroni correction as well. The corrected critical p-values are listed in Table 3.2:

Table 3.2: Adjusted Critical P-Values

Area of Interest	Questions	Dimensions	Metrics	Tests Conducted	Corrected p-value
Video Details	3	3	1	1	0.006
Comment Length	1	3	3	1	0.006
Comment RBO	1	3	1	1	0.016
Sentiment Analysis	2	3	1	5	0.0016
Title - Keyword Similarity	1	3	1	1	0.016

Also, the hypotheses and decision criteria can be summarized as follows:

H_0: The difference between two groups is negligible.

H_1: The difference between two groups is statistically significant.

Decision Criteria

$p_{test} < p_{crit}$: Reject H_0

$p_{test} \geq p_{crit}$: Can not reject H_0

3.1.5 Independent Welch's t-test

Welch's t-test is similar to the Student's t-test. The only difference between these two is that Welch's t-test does not assume the sample groups have equal variances [28]. In Welch's t-test, t-statistic can be calculated using equation 3.6:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{3.6}$$

The term *equal variances* is somewhat abstract and subjective. Therefore, we used the variance rule of thumb to determine whether two sets of samples have equal variances or not. The rule of thumb indicates that if the variance ratio of a set of sample means to the other is greater than a predetermined constant, the variances are not equal [8]. If otherwise, one could assume that variances are equal. We choose that predetermined constant as 4 since this magnitude allows us to capture unequal variances more robustly.

3.1.6 Chi-Square Test of Independence

The chi-square test of independence is a fundamental non-parametric statistical test to examine whether there is a significant difference between observed and expected frequencies [5]. It is often used for categorical data therefore a contingency table, a table containing the value counts of the specific category, is required. The chi-square is calculated via Equation 3.7:

$$\tilde{\chi}^2 = \frac{1}{d} \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{3.7}$$

where O_k and E_k refer to observed and expected frequency respectively. The expected frequency is dependent on the sample size therefore, the chi-square test of independence is sensitive to sample size. There are two guidelines to apply the chi-square test. These are

- Any of the observed frequencies should be at least 5. If that is not the case Yates correction should be applied [7].
- The sample size should be between 100 and 200 for the chi-square measure [30]. Less than 100 makes test results biased towards non-significance and larger than 200 does the reverse. In other words, if the sample size is small, even if there are large differences, the test results could be not significant and the opposite could occur if the sample size is large.

Considering these, we selected our sample sizes as 200 and since our observed frequencies were at least 5, we did not apply Yates correction in our analysis.

3.1.7 Wilcoxon Signed Rank Test

Wilcoxon Signed Rank Test is a non-parametric test that can be used for paired samples [3]. It compares the medians to determine whether two distributions are statistically different from each other. The test ranks the absolute differences for each data point and then creates one ranking for positive differences and one for negative differences.

If we use T^+ to denote the sum of positive ranks, when the number of paired samples exceeds 10, the distribution of positive ranks can be approximated to a normal distribution having parameters:

$$\mu_{T^+} = \frac{n(n+1)}{4} \tag{3.8}$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{3.9}$$

With these parameters, the test statistic is calculated for the sum of positive ranks and it is compared against the critical values of the normal distribution to determine whether the difference between paired samples is significant.

3.2 Ranking Algorithm

The scope of this thesis project involves ranking comparison. To conduct this comparison operation objectively, we need a ranking similarity measure. Before determining our approach, we need to consider the following properties of a ranking problem [33]:

- **Incompleteness**: Rankings often gathered from a big data source, which can lead to challenges, including missing or incomplete information.
- **Uneven importance**: The difference in the top part of a ranking is usually more important than the ones that appear in the bottom part.
- **Uneven Ranking Depth**: Calculating similarity between two rankings on a fixed evaluation depth is straightforward yet not realistic.

The bullet points we have defined above make the most commonly used ranking similarity measures inefficient. For instance, the Kendall Tau distance measure automatically assumes all the variables between two ranks are conjoint. Moreover, it is an unweighted measure i.e., it does not give a higher importance to the top part of the ranks. Another one is the Spearman Correlation Coefficient. This algorithm also can not handle uneven ranking depth effectively and can have sensitivity and scaling problems.

3.2.1 Rank Biased Overlap Algorithm

Rank Biased Overlap (RBO), is a ranking similarity measure that considers the bullets we defined above [33]. Let us explain how the algorithm works:

$I(S, T; d)$: Intersection of rank list S and T to the depth d
$X(S, T; d)$: Intersection length or $|I(S, T; d)|$
k : Evaluation depth
d : Current depth

$$\text{Proportion of overlap at depth d: } A(S, T; d) = \frac{X(S, T; d)}{d} \tag{3.10}$$

$$\text{Average Overlap: } AO = \frac{1}{k} \sum_{d=1}^{k} A_d \tag{3.11}$$

$$\text{Similarity: } SIM(S,T;w) = \sum_{d=1}^{\infty} w_d A_d \tag{3.12}$$

for $0 < p < 1$ the geometric series applies:

$$\sum_{d=1}^{\infty} p^{d-1} = \frac{1}{1-p} \tag{3.13}$$

Combine Equation 3.12 & 3.13 and consider $\sum_d w_d = 1$, then Equation 3.12 becomes:

$$RBO = (1-p) \sum_{d=1}^{\infty} p^{d-1} A_d \tag{3.14}$$

where the parameter p is used to determine how fast we want to ranking weights to decline. In other words, as p increases, the length of the most top-rankings decreases.

Although the RBO algorithm effectively handles the bullets we defined in the previous section, the algorithm is intrinsically biased toward longer lists. In other words, when the ranking length changes across the groups, the similarity measure always rewards the longer groups, even if the total similarity is the same. To mitigate this problem Equation 3.15 (RBO Extrapolated) is proposed by the authors:

$$RBO_{ext}(S,T,p,k) = \frac{x_k}{k} p^k + \frac{1-p}{p} \sum_{d=1}^{k} \frac{x_d}{d} p^d \tag{3.15}$$

Computing the ranking similarity using equation 3.15 allows the algorithm to end up with a higher score if the extra items of the longer lists have commonalities and end up with a lower score if otherwise is the case. In this work, we used RBO_{ext} because of its robustness.

We selected a p value of 0.9 for our analysis. This choice is based on the fact that, with this p value, approximately 86% of the weight in the similarity measure is assigned to the top 10 rankings. This effectively emphasizes the importance of higher rankings in our calculation of similarity.

3.3 Measuring Vector Similarity

Regardless of the task, all the machine learning algorithms work with numbers. When the task involves words or sentiments, this still applies. To represent the vocabulary terms, we use word embeddings [34]. Word embeddings are generated in various ways such as one-hot encoding, Term Frequency-Inverse Document Frequency (TF-IDF), and neural network-based models like Word2Vec. In our analyses, we used the ”Twitter RoBERTa Base for Sentiment Analysis” model to generate the word embeddings. The details about the model will be explained in the Section 4.2.2.

One popular similarity metric in the context of information retrieval is cosine similarity [25]. It is useful for calculating how similar each word, phrase, or paragraph is to another. Once the texts in consideration have been converted to embeddings, the cosine similarity formula (Equation 3.16) can be used:

$$Sim(\vec{g}, \vec{q}) = \frac{\vec{g} \cdot \vec{q}}{|\vec{g}||\vec{q}|} \tag{3.16}$$

4 Problem Definition and Methodology

In this chapter, we set up the basis of our study. Section 4.1 covers the problem we're focusing on and the questions related to it. Then, in Section 4.2, we describe the methods we used to tackle these questions and examine our data.

4.1 Problem Definition and Research Questions

The primary goal of this thesis is to investigate the presence of gender bias on digital education platforms as well as understand the impact of the COVID-19 pandemic on the video and comment characteristics shared on the digital platforms. In this regard, we worked through a set of educational videos as well as their top-ranked comments based on content relevancy and posting time. We focused on understanding the role of three distinct dimensions in this problem. These are time (Precovid vs. Postcovid), the gender of the video narrator (Male vs. Female), and the query field (STEM vs. non-STEM). Through a detailed analysis of a large dataset, we aimed to gain meaningful insights into gender bias on digital education platforms.

During our analysis, to make it more detailed, we retrieved some additional information. The information we had, led us to approach this problem from two perspectives: the video side and the comment side. As our investigation progressed, our research questions evolved and became more defined. The final form of our research questions both in terms of comment and video side can be listed as follows:

Video Based Research Questions:

- **RQ1:** Did the COVID-19 pandemic impact the video durations?
- **RQ2:** Did the COVID-19 pandemic influence the video release rates?

- **RQ3:** Does the viewer engagement vary between dimensions?

Comment Based Research Questions:

- **RQ4:** Does the length of the comments significantly differ from each other between the periods?
- **RQ5:** Do the most frequent words and emojis used in the comments differ considerably between the videos having male narrators and female narrators?
- **RQ6:** How similar are the most frequently used emoji rankings between dimensions?
- **RQ7:** Does the relevancy of the recent comments change across the dimensions?
- **RQ8:** Is the gender of the video narrator an impactful parameter on comment polarity or emotion?
- **RQ9:** Does the title - keyword similarity changes between dimensions?
- **RQ10:** In terms of Named Entities and nouns, is there any outstanding word or phrase that gives a clue for us to understand the dynamics of gender bias?

