St. Joseph's University

Bengaluru, Karnataka - 560027

Project on Advanced Statistical Methods

Submitted by:

Jason Alvin Chesney (222BDA03)

Yobah Bertrand Yonkou (222BDA15)

Alvina Joanna (222BDA63)

Saba Santhosh (222BDA65)

Submitted to:

Jayati Kaushik

Assistant Professor

Department of Advanced Computing

St. Joseph's University

Aim

We live in a world where we can easily be anonymous in the digital space. This breaks down barriers and enables people to express themselves on online social spaces. Controversies usually garner a lot of interest and people feel the urge to participate in conversations that surround these controversies. In our project, we aimed to build a machine learning model that can predict the sentiment (either for/against the topic) of a comment said regarding a certain controversial topic. This type of analysis can help with checking the general public opinion on a certain topic or issue.

Domain

Sentiment Analysis, Machine Learning.

Problem statement

Performing sentiment analysis on comments regarding a controversial topic.

Introduction

Our thought process going into the project was to first find a topic that was controversial: flat earth vs. terra ball, moon landing faked or real, blue bubbles vs green bubbles in messages, and many more. All of these were equally interesting, but the nature of these comments were not what we expected. The general sentiment of the majority of the comments weren't clearly for/against, and that was pretty interesting. Many people seemed to thank the content creator for the information, others were just ranting regarding the issue or making fun of something totally unrelated, and the occasional random comments (i.e. trolls). Then we had to have at least 1000 comments on the video, but YouTube has a counting mechanism that includes replies in the total count but scraping those replies isn't an easy feat. Our last hurdle was having enough for and against comments so as to not bias our training model.

After much back and forth, we finally selected on a video titled "Three Biblical Questions For Fans Of The Chosen | Joseph Smith, Dallas Jenkins, Pope Francis" by Wretched. The speaker in the video details three main reasons why he does not support a Christian web-TV series titled "The Chosen" and poses those thoughts as questions to the fans of the show. Our analysis was based on the main comments of this video.

Literature Survey

The increase in the popularity of social media as an integral part of our lives have been a significant cause of the increase in research on sentiment analysis (Evaluating the effectiveness of Text preprocessing in Sentiment Analysis). The following are some related work done on sentiment analysis of YouTube comments.

[13] Alhujaili et. al carried out research to identify sentiment analysis methods and techniques that can be used on YouTube content. The approaches here were explained and categorised and are useful for research on data mining and sentiment analysis. After experimenting on multiple datasets (like twitter, YouTube, facebook,...), it was discovered that Machine Learning and Deep Learning techniques tend to have high accuracies in sentiment analysis. Machine learning methods used here include, but are not limited to Naive Bayes, SVM, and KNN.

[12] Furthermore, a research on the Classification of YouTube data based on sentiment analysis Bamane et. al made use of the Naive bayes algorithm, following the standard sentiment analysis approach. The Naive Bayes classifier was trained on a set of opinions derived from YouTube comments and using it to determine the sentiment of other comments (in the test set). They stored positive and negative opinions in separate dictionaries and calculated the polarity of each word by calculating the number of times it appears in the positive and negative dictionaries. The method used here considers comments as independent words and does not consider their ordering.

[14] Chongtham Rajen Singh and R. Gobinath presented in their paper an approach to sentiment analysis by examining the climate and population in accordance with economic growth to derive a statistical hypothesis. They also added a subset of the data using predefined seeding features and rules before adding the best performing supervised learning model to predict the unlabeled tweets. Chongtham Rajen Singh and R. Gobinath then labelled tweets as ‘believer’ or ‘denier’ for each country and established a hypothesis testing based on the reviews by rich and poor countries.

The statistical analysis presented in the paper demonstrates a positive correlation between the GDP growth rate and the number of deniers and believers in each country. They go on to explain in detail description of the techniques used in their experimental setup along with their findings and statistical analysis.