4.2 Methodology

In this section, the methodology behind this work is explained. Section 4.2.1 highlights the steps taken to prepare the data. Section 4.2.2 provides an overview of the sentiment analysis model employed in our study. Section 4.2.3 outlines the steps we took for Named Entity and Noun tagging and finally, Section 4.2.4 describes the methodology followed to assign the gender of the video narrators.

4.2.1 Data Preprocessing:

The raw dataset we used in our studies, required some preprocessing steps to get rid of the irrelevancies. Therefore, before the analysis, we applied many data preprocessing steps to improve the quality of the research. In this section, we will describe these irrelevancies and the steps we took to deal with them.

The Data Preprocessing Steps:

- **Removal of Duplicate Video Identifiers (ID):** To prevent potential bias in our analysis, we first identified and removed any duplicate video IDs from the dataset. Table 1.5 illustrates the point we mention in this bullet.

- **Elimination of Timestamp Comments:** All the comments do not necessarily contain important sentiments for our research area. Timestamp comments, the ones only containing timestamps, are one of them. Here is an example comment that we eliminated in our dataset:

  02:42 Brain development.
  04:00 Function.
  05:07 The thalamus and hypothalamus.
  06:38 Using functional MRI
  07:56 The basal ganglia.
  09:05 Parietal lobe.

  These comments are posted to indicate specific moments in a video for various reasons. They add little to no value to our research and should be removed.

- **Task-Specific Preprocessing:** Additional preprocessing steps were applied to the specific analytical tasks at hand. Details of these steps will be provided in the sections of this thesis where each specific task is described. For instance, the both Name Entity Recognition and Noun Extraction algorithms are not perfect. Therefore they sometimes gave meaningless outputs to us. To deal with that problem we had to perform additional text preprocessing steps.

4.2.2 Sentiment Analysis and The Model We Used

Sentiment analysis is a sub-field of Natural Language Processing (NLP) that attempts to discover the sentiment in the text [31]. Advancements in computer sciences allowed us to perform this analysis better over time. Especially in recent years, the improvements have been remarkable. Currently (2024), the state-of-the-art model for many NLP applications -including sentiment analysis- is Bidirectional Encoder Representations from Transformers (BERT). The key features contributing to its effectiveness are:

- **Large training data:** Aside from the nuances, generally more data improves a model's prediction performance. BERT was trained on a vast amount of text data. With around 3.3 billion words in its training corpus, this extensive data foundation is a major reason for BERT's high effectiveness.

- **Masked Language Modeling:** BERT's training does not rely on labeled data. Instead, it learns by masking certain words in a sentence, predicting these words, and then updating its predictions, accounting for about 50% of its training.

- **Next Sentence Prediction:** To better understand the relationship between sentences, BERT incorporates next sentence prediction in its training process, which also constitutes 50% of its learning mechanism.

- **Multi-head Attention Mechanism:** The multi-head attention allows BERT to focus on different parts of the text simultaneously allowing it to understand the context of the text even better.

The encoded texts are currently being used for performing a variety of tasks such as sentiment analysis, question answering, text prediction, and summarization.

To conduct our sentiment analysis, we employed the Robustly Optimized BERT Pretraining Approach (RoBERTa), developed by researchers at Facebook and Washington University, due to its superior performance capabilities over other BERT models [19]. The researchers tweaked the BERT model based on the observations they made on BERT, and RoBERTa was created. The main observations seen and corresponding actions taken during the development of RoBERTa include:

Observation: Next sentence prediction is not improving the learning process.
Action: Remove the next sentence prediction objective and learn only by using Masked Language Modeling.

Observation: BERT underfits a lot.
Action: Train it for longer.

Observation: BERT can handle more data.
Action: Train it with more data (2.5. terabytes of text data).

Observation: BERT training can be unstable.
Action: Use larger batches while training the model.

Transfer learning refers to applying knowledge obtained from one data to another [32]. In the case of BERT models, they have already been trained with a vast amount of data thus one can use them directly. However, to get better results, one can change the pretrained model slightly by training the pretrained model for a short amount of time with a more specific dataset. This process is called *fine-tuning*. That being said, in this thesis study, we employed a fine-tuned RoBERTa called ”Twitter RoBERTa Base for Sentiment Analysis” [6]. This model was specifically fine-tuned for a variety of text classification tasks. The tasks and the corresponding labels are:

- **Emoji Prediction:** Classifies texts based on the emoji used, i.e. the labels are emoji names.

- **Polarity Analysis:** Categorizes texts as ”positive”, ”neutral”, or ”negative”.

- **Emotion Analysis:** Identifies the dominant emotion in the input sentence.

The labels are ”joy”, ”anger”, ”optimism”, or ”sadness”.

- **Offensive Speech Detection:** Detects whether a text is ”offensive” or ”not-offensive”.
- **Hate Speech Detection:** Determines if a text contains hate speech or not. The labels are ”hate” and ”not-hate”.
- **Stance Analysis:** Assesses the stance of the text as ”none”, ”against”, or ”favor”.
- **Irony Analysis:** Classifies as ”irony” or ”non-irony”, depending whether input text contains an irony.

We used both the polarity and the sentiment analysis task in our analyses. The comment label is determined based on the scores that RoBERTa returns. For instance, the comment ”A teacher is always like a parent, the education system in somalia where ever this is, it s really very very poor....” gets a negative score of 0.903058, a positive score of 0.007080 and a neutral score of 0.089862. Therefore the label of this comment becomes ”Negative” since 0.903058 is by far the greatest score of all labels.

4.2.3 Name Entity Recognition and Noun Extraction

Name Entity Recognition (NER) is a task of NLP, which is used for assigning the phrases and words some tags like ORGANIZATION, LOCATION, or PERSON [16]. There are multiple ways to perform this task such as using Hidden-Markov Model and Conditional Random Field.

In this thesis work, we used a different RoBERTa model fine-tuned for this particular task. We only used the words having at least 95% probability of being a named entity because from our observations we saw that the lower probability tags frequently be wrong. Moreover, even if the baseline NER algorithm (Algorithm 1) has an accuracy score of 92%, so incorporating low-probability named entities is not appropriate.

Algorithm 1 Baseline Named Entity Recognition Algorithm

```
if the given word is ambiguous then:
    Choose the most frequent tag in the training corpus.
else:
    pass
end if
```

For the noun extraction, we used the TextBlob library. Both algorithms are not perfect, due to this reason we had to apply a couple of further text preprocessing steps to the results. These include:

- **Removing ' s and ' m:** For example, noun extraction result for the comment "“It’s the entire purpose of life, so there’s no reason to blush” just found the best pick up line” is "' s entire purpose ' s blush "". The "' s" does not contain any meaningful information, therefore, should be eliminated.
- **Removing extra white spaces:** In case of the presence of extra white spaces either in the beginning or in the end, we checked and stripped these empty lines.
- **Lemmatization:** While analyzing the common and differing nouns and Named Entities, we created Word Clouds. These Word Clouds were prepared using only the most common (or differing) 50 words. We did not want words having similar meanings to be displayed in these Word Clouds therefore we used lemmatized the outputs.

4.2.4 Gender Assignment

One of the main goal of this work is to explore the presence of gender bias on online education platforms. To do that it is necessary to categorize the videos based on the gender of the video narrator. For this purpose, we used two features that are included in our dataset. The first one, sampling_biased_audio_male_ratio, refers to the probability of the video narrator being male, and the second one, sampling_biased_audio_female_ratio, corresponds to the probability of the video narrator being female. The labeling process was fairly simple. When the probability of male exceeds the probability of female, the gender assigned to a particular video would be "male" and female otherwise. Although this approach sounds reasonable, one can suspect that when the probabilities are close to each other, the classification algorithm becomes prone to make mistakes. This is a valid concern and since we can not manually label all the videos, we checked whether the test results change if the classification algorithm also changes. Starting from a threshold probability of 70% to 99%, we classified all the videos and repeated the experiments. Not surprisingly, as the classification probability threshold increases, the number of classified videos decreases.

Table 4.1: Classification Thresholds and the Video Counts

Threshold	Male Video Counts	Female Video Counts	Not Classified
0.7	10107	4568	5192
0.75	9912	4454	5501
0.8	9650	4319	5898
0.85	9346	4163	6358
0.9	8982	3982	6903
0.95	8374	3704	7789
0.99	7029	3063	9775
Default Settings	10782	5086	3999

Although the number of classified videos changes considerably, this change does not reflect the experiment results.

5

Descriptive Analysis & Test Results

Performing a descriptive analysis before the main study is important because it allows the researcher to get a clear overview of the dataset. Once the researcher *understands* the data, (s)he can shift his/her focus to the correct places.

Because of this reason, during this thesis project, we also conducted a descriptive analysis. The insights gained from this analysis shaped our decisions in terms of statistical test preferences and data preprocessing methods to apply before proceeding with more complex analyses. In this section, we will explain the findings of the descriptive analysis we conducted as well as the statistical tests we performed.

5.1 Video Duration Analysis

We started our analysis by looking at the video durations. We realized that video counts change significantly in the gender dimension (Table A.1). Besides that the variation in minimum and median durations are considerably less than the maximum durations. The same explanation can be made while explaining the relationship between videos posted before and after the COVID-19 announcement date as well as the non-STEM and STEM videos.

To see whether video durations statistically differ from each other, we split the data according to the dimensions, sample from these datasets, and finally calculated the confidence intervals (CI) to generate our hypotheses. Since the maximum values vary significantly, we compared sample medians.

We found that the video durations vary across the dimensions (Table A.2). Videos having male narrators tend to be longer compared to videos having female narrators. STEM videos and non-STEM videos have somewhat similar CI's. The only difference is that the variance of STEM videos are a bit less compared to the other. Lastly, the videos published in the Postcovid period are significantly longer than the ones published in the Precovid period. The next step is to test the hypotheses using t-test:

Table 5.1: Video Duration Test Results

Experiment	Result	Test Statistic	p-value
Male vs Female	Male > Female	183.182	0.0
Precovid vs Postcovid	Precovid < Postcovid	-687.363	0.0
STEM vs non-STEM	STEM > non-STEM	21.284	4.494E-91

Table 5.1 proves that the hypotheses we made are likely to be true. Coming back to the research question:

RQ1: Did the COVID-19 pandemic impact the video durations?
To answer this question, we plotted video durations and release dates in a scatter plot since scatter plots allow us to observe data points individually (Figure 5.1).

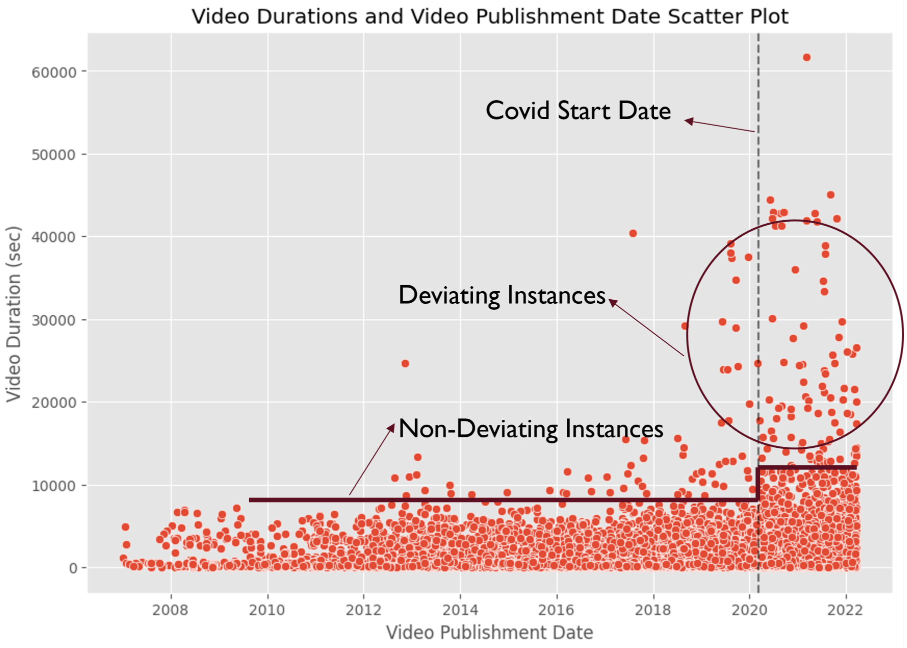


Figure 5.1: Video Durations and Release Dates

The scatter plot shows that many videos existing in our dataset are considerably longer than the average video duration (Deviating Instances). On the other hand, in the COVID-19 period, both deviating and non-deviating instances either became longer or more frequent on average. After these findings, the following research questions appeared:

- What are these deviating videos and what are their subjects?

- What is the main motivation behind making longer videos?
- Are there any biases towards one dimension?

To find answers to our questions, we investigated and manually labeled the longest videos that appeared in our dataset (Table A.3 & Table A.4). We found that almost all the longest videos released Precovid and Postcovid periods are lecture videos (Figure A.1). Moreover, these types of videos are not common until the year 2019. In the lighting of these results, we concluded that although COVID-19 is not the main reason, it might be one of the motivators behind making longer videos more frequent. Aside from the reason, we also found that the increase in video duration is reflected in each dimension (Table 5.2).

Table 5.2: Video Duration Comparison

Dimension	Result	Test Statistic	p-value
Male	Precovid < Postcovid	-453.402	0.0
Female	Precovid < Postcovid	-426.752	0.0
STEM	Precovid < Postcovid	-611.465	0.0
non-STEM	Precovid < Postcovid	-360.860	0.0

Lastly, we examined the query sub-fields of the longest videos. We found that these videos tend to be STEM videos and they are mostly related to Math and Computer Sciences (Figure A.2).

5.2 Video Release Rate Analysis

To analyze how the recent pandemic changed the video posting rates, we grouped the videos based on their channel and release date. Then, we computed the yearly difference in video counts per channel. After that, we summed up the differences.

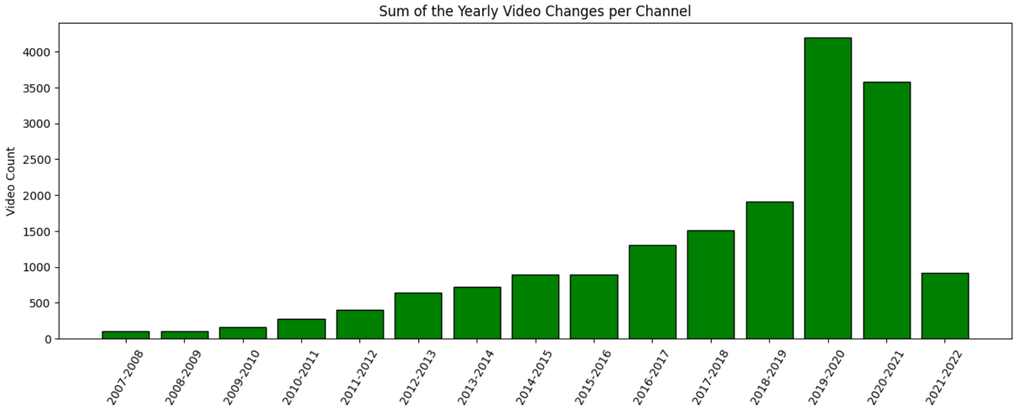


Figure 5.2: Sum of the Yearly Video Changes per Channel

We detected a significant increase in video posting rates in 2020 and 2021. However, this increase does not continue in 2022. The mismatch between 2022 and the previous two years is expected because our dataset only covers the videos published until March 21, 2022. More importantly, our plot *may* indicate a correlation between the COVID-19 pandemic and the fluctuating rates of video releases.

RQ2: Did the COVID-19 pandemic influence the video release rates?

To answer this question objectively, we analyzed *only* the channels present in both the Precovid and Postcovid periods. The video release frequency for each channel during each period was calculated by dividing the total number of videos by the number of days the channels are active in those periods.

The active day calculations differ slightly in Precovid and Postcovid conditions. For example, if a channel posted it's first video at 01/01/2018 and its last video at 01/01/2021, the the active days of that channel in Precovid condition is calculated by subtracting COVID-19 announcement date from 01/01/2018 (800 days). For the Postcovid condition on the other hand, the number of active days is found by subtracting 01/01/2021 from the COVID-19 announcement date (296 days).

The results we get from these calculations were highly skewed and therefore the sample mean distribution failed to satisfy the Gaussian assumption of the t-test. For this analysis, we used the Wilcoxon Signed Rank Test instead. Regarding that, we performed two experiments. The first experiment covers all the data. The upside is we have more data points compared to the second experiment; however, including all data disregards the duration imbalance between the Precovid and Postcovid periods (4811 and 740 days respectively). The experimental procedure of the second experiment solves this issue by only considering the last 740 days in the Precovid period with the price of eliminating some of the data points.

From the test results (Table 5.3 & Table 5.4), we saw that the video posting rate indeed increased. However, when the period lengths are equated, these differences disappear. In summary, we concluded that the video release rates increase over time but, even if COVID-19 has an impact on this situation, the influence of the event is undetectable.

Table 5.3: Video Release Rates Comparison

Dimension	**Result**	**Test Statistic**	**p-value**
Male	Precovid < Postcovid	12691	2.5E-13
Female	Precovid < Postcovid	3702	0.0002
STEM	Precovid < Postcovid	13390	9.46E-5
non-STEM	Precovid < Postcovid	12937	1.02E-9

Table 5.4: Time Equated Video Release Rates Comparison

Dimension	Result	Test Statistic	p-value
Male	Precovid = Postcovid	11310	0.15
Female	Precovid = Postcovid	2643	0.90
STEM	Precovid = Postcovid	10726	0.69
non-STEM	Precovid = Postcovid	11752	0.87

5.3 Video Like - Video View Ratio Analysis

A common expectation is that videos having more likes would also have more views, and this is also the case in our dataset (Figure A.3). We used the ratio between likes and views to gauge viewer engagement.