Hypothesis

We chose the following hypothesis:

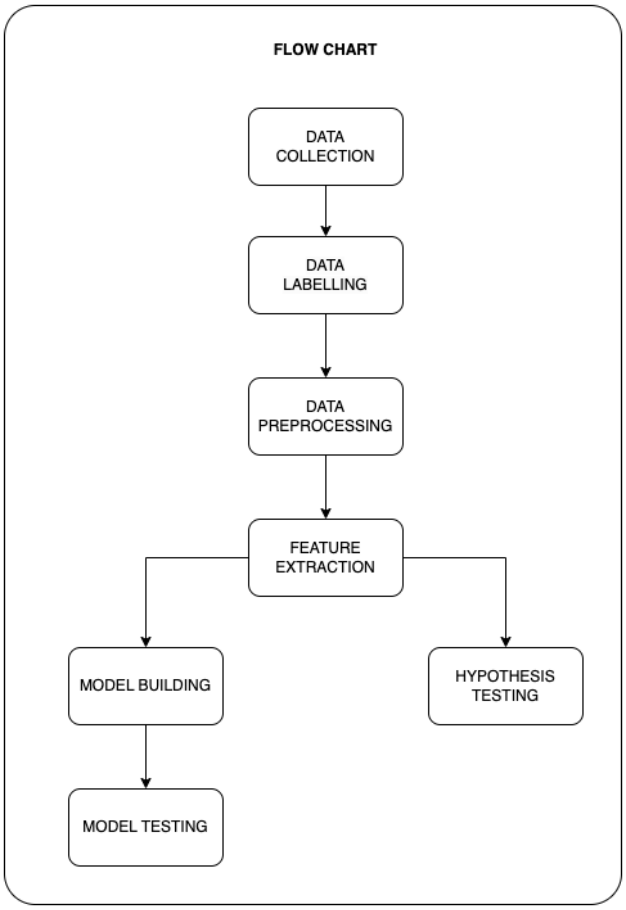
$$H_0: \ Number\ of\ likes\ for\ 'for'\ comments \ = \ Number\ of\ likes\ for\ 'against'\ comments$$

$$H_0: \ Number\ of\ likes\ for\ 'for'\ comments \ \neq \ Number\ of\ likes\ for\ 'against'\ comments$$

In the context of analysing likes on for/against comments, we used ANOVA to determine whether there is a statistically significant difference in the mean number of likes and positive/negative sentiment scores (referring to for/against comment sentiment scores), i.e, the comments liked by viewers.

Methods and Materials

Our project followed standard sentiment analysis steps (as mentioned in the flow chart) and made use of popular Python packages in this domain.



1. Data collection:

   We collected 803 YouTube comments centred around a video on *three Biblical questions for fans of the chosen show*, presented by Mr. Todd Friel on his channel 'Wretched'. The main comments (not including the replies) of this video were scrapped using the WebHarvy tool. Along with the comments, we scraped the name of the user, when the comment was posted (we called it time since the video was posted), number of likes, and the number replies.

2. Data labelling:

   As performing sentiment analysis is a supervised learning algorithm, we had to manually label our dataset. The comments were labelled into one of two sentiments; For or Against, based on a set of well defined rules. Comments that were not related to the video were flagged as drop. This process took up to 2 days to complete.

3. Data preprocessing:

   The following operations were performed (using Python) on the dataset to make it ready for EDA and modelling:

   - *Dropped unwanted data*: Comments flagged as drop during the data labelling phase were dropped as well as the 'replies' column as more than 80% was missing.
   - *Extracting month & year from time:* The time of posting for a comment on YouTube increases from seconds to days, to weeks, to months and lastly to years. The actual timestamp when a comment was posted is not mentioned anywhere. Time is an important factor in visualising trends (like change in the number of comments for each sentiment over time). We built a custom function to extract the month and year from the time given by default.
   - *Character replacement:* The characters, “, ”, ‘, and ’ were replaced with ", ", ' and ', respectively. This is because “, ”, ‘, and ’ are not special characters, rather they are considered as regular text.
   - *Word spelling correction:* Comments were checked, word by word for misspelt words using a Python library called SpellChecker. The misspelt words were later replaced with the best likely suggested word. For example, 'Chrch' was converted to 'Church'.
   - *Expansion of contractions:* Contractions such as “isn't” were expanded to their original words such as “is not”. This is because contractions introduce inconsistency in the dataset as the computer would interpret “isn't” and “is not” as having two separate meanings. On the other hand, they would increase the dimension of the dataset as “isn't” and “is not” will be considered as two separate features (Dealing with Contractions in NLP).
   - *To lowercase:* Here, every letter is converted to lowercase. Just like contractions, the computer would interpret the words “Teach” and “teach” as two different features. Hence, converting all letters to lowercase will reduce

the dimensionality of our dataset (A review: preprocessing techniques and data augmentation for sentiment analysis).