RQ3: Does the viewer engagement vary between dimensions?

To explore whether this relationship differs across the dimensions, we split the dataset according to the dimensions and calculated the like-view ratios of the videos. After that computed the CI for each dimension (Table A.5).

Videos with female narrators had a higher average of likes per view than those with male counterparts. Similarly, videos on non-STEM topics outperformed STEM topics in the like-view ratio. We also realized that the like-view ratio increased in the Postcovid period. Lastly, we tested these hypotheses (Table 5.5) and verified that all of our hypotheses are statistically significant.

Table 5.5: Like-Count Ratios Among Dimensions

Experiment	Result	Test Statistic	p-value
Male vs Female	Male < Female	-84.31	0.0
STEM vs non-STEM	STEM < non-STEM	-36.02	1.11E-219
Precovid vs Postcovid	Precovid < Postcovid	-866.79	0.0

In addition to these findings, we also looked at the similarity of the dimensions in terms of viewer engagement in Precovid and Postcovid conditions by considering *only* the channels that are present in both periods. Concerning that, we performed two experiments similar to the ones we conducted in the previous section. While the first one includes all the data in the Precovid period and leaves the durations imbalanced between periods, the second one equates the number of days in both of the periods with the price of eliminating some of the data points in the Precovid period.

The results of the experiments show that the increase in viewer engagement is reflected in each dimension (Table A.6, Table A.7).

5.4 Comment Length Analysis

Analyzing comment lengths is a sensitive task because focusing on just one aspect can lead to skewed results. For example, only counting words might miss nuances that we can potentially get from characters, counting characters might ignore word importance, and not excluding spaces could give misleading length estimates. Therefore, we analyzed the comment lengths in three different ways: word counts, character counts, and character counts without spaces.

RQ4: Does the length of the comments significantly differ from each other between dimensions?

To find out whether there are any significant differences across dimensions, we sampled our dataset and computed confidence intervals (Table A.8). From the confidence intervals, we hypothesized that, for all the comparison parameters, comments posted on the videos having male narrators are longer than the ones having female counterparts. Similar hypotheses can be made to explain the relationship between non-STEM and STEM videos as well as videos released in the Precovid and Postcovid periods. Lastly, we tested our hypotheses and verified them statistically (Table 5.6).

Table 5.6: Comment Length Experiment Confidence Intervals

Parameter	Result	Test Statistic	p-value
Sentence Length	Male > Female	612.11	0.0
Sentence Length	Precovid > Postcovid	775.76	0.0
Sentence Length	STEM < non-STEM	2630.68	0.0
Character Length	Male > Female	639.48	0.0
Character Length	Precovid > Postcovid	800.68	0.0
Character Length	STEM < non-STEM	2600.19	0.0
Character Length w.o. ' ' *	Male > Female	643.80	0.0
Character Length w.o. ' '	Precovid > Postcovid	812.54	0.0
Character Length w.o. ' '	STEM < non-STEM	2603.25	0.0

***** "Character Length w.o. ' ' stands for Character Length without Spaces

When we looked at how the measurement results changed between the periods, we encountered a consistent pattern. In other words, regardless of the dimension and measure, the Precovid results outperformed the Postcovid results (Table A.9).

5.5 Frequency Analysis

RQ5: Do the most frequent words and emojis used in the comments differ considerably between the videos having male narrators and female narrators?

For this question, we investigated the word counts and emoji counts of the videos and grouped them based on the gender of the narrator. Before conducting the analysis, we applied some basic NLP operations to the comments. These include:

- **Eliminating grammar-related elements, such as punctuations, contractions, and extra white spaces:** While splitting words in the comments, it is important to remove punctuations. Otherwise, the last word could be recognized as a different word due to punctuation at the end. For example, if the punctuation removal step is skipped, the word "ball." and "ball" will be recognized as different words. The same problem arises when contractions and extra white space elimination steps are skipped.
- **Removing the stopwords and the numbers:** Stopwords like "a", "the", "is", and "are" contain little to no valuable information, and they are used frequently in the sentences. To prioritize the words having sentimental meanings, we eliminated the stopwords beforehand. The same is true for the numbers.

Since the comment counts between videos having male and female narrators are not equal, we preferred word frequencies over word counts as the main evaluation metric. We found that word usage percentages are not that different between genders except for a few. Table 5.7 shows only the words having a frequency difference above 1%:

Table 5.7: Word Frequency Difference

Word	**Male %**	**Female %**	**Difference %**
mam	0.012	5.306	5.294
sir	5.015	2.313	2.701
would	9.958	7.509	2.449
madam	0.012	2.166	2.154
man	2.246	1.015	1.231
could	4.607	3.552	1.054
guy	1.484	0.439	1.044

Naturally, the words *mam* and *sir* are the words having the most frequency difference as these words are used for a specific gender exclusively. Similarly, the frequency difference between the words *madam*, *man*, and *guy* can be explained due to the same gender-specific context. Since the remaining words (would and could) have no bias-related meaning, we did not investigate them even further.

Following this argument, we applied the same procedure only the comments containing emojis. Since the number of comments containing emojis are lower, some of the percentages attenuated. Table 5.8 reveals the words having frequency difference above 1%:

Table 5.8: Word Frequency Difference in Emoji Containing Comments

Word	Male %	Female %	Difference %
mam	0.028	9.901	9.873
sir	11.228	4.157	7.071
madam	0.021	4.574	4.554
thank	15.176	16.865	1.689
man	2.031	0.619	1.412
much	9.255	10.439	1.184
would	4.54	3.364	1.176

Again, excluding the gender-specific and the modal words, only the word *thank* remains. Solely the meaning of the word *thank* tells something however, the difference is too small (1.6%) and can be explained by the difference between comment counts.

The next step is to analyze the emoji frequencies. In this regard, we pulled the emojis from the video comments and analyzed the differences in gender dimension. Table 5.9 shows the union of the top 10 most frequent emojis posted on the comments of the videos having male and female narrators as well as their usage frequencies.

We found that 8 of the 10 most frequent emojis are common. This means we can't interpret much from the differences except the heart emoji. Its frequency is twice as much in female group compared to the male group. Also, it's important to note that the top frequent emojis are predominantly positive which might be an indication of a positive mood in the comments section.

Table 5.9: Emoji Frequencies

Emoji	Male %	Female %
❤️	0.482	0.862
😂	0.446	0.443
👍	0.34	0.456
🙏	0.267	0.385
😊	0.265	0.466
😅	0.119	0.128
🤔	0.108	0
🤣	0.107	0
😁	0.1	0.114
😭	0.099	0.148
☺️	0	0.124
🙂	0	0.121

5.6 Rank Analysis

Emoji Rank Analysis

RQ6: How similar are the most frequently used emoji rankings between dimensions?

As discussed before, the scope of this thesis project involves ranking comparison. This section explains our efforts as well as the main findings we obtained from the ranking comparison. We compared the ranks of both emojis and comments. While comparing emoji rankings, the steps we took are as follows:

- **Sampling videos:** We randomly selected videos from relevant datasets, such as STEM vs non-STEM.
- **Extracting emojis:** For each video, we reviewed all the comments we had, and extracted the emojis as well as their respective counts.
- **Sorting emojis:** We sorted the emojis in descending order based on their counts. To standardize the ranking for emojis occurring the same number of times, we applied a secondary sort based on the emojis' Unicode values.
- **Calculating ranking similarities:** We employed the RBO Extrapolated method with $p = 0.9$ to calculate the similarity in emoji rankings between different lists (Section 3.2).
- **Obtaining a sample distribution:** We repeated this process 250 times to obtain a reasonable amount of data points.
- **Establishing control groups:** To validate our results, we replicated a similar procedure to create control groups. The only difference is that samples are drawn from the same datasets.
- **Bootstrapping:** To compare the groups and obtain a confidence interval we applied bootstrap sampling among the RBO datasets.
- **Comparing with control groups:** Finally, we compared the emoji ranking distributions from our main analysis with those obtained from the control groups using t-tests.

Table 5.10 demonstrates the experiment results. As expected, the emoji rank similarities of the control group are higher than the test groups. All emoji ranking similarities range somewhere between 0.18 and 0.22. From the results we obtained, we concluded that emoji ranking similarity is not closely correlated with any specific dimension.

Table 5.10: Control and Test Group Comparison

Dimension	Control Group CI	Test Group CI	Accepted Hypothesis
Male vs Female	(0.2, 0.23)	(0.19, 0.22)	Test Group < Control Group
STEM vs non-STEM	(0.23, 0.26)	(0.19, 0.22)	Test Group < Control Group
Precovid vs Postcovid	(0.19, 0.22)	(0.18, 0.21)	Test Group < Control Group

Comment Rankings

RQ7: Does the relevancy of the recent comments change across the dimensions?

While comparing the comment ranks, our goal was to understand the influence of time on the relevancy of comments. The higher ranking similarity indicates that the recent comments are more relevant and less relevant if otherwise is the case. For that purpose, we collected comments from videos in two distinct datasets, using *Time* and *Relevancy* options of YouTube API. In the next step, computed RBO_{ext} score of the randomly selected samples by comparing the comment rankings in two datasets. Lastly, we applied bootstrap sampling and calculated the 95% CI for RBO_{ext} scores.