- *Removal of extra spaces:* The normal number of spaces between words is one. Here, we removed extra spaces from each comment. Extra Spaces could introduce inconsistency in our dataset as " teach" and "teach" would be interpreted as different features.
- *Lemmatization:* This is the process of converting a word to its root form. Lemmatization normalises the words in our dataset by transforming them to their lemma or dictionary form.

4. Feature Extraction:

   A document Term Matrix (precisely, Term Frequency-Document Inverse Frequency) was used to represent comments into a numerical format that can be used for further analysis and model building. Here, the occurrence of a word in different document corpus is analysed (A beginners guide to EDA).

5. Modelling building:

   The document term matrix was split into train and test datasets in the ratio 80:20, respectively. Then, the naive bayes model, precisely, MultinomialNB was used to train a model on the train dataset. Furthermore, the test dataset was used to evaluate our model.

6. Hypothesis testing:

   After framing our null hypothesis, we used ANOVA to test our hypothesis.

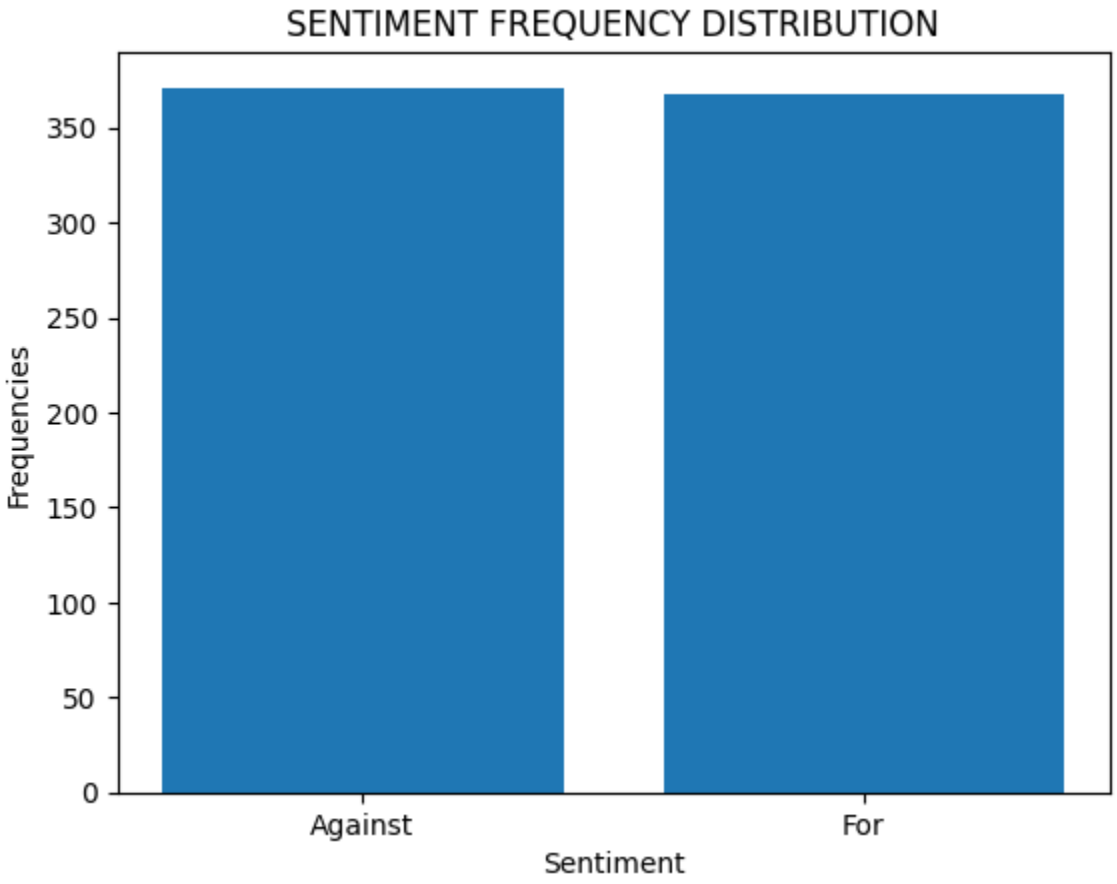
Study Design

Our study design for sentiment analysis using natural language processing involved collecting a large dataset of text data from sources, such as YouTube reviews. The goal was to analyse the sentiment expressed in the text, classifying it as positive or negative. We scraped data such as name, comments, likes, number or replies to comments and time and then used NLP techniques to preprocess the text, such as removing stop words contradictions, lemmatization, before applying a machine learning algorithm, such as a Naive Bayes classifier, to classify the sentiment. Then we performed the exploratory data analysis on the cleaned data. The results