The confidence intervals (Table A.10) highlight significant differences in comment relevance across the dimensions. Recent comments posted on videos having female narrators tend to be more relevant than videos having male counterparts. The same interpretation can be done when comparing comments posted on Precovid and Postcovid periods, as well as comments posted on STEM and non-STEM videos respectively. The last step is to test our hypotheses statistically.

Table 5.11: Comment Rank Experiment Test Results

Experiment	Result	Test Statistic	p-value
Male vs Female	Male < Female	-142.378	0.0
STEM vs non-STEM	STEM > non-STEM	10.71	6.81E-26
Precovid vs Postcovid	Precovid < Postcovid	-1076.273	0.0

The results of Table 5.11 show that all the hypotheses we came up with are statistically significant.

5.7 Sentiment Analysis

RQ8: Is the gender of the video narrator an impactful parameter on comment polarity or emotion?

Polarity analysis tells us if a comment is positive, negative, or neutral. Emotion analysis, on the other hand, helps us to find out the dominant feeling in a comment, like optimism or sadness. We started with polarity analysis because the outcomes of this analysis are a bit more general compared to detecting specific feelings. This way, we first get a big picture of the comment's tone and then proceed with a more detailed analysis.

Polarity Analysis

We used RoBERTa in polarity settings to get the sentiment scores and labels of all the comments. Then we divided the comments according to the gender of the narrator that the comments were posted, and plotted the results on a bar chart to get an overview of the label distribution across the dimensions.

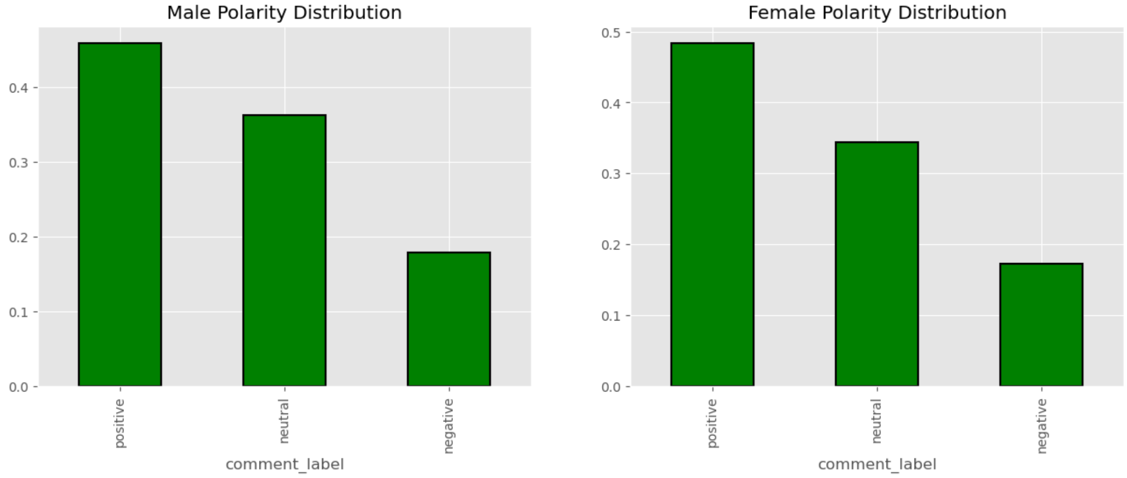


Figure 5.3: Polarization Distribution

We found that the polarity distributions between the comments posted on the videos having male and female narrators are similar. Lastly, we repeated multiple chi-square tests to see whether these distributions indeed do not differ from each other.

The reader may ask the reason for the multiple hypothesis tests. Repeating multiple statistical tests increases the Type 1 error probability. However, compared to the size of our dataset, the sample size we used was small due to the sample size criteria and we wanted our samples to represent the data in a strong sense. Also, since we used adjusted p-values in our tests, we protect our analysis to suffer from type 1 error probability.

Nevertheless, the results of our tests indicate that the difference between the distributions is not statistically significant (Table 5.12). Thus we concluded that polarity is not affected by the gender of the narrator.

Table 5.12: Polarity Comparison Test Results

Polarity Test	Test Statistic	p-value
Test Results 1	1.085	0.581
Test Results 2	0.365	0.833
Test Results 3	1.064	0.587
Test Results 4	0.485	0.785
Test Results 5	1.294	0.524

Emotion Analysis

In the previous section, we concluded that the narrator's gender is not an important parameter for comment polarity. Thus, we extended our analysis to investigate whether the narrator's gender plays a role in the emotional tone of comments. At first glance, we employed RoBERTa in emotion labeling settings to obtain the emotion sentiment probabilities and the labels of the comments. After that, we split the dataset across the dimensions and plotted the label distributions. From the plot (Figure 5.4), we hypothesized that there are no major differences between the comments posted on the videos having male and female narrators in terms of emotion distribution.

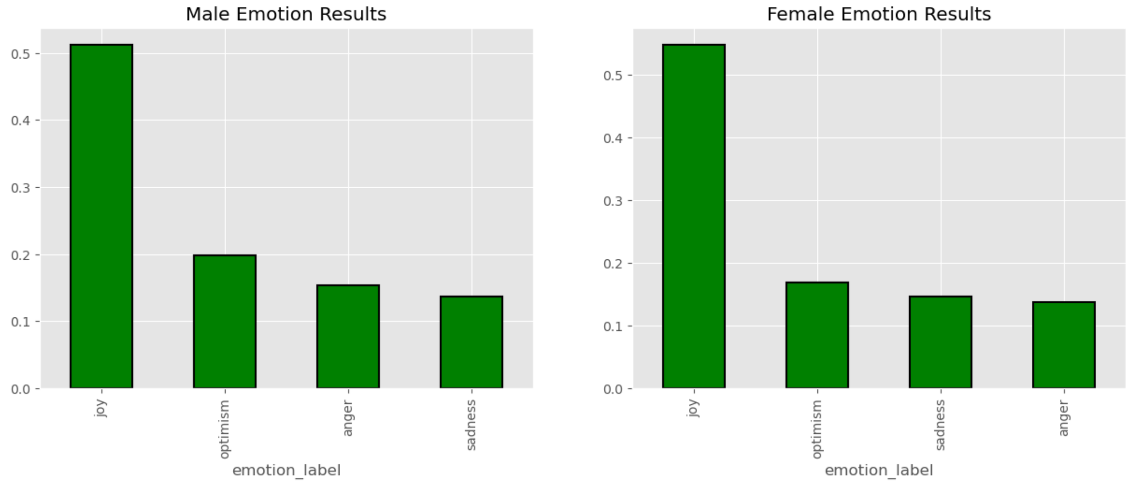


Figure 5.4: Gender Emotion Distribution

The last step was to test this hypothesis statistically. To do that, we applied multiple chi-square of independence tests to our samples.

As test results suggest, differences in emotions are not statistically significant. Thus we concluded that the narrator's gender has no meaningful impact on the comments section emotion-wise.

Table 5.13: Emotion Comparison Test Results

Emotion Test	Test Statistic	p-value
Test Results 1	0.147	0.986
Test Results 2	1.431	0.698
Test Results 3	2.751	0.432
Test Results 4	0.394	0.941
Test Results 5	0.565	0.904

5.8 Title - Keyword Similarity

RQ9: Does the title - keyword similarity changes between dimensions?

To compute the similarity between video tags and video titles, we first got the embedding vectors of both the video titles and video tags using RoBERTa, then eliminated zero vectors that appeared due to lack of keywords, and finally calculated the cosine similarities of the remaining videos in our dataset. For the videos having more than one keyword, we calculated the similarity between the video title and each keyword and then averaged all the results we obtained.

The next step is to compare the similar results we obtained between the dimensions. For this purpose, we calculated the confidence intervals for each dimension and generated our hypotheses (Table A.11).

Considering the gender dimension, the cosine similarity is slightly higher in females. Similarly, when examining the query dimension, STEM has a slight edge compared to non-STEM. Lastly, between the Precovid and Postcovid periods, the difference is negligible.

Table 5.14: Similarity Test Results

Experiment	Result	Test Statistic	p-value
Male vs Female	Male < Female	-8.976	6.80E-19
STEM vs non-STEM	STEM > non-STEM	15.515	4.20E-49
Precovid vs Postcovid	Precovid > Postcovid	2.847	0.002

The test results indeed show that videos having female narrators have more keyword title similarity. The same conclusion can be told for the STEM videos. Interestingly, even though the magnitude is small, according to the statistical test results, the difference in keyword-title similarity between the periods is not negligible.

5.9 NER and Noun Analysis

RQ10: In terms of Named Entities and nouns, is there any outstanding word or phrase that gives a clue for us to understand the dynamics of gender bias?

In this part of our study, we wanted to incorporate more advanced NLP processes to detect gender bias. After applying the steps we discussed in Section 4.2.3, we created word clouds and observed both intersecting and differing words between the dimensions. For word analysis, we considered the most frequent 50 instances.