Exploratory Data Analysis

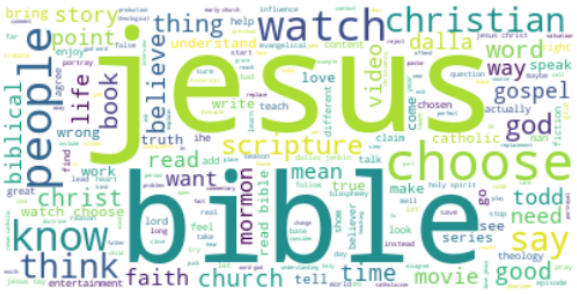
We conducted an EDA on the comments acquired in order to understand the data further.

Following is a distribution of sentiments in the entire dataset. This showed that our dataset was fairly well balanced.



Analysed the word clouds of the cleaned words from the ‘for’ and ‘against’ comments to gain an insight into the most frequently utilised words in the comments.

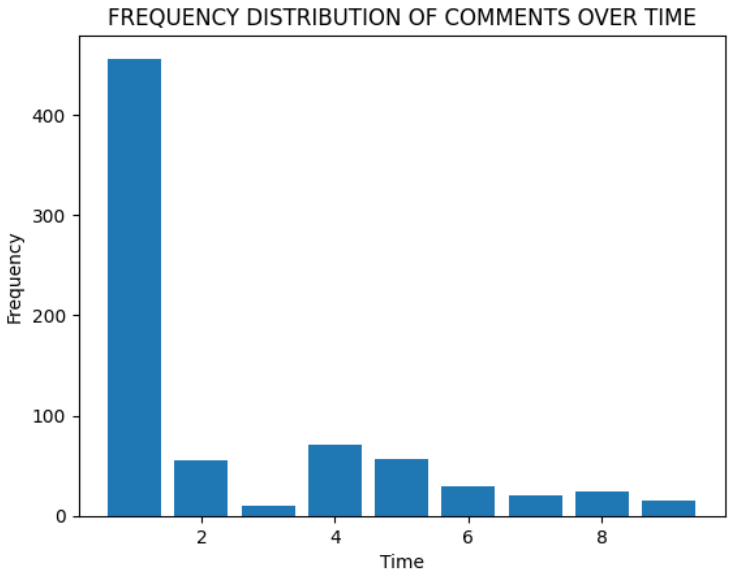
Word cloud for FOR comments



Word cloud for AGAINST comments

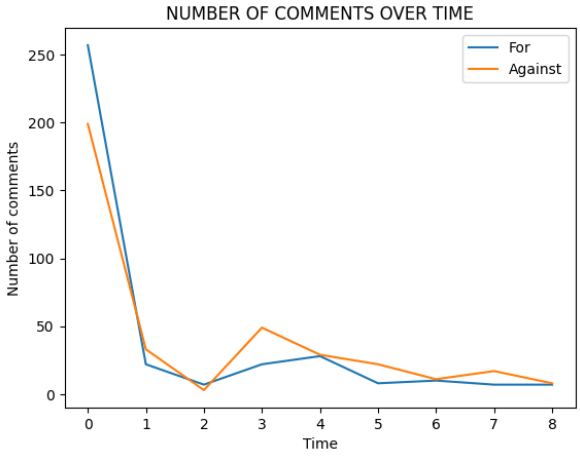
Both of the above show that the main words are similar in nature.

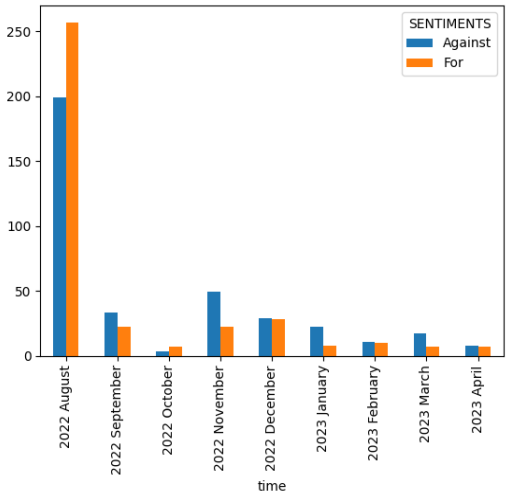
A simple frequency distribution of all comments from the time that the video was posted shows that the most comments were made in the very month of the video being posted. Then a natural decline over the following three months. A slight increase is observed in the fourth month from posting (i.e. November 2022) and this can be attributed to increased attention towards the show due to the promotional activities and release of the third season in theatres (November 18, 2022).