From our observations, we found that videos having male narrators are more likely to cover topics related to wars. The reason for this is, that some of the most frequently Named Entities under these videos are related to wars and they are only used under the videos having male narrators. Here are the words only found in the comments of the videos having male narrators:

- africa, **biden**, grant, **nato**, **nazi**, netflix, **putin**, python, roman, rome, scotland, **trump**, **ukraine**, **ukrainian**, **ww2**, **zelensky**

Not surprisingly, In STEM NER results, we found a bias towards company or scientist names and technical terms. The words only found in comments of the STEM videos are as follows:

- adam, **albert einstein**, alice, allah, **amazon**, **apple**, bob, calc, **calculus**, charlie, dave, earth, **einstein**, **feynman**, **google**, grant, greek, hindi, java, jim, **khan academy**, leonard, **linux**,**mac**, matt, **mit**, mosh, **newton**, oop, pakistan pc, plz, **python**, richard, **schrodinger**, **sql**, **ted**, **windows**, **youtube**

Besides these findings, during our analysis, we also observed that the word "India" appears so frequently in our word clouds.

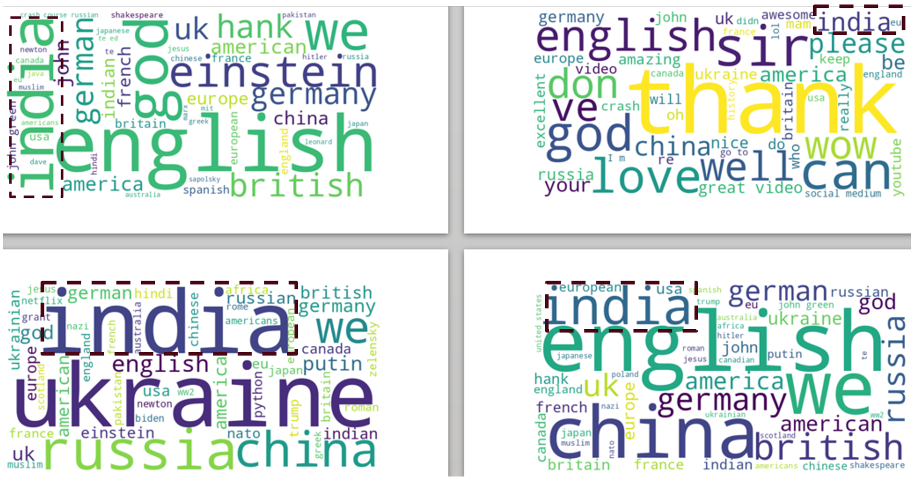


Figure 5.5: India in Various Word Clouds

We investigated the reason behind this situation by looking at the comments containing the word ”India” and manually reasoning these one by one. However, we couldn't find a definite answer to this question (Table A.12). The reasons have so much variety and therefore we can not categorize them and create a reasoning.

Another manual investigation we performed is about the word ”Ukraine”. For that purpose, we looked at all the videos having comments containing the root ”Ukra”. This approach allowed us to capture all the words about Ukraine. We found that almost all these comments are related to the recent Russia-Ukraine war (Table A.13).

5.10 The Impact of Covid-19

Up to this point, our studies followed the same procedure. Process and split the data and test the difference. Indeed, this is a valid approach because it allows us to use all the data points we have. In addition to that, we compared the statistics gender and query dimension in Precovid and Postcovid conditions. In this section, we focused on giving factual information about Precovid and Postcovid conditions.

Over the channels that had been established before the Postcovid period, we found the following,

567 channels in our dataset posted videos both in the Precovid and Postcovid periods. Among these,

- 234 channels published less, 136 channels published more videos in the Postcovid period. 197 of them on the other hand, published an equal amount of videos.
- On average 215 channels published shorter, 352 channels published longer videos in the Postcovid period.
- 146 channels produce both STEM and non-STEM videos. 198 channels produce only STEM videos and 223 channels produce only non-STEM videos.

There are 342 videos containing comments that exist in both periods in our dataset. Among these,

- 173 channels have a lower, 168 channels have higher mean comment length in the Postcovid period. Lastly, 1 channel has equal mean comment length in both periods.
- 174 channels have lower, 168 channels have higher mean character length in Postcovid period.

- 173 channels among these have lower, 168 channels have higher mean character length (spaces excluded) in the Postcovid period. Lastly, 1 channel among these has an equal mean character length in Postcovid and Precovid periods.

6
Conclusion & Future Work

In this thesis study, we investigated footprints about the presence of gender bias in online educational videos as well as how the COVID-19 pandemic impacted the video and comment characteristics. For this purpose, we first introduced 10 research questions and answered them by analyzing the video details of 19867 educational videos as well as their top 500 relevant or recent comments. Our analyses involve many data preprocessing steps, statistical tests, and sentimental analyses. Besides that, in some our our experiments we compared the rankings of the most frequently used emojis and words to detect the dynamics of behavioral change between the Precovid and Postcovid periods. Our efforts were concentrated on three key dimensions: gender of the video narrator, time, and query field.

Our research questions were formed in two parts. Video-based research questions emphasize more on the video details such as video duration, the number of likes and views a video receives, and video release rates. On the flip side, comment-based research questions delve more into comment details like comment lengths, word rankings, and sentimental differences. Aside from answering the research questions, we provided descriptive plots and analyzed the Named Entities and nouns.

From our analyses of video details, we found that videos started to become longer after COVID-19 became a part of our lives. Many long lecture videos contributed to this result **(RQ1)**. Although the videos become longer on average, the video release rates were not affected much by COVID-19 pandemic **(RQ2)**. We used the like-count ratio to gauge viewer engagement. Our results indicate that videos belonging to Female, non-STEM, and Postcovid dimensions receive more likes per view compared to the videos in opposite dimensions **(RQ3)**.

From the comment details, we concluded that regardless of how it is measured, comments of the videos in Male, Precovid, and non-STEM dimensions are longer compared to the Female, Postcovid, and STEM dimensions **(RQ4)**. The most frequently used emojis are highly common between videos having male and female narrators. On the other hand, words that have been used most differently are generally gender-specific **(RQ5)**. The dimensions of this thesis study are not closely related to emoji rankings. Our results indicate an approximately 20% similarity in each dimension pair **(RQ6)**. In terms of the relevancy of the recent comments, the results indicate that Female, non-STEM, and Postcovid dimensions outperformed Male, STEM, and Precovid dimensions **(RQ7)**. The polarity and emotion distribution does not change significantly with the gender of the narrator **(RQ8)**. Keyword-title similarities also change across the dimensions though the difference between the periods is small **(RQ9)**. From our NER and noun analysis, we detected a frequent use of the word ”India” but we couldn’t find a specific reason for this situation. The word ”Ukraine” is also frequently used due to the recent war between Ukraine and Russia **(RQ10)**.

To sum up, we found that videos are getting longer. More frequent release in long lecture videos contribute to this result so it can be concluded that education in digital education platforms improve over time. A higher viewer engagement in videos having female narrators indicates a positive inclination towards women. Moreover this behaviour is more pronounced in the Postcovid condition (Table A.7). The increase in viewer engagement did not reflected to the comment lengths. In other words, although viewers like the videos more, they do not represent their appreciation with longer comments. The higher relevancy in the recent comments for the videos having female narrators supports the interpretation we have done on viewer engagement. Lastly, the emotion and polarity distributions are not different in the gender dimension. Considering this result, together with the other ones, our findings do not support a negative tendency towards women in digital educational platforms. Conversely, viewers generally adopt either a neutral or somewhat positive attitude towards women.

6.1 Limitations and Future Work

In this study, we performed our analyses over details of 19867 videos and their respective comments. The dataset is not small yet it could be inefficient to see the whole picture because first, many of the videos in our dataset are not available any more on the YouTube platform and second, the number of videos belonging to male gender is significantly higher.

One of the motivations for this study is to detect the impact of COVID-19 on the video and comment characteristics. To do that, we considered the channels only present in both periods. The differences we found might be attributed to COVID-

19 but it is not guaranteed. COVID-19 is not the only novel thing that has been introduced to our daily life in the period this study investigates.

The model we used for our sentiment analysis ”Twitter RoBERTa Base for Sentiment Analysis” is fine-tuned for classifying the sentiments of the tweets. We used that model for YouTube comments. The mismatch between platforms can lead to many wrong classifications.

In future work, a more balanced and up-to-date dataset can be used for the same analysis. Also for sentiment classification, a more suitable model can be utilized.

Bibliography

[1] Reem Alkhammash. “It is time to operate like a woman: a corpus based study of representation of women in STEM fields in social media”. In: *International Journal of English Linguistics* 9.5 (2019), p. 217.

[2] Inoka Amarasekara and Will J Grant. “Exploring the YouTube science communication gender gap: A sentiment analysis”. In: *Public Understanding of Science* 28.1 (2019), pp. 68–84.

[3] David R Anderson, Dennis J Sweeney, and Thomas A Williams. *Essentials of statistics for business and economics*. Cengage Learning, 2020.

[4] Chittaranjan Andrade. “Sample size and its importance in research”. In: *Indian journal of psychological medicine* 42.1 (2020), pp. 102–103.

[5] Charles Kojo Assuah et al. “Walking Mathematics Students through the Maze of Chi-square Test of Independence and Homogeneity, Test Involving Several Proportions, and Goodness-of-fit Test”. In: *Asian Journal of Probability and Statistics* 18.4 (2022), pp. 22–35.

[6] Francesco Barbieri et al. “Tweeteval: Unified benchmark and comparative evaluation for tweet classification”. In: *arXiv preprint arXiv:2010.12421* (2020).

[7] James Dean Brown. “Yates correction factor”. In: *Shiken: JALT Testing & Evaluation SIG Newsletter* 8.1 (2004), pp. 22–27.

[8] Angela Dean and Daniel Voss. *Design and analysis of experiments*. Springer, 1999.

[9] Nicola Döring and M Rohangis Mohseni. “Gendered hate speech in YouTube and YouNow comments: Results of two content analyses”. In: *SCM Studies in Communication and Media* 9.1 (2020), pp. 62–88.

[10] Bradley Efron and Trevor Hastie. *Computer age statistical inference, student edition: algorithms, evidence, and data science*. Vol. 6. Cambridge University Press, 2021.

[11] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[12] Naomi Ellemers. “Gender stereotypes”. In: *Annual review of psychology* 69 (2018), pp. 275–298.

[13] Jorge Faber and Lilian Martins Fonseca. “How sample size influences research outcomes”. In: *Dental press journal of orthodontics* 19 (2014), pp. 27–29.

[14] Shereen Fouad and Ezzaldin Alkooheji. "Sentiment analysis for women in stem using twitter and transfer learning models". In: *2023 IEEE 17th international conference on semantic computing (ICSC)*. IEEE. 2023, pp. 227–234.

[15] Gizem Gezici and Yucel Saygin. "Measuring gender bias in educational videos: A case study on youtube". In: *arXiv preprint arXiv:2206.09987* (2022).

[16] Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*

[17] Sang Gyu Kwak and Jong Hae Kim. "Central limit theorem: the cornerstone of modern statistics". In: *Korean journal of anesthesiology* 70.2 (2017), pp. 144–156.

[18] David Lane et al. *Introduction to statistics.* Citeseer, 2003.

[19] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[20] Yingying Ma, Chenlei Leng, and Hansheng Wang. "Optimal subsampling bootstrap for massive data". In: *Journal of Business & Economic Statistics* 42.1 (2024), pp. 174–186.

[21] Katherine L Milkman, Modupe Akinola, and Dolly Chugh. "What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations." In: *Journal of Applied Psychology* 100.6 (2015), p. 1678.

[22] Prabhaker Mishra et al. "Application of student's t-test, analysis of variance, and covariance". In: *Annals of cardiac anaesthesia* 22.4 (2019), p. 407.

[23] Corinne A Moss-Racusin et al. "Science faculty's subtle gender biases favor male students". In: *Proceedings of the national academy of sciences* 109.41 (2012), pp. 16474–16479.

[24] Rahmatika Rahmatika, Munawir Yusuf, and Leo Agung. "The effectiveness of YouTube as an online learning media". In: *Journal of Education Technology* 5.1 (2021), pp. 152–158.

[25] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. "Semantic cosine similarity". In: *The 7th international student conference on advanced science and technology ICAST*. Vol. 4. 1. University of Seoul South Korea. 2012, p. 1.

[26] Shruti Raina. "Gender bias in education". In: *International Journal of Research Pedagogy and Technology in Education and Movement Sciences* 1.02 (2012).

[27] Rachel L Roper. "Does gender bias still affect women in science?" In: *Microbiology and Molecular Biology Reviews* 83.3 (2019), e00018–19.

[28] Tetsuya Sakai. "Two Sample T-Tests for IR Evaluation: Student or Welch?" In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '16. Pisa, Italy: Association for Computing Machinery, 2016, pp. 1045–1048. ISBN: 9781450340694. DOI: 10.1145/2911451.2914684. URL: https://doi.org/10.1145/2911451.2914684.

[29] Heather Sarsons. “Recognition for group work: Gender differences in academia”. In: *American Economic Review* 107.5 (2017), pp. 141–145.

[30] Kamran Siddiqui. “Heuristics for Sample Size Determination in Multivariate Statistical Techniques”. In: *World Applied Sciences Journal* 27 (Jan. 2013), pp. 285–287. DOI: `10.5829/idosi.wasj.2013.27.02.889`.

[31] Maite Taboada. “Sentiment analysis: An overview from linguistics”. In: *Annual Review of Linguistics* 2 (2016), pp. 325–347.

[32] Edna Chebet Too et al. “A comparative study of fine-tuning deep learning models for plant disease identification”. In: *Computers and Electronics in Agriculture* 161 (2019), pp. 272–279.

[33] William Webber, Alistair Moffat, and Justin Zobel. “A similarity measure for indefinite rankings”. In: *ACM Transactions on Information Systems (TOIS)* 28.4 (2010), pp. 1–38.

[34] Hamed Zamani and W Bruce Croft. “Estimating embedding vectors for queries”. In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. 2016, pp. 123–132.

A

Appendix

Video Durations

Table A.1: Summary Statistics of the Dimensions

Dimension	count	mean	median	min	max	std
Female	4641	1386.09	676.0	9.0	42827.0	2110.71
Male	9739	1726.61	774.0	15.0	61662.0	2796.57
non-STEM	8803	1566.14	749.0	3.0	42900.0	2118.97
STEM	9134	1698.65	759.0	5.0	61662.0	3088.43
Precovid	9709	1310.39	614.0	7.0	40933.0	2053.86
Postcovid	8228	2015.02	1001.0	3.0	61662.0	3186.81

Table A.2: Video Length Confidence Intervals

Dataset	95% CI
Male	(690.0, 897.525)
Female	(596.488, 775.0)
STEM	(669.5, 870.512)
non-STEM	(660.463, 873.037)
Precovid	(546.5, 687.037)
Postcovid	(871.0, 1199.55)

Table A.3: Top longest videos released before Covid-19

Video Name	Channel	Type
Pioneers of Science Full Audiobook by Oliver LODGE	Full Audiobooks	AUDIO BOOK
Calculus for Beginners full course	Academic Lesson	LECTURE
Object Oriented Programming (OOPs) Concepts In Java	Durga Software Solutions	LECTURE
AWS Certified Solutions Architect - Associate 2020	freeCodeCamp.org	LECTURE
Data Science Full Course - Learn Data Science in 10 Hours	edureka!	LECTURE
Machine Learning Full Course - Learn Machine Learning 10 Hours	edureka!	LECTURE
Statistics - A Full University Course on Data Science Basics	freeCodeCamp.org	LECTURE

Table A.4: Top 10 longest videos published (sorted by durations)

Video Name	Channel	Type
Database Systems - Cornell University Course	freeCodeCamp.org	LECTURE
Data Structures and Algorithms in Python	freeCodeCamp.org	LECTURE
Python for Data Science - Course for Beginners	freeCodeCamp.org	LECTURE
Modern Physics || Modern Physics Full Lecture	Academic Lesson	LECTURE
Longplayer Assembly 2020	Artangel	INTERVIEW
12 Hours Non-Stop Class | Maths Marathon by Dhasu Sir	wifistudy by Unacademy	LECTURE
Calculus 1 - Full College Course	freeCodeCamp.org	LECTURE
Full Course Image Processing and OpenCV	Ask It Loud	LECTURE
Biology 12 Hours Marathon Special Class By - Kajal Ma'am	Futurekul Coaching	LECTURE