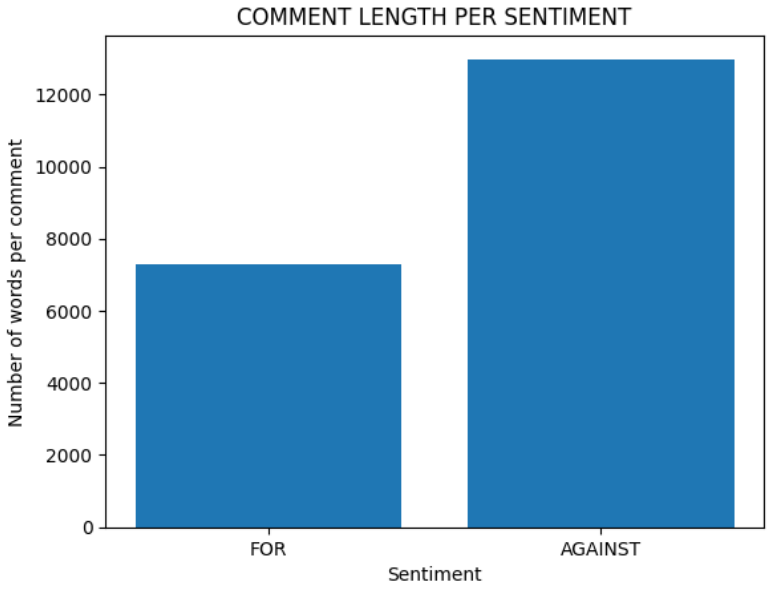


The following two charts visualise the break-up of the number of comments of time for each sentiment. The number of ‘for’ comments are in the lead only in the initial month of the video and two months hence. Whereas, the ‘against’ comments are

consistently more than the ‘for’ comments since then. This leads us to conclude that those generally not in favour of the main arguments of this video were more vocal about it over time.