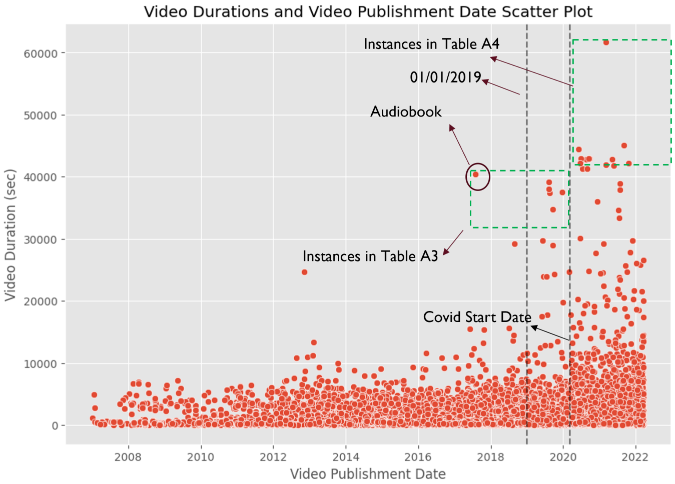


Figure A.1: Video Durations and Release Date 2

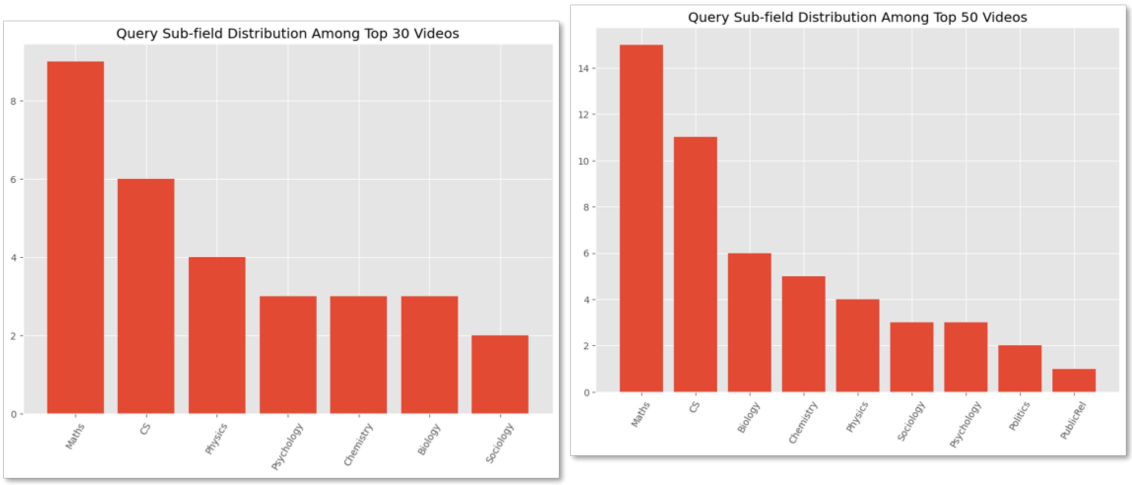


Figure A.2: Most Common Query Sub-fields Among 30 Videos (left) 50 videos (right)

Video Like - Video View Ratio

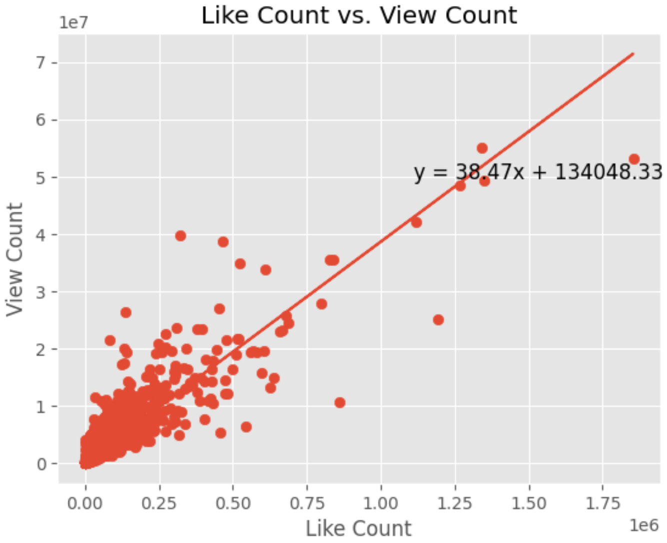


Figure A.3: Like Count vs View Count

Table A.5: Like-View Ratio CI

Dimension	95% CI
Male	(0.0192, 0.0203)
Female	(0.0202, 0.0217)
PreCovid	(0.0141, 0.0147)
PostCovid	(0.027, 0.0288)
STEM	(0.0192, 0.0206)
non-STEM	(0.0198, 0.0209)

Table A.6: Like-View Ratio Comparison Between Periods

Dimension	Precovid CI	Postcovid CI	Accepted Hypothesis
STEM	(0.015, 0.017)	(0.024, 0.026)	Precovid < Postcovid
non-STEM	(0.016, 0.017)	(0.02, 0.024)	Precovid < Postcovid
Male	(0.015, 0.016)	(0.022, 0.023)	Precovid < Postcovid
Female	(0.016, 0.018)	(0.023, 0.028)	Precovid < Postcovid

Table A.7: Time Equated Like-View Ratio Comparison Between Periods

Dimension	Precovid CI	Postcovid CI	Accepted Hypothesis
STEM	(0.02, 0.022)	(0.024, 0.027)	Precovid < Postcovid
non-STEM	(0.019, 0.021)	(0.02, 0.022)	Precovid < Postcovid
Male	(0.019, 0.021)	(0.022, 0.023)	Precovid < Postcovid
Female	(0.02, 0.023)	(0.023, 0.027)	Precovid < Postcovid

Comment Length Analysis

Table A.8: Comment Length Experiment Confidence Intervals

Dimension	**95% CI**
Male Sentence Length	(23.605, 29.716)
Female Sentence Length	(20.322, 27.975)
Precovid Sentence Length	(20.233, 28.775)
Postcovid Sentence Length	(20.116, 25.690)
non-STEM Sentence Length	(24.629, 31.807)
STEM Sentence Length	(17.869, 22.876)
Male Character Length	(131.822, 167.284)
Female Character Length	(112.673, 156.236)
Precovid Character Length	(123.067, 162.385)
Postcovid Character Length	(111.443, 143.425)
non-STEM Character Length	(137.435, 178.518)
STEM Character Length	(98.092, 128.381)
Male Character Length without Spaces	(108.725, 138.026)
Female Character Length without Spaces	(92.857, 128.286)
Precovid Character Length without Spaces	(101.654, 133.821)
Postcovid Character Length without Spaces	(91.371, 118.226)
non-STEM Character Length without Spaces	(113.428, 147.226)
STEM Character Length without Spaces	(81.000, 105.691)

Table A.9: Comment Length Change Comparison

Dimension	**Measure**	**Precovid CI**	**Postcovid CI**	**Accepted Hypothesis**
Male	Sentence Length	(23.25, 30.84)	(23.11, 30.11)	Precovid > Postcovid
	Character Length	(130.11, 174.23)	(128.75, 169.43)	Precovid > Postcovid
	Character w.o. ’ ’ *	(107.43, 143.63)	(106.26, 139.7)	Precovid > Postcovid
Female	Sentence Length	(19.81, 29.77)	(19.79, 28.27)	Precovid > Postcovid
	Character Length	(110.32, 167.59)	(109.37, 157.71)	Precovid > Postcovid
	Character w.o. ’ ’	(91.01, 137.84)	(90.13, 129.83)	Precovid > Postcovid
STEM	Sentence Length	(18.63, 25.43)	(17.1, 22.56)	Precovid > Postcovid
	Character Length	(103.41, 143.51)	(94.37, 126.23)	Precovid > Postcovid
	Character w.o. ’ ’	(85.37, 118.48)	(77.89, 104.12)	Precovid > Postcovid
non-STEM	Sentence Length	(24.92, 34.09)	(23.66, 31.71)	Precovid > Postcovid
	Character Length	(139.77, 192.55)	(131.66, 177.66)	Precovid > Postcovid
	Character w.o. ’ ’	(115.62, 159.0)	(108.58, 146.29)	Precovid > Postcovid

* ”Character Length w.o. ’ ’” stands for Character Length without Spaces.

Rank Analysis

Table A.10: RBO Extrapolated Comment Ranking Results

Dimensions	**95% CI**
Male	(0.389, 0.463)
Female	(0.425, 0.493)
Precovid	(0.186, 0.247)
Postcovid	(0.364, 0.431)
STEM	(0.353, 0.421)
non-STEM	(0.381, 0.453)

Title - Keyword Similarity

Table A.11: Cosine Similarity Confidence Intervals

Dimension	**95% CI**
Male	(0.912, 0.916)
Female	(0.913, 0.917)
STEM	(0.913, 0.918)
non-STEM	(0.912, 0.916)
Precovid	(0.913, 0.917)
Postcovid	(0.913, 0.917)

NER and Noun Analysis

Table A.12: Manual Labeling for the Comments Containing India

Comment	Manual Labels
You are always wrong and will ...	Comment about the topic
Hello ,I have Master's of Anal...	Commentary about degree from India
Sir, Can you help me \nI have ...	Question about India
Your videos will keep remainin...	The lack of education in India
B.Sc degree in Data Science fr...	Commentary about degree from India
india is not next china \n\nin...	Commentary about the future of India
wth it is war...India said it ...	Comment about India and war
Hello everyone, how was the vi...	Link contains the word India
My father also suffer from sch...	Commentary (Indicating Location)
In India every lecturer should...	The lack of education in India
Please start these type of ser...	The lack of education in India
ye nafsiyat kya h bhai..please...	Request for language translation
Disagree. \nEconomic advanceme...	Commentary about topic
Great video. I can't help but ...	Commentary about education in India

Table A.13: Comments Containing the Root Ukra

Comments	Manual Label
Every Weapon We Give Ukraine I...	Russia - Ukraine War
Typical of the modern left'...	Russia - Ukraine War
Ukraine's guerrilla warfare...	Russia - Ukraine War
Abhijit Iyer-Mitra Explains ...	Russia - Ukraine War
The Realist View of Ukraine/...	Russia - Ukraine War
The Untold Story Of Volodymy...	Zelensky Biography
The War in Ukraine Could Cha...	Russia - Ukraine War
Ukraine War: 'We want to be ...	Russia - Ukraine War
Ukraine and Russia: What Cau...	Russia - Ukraine War
Ukraine's Civilians Take Up ...	Russia - Ukraine War
Ukrainian President Zelensky...	Russia - Ukraine War
War in Ukraine – and What I...	Russia - Ukraine War
War in Ukraine: Zelenskyy te...	Russia - Ukraine War
Watch Joe Biden's Full Speec...	Russia - Ukraine War