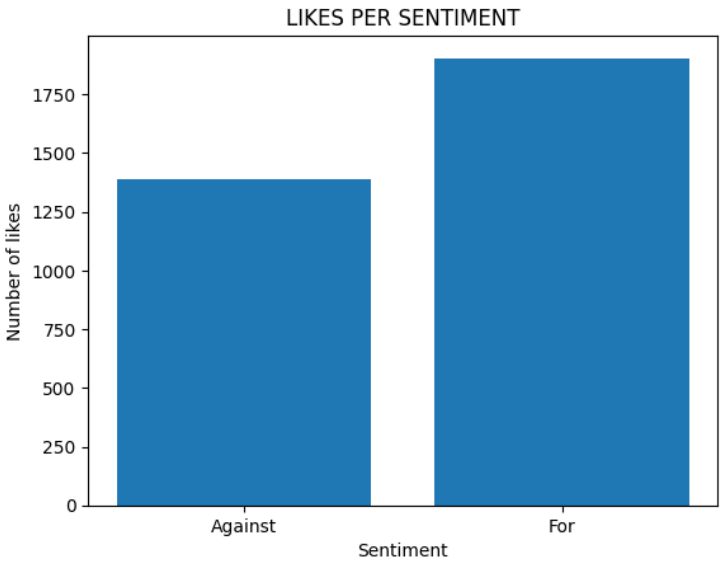




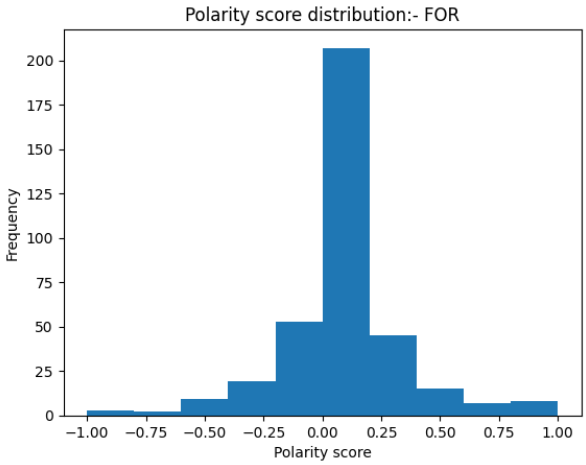


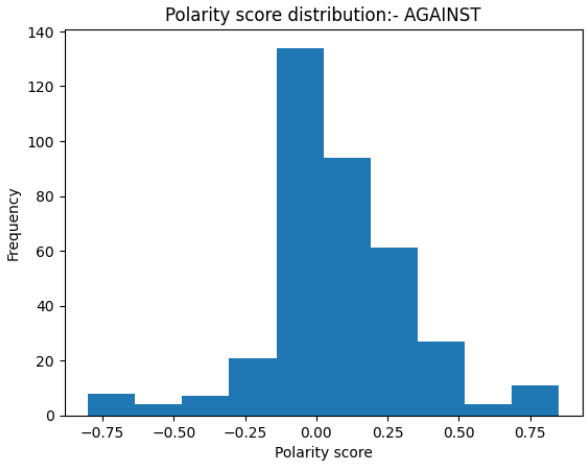
An interesting insight was to see that the ‘against’ comments contained way more words than the ‘for’ comments. This can be explained by the nature of many ‘for’ comments that simply stated their agreement, while ‘against’ comments went into much detail regarding their disagreement and their reasons for disagreement.

The like per comment was analysed to see the strength of support received for either sentiment. We see that the 'for' comments received more likes than the 'against' comments (even though our dataset included an almost equal number of either class of comments). From this, we can infer that people are more likely to express their agreement for a statement that is favourable to the main argument.

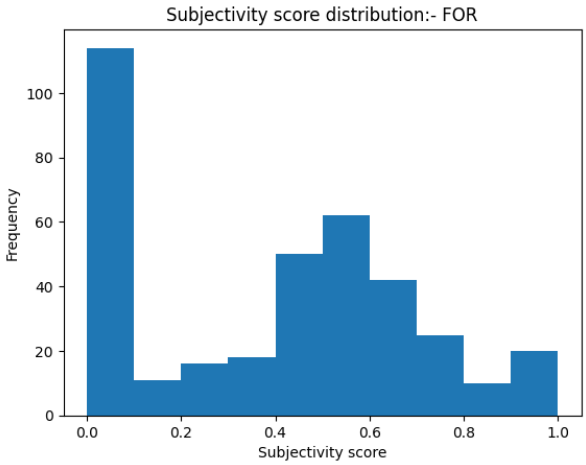


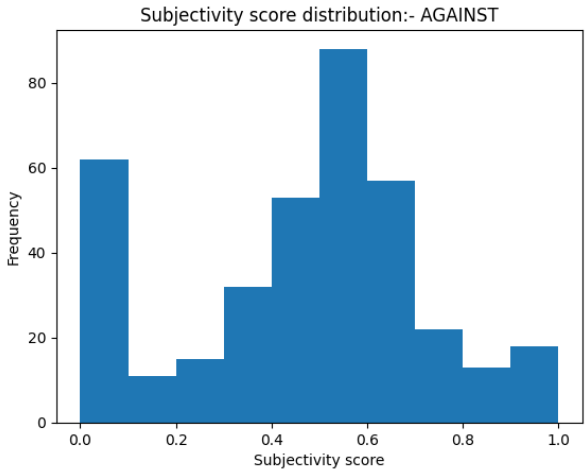
"Polarity refers to the overall sentiment conveyed by a particular text, phrase or word. This polarity can be expressed as a numerical rating known as a *sentiment score*." The below two charts show the polarity scores for either sentiment. We see that the 'against' comments had a more diverse distribution of polarity scores in comparison to the 'for' comments. This could be because many of the 'against' comments weren't clear-cut against, but expressed a broad explanation for their reasoning and even their gratitude in some cases for this content creator's questions.





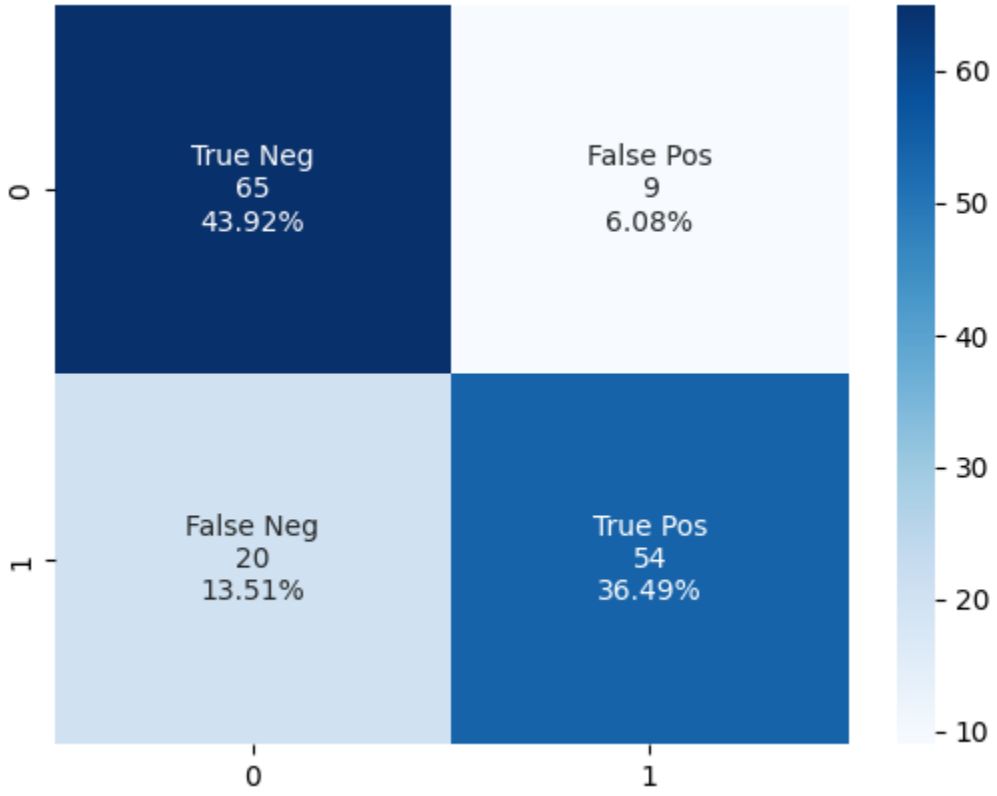
The subjectivity score as provided by the TextBlob package indicates the 'subjectivity' (in other words, how likely a statement is to be a personal opinion) of a comment. The two charts expressing the subjectivity scores for the two sentiments show that the 'for' comments were less likely to be personal opinions with a majority of the comments getting a zero or close to zero score, while the 'against' comments were more likely to be a personal opinion, with most of these comments lying in the midsection of 0.3 - 0.7 score.





Results

Metrics on the performance of the model



Confusion matrix for the model

	precision	recall	f1-score	support
Against	0.88	0.76	0.82	85
For	0.73	0.86	0.79	63
accuracy			0.80	148
macro avg	0.80	0.81	0.80	148
weighted avg	0.82	0.80	0.81	148

Classification report of the model

Hypothesis Testing

When conducting a hypothesis test using ANOVA, the null hypothesis is that there is no significant difference between the means of the groups being compared, and the alternative hypothesis is that there is a significant difference. In our case, since the p-value is 0.001 which is smaller than 0.05, we rejected the null hypothesis and concluded that there is a significant difference between the means. The f-statistic of 11.389 indicates the ratio of variance between the groups to variance within the groups, and the larger the value of the f-statistic, the more likely it is that there is a significant difference between the means. The combination of a small p-value and a large f-statistic provides strong evidence that there is a significant difference between the means of the groups being compared, which supports the rejection of the null hypothesis.

Conclusion & Inferences

In this project, we performed sentiment analysis on the comments of the YouTube video “Three Biblical Questions For Fans Of The Chosen”. Our process involved pre - processing of the data and EDA in order to better understand the dataset. This revealed that our data was balanced. By doing the word cloud for the cleaned dataset, it was interesting to see few of the common words that were present in both “for” and “against” classification. Through our entire process, we found that the comments on social media websites usually contain polarising views and opinions regarding the topic of discussion and this can be an interesting indicator of the general public sentiment of the topic. Controversial topics are the best way to study and understand this phenomena.

Discussion & future work

Our current model classifies comments being either "for" or "against" the main argument in the video we chose. A potential expansion of this could be to obtain a transcript of any video - to then analyse it in order to identify the main arguments. This could then be used to create a threshold to classify the comments on the video as either for/against.

A potential business perspective could be for video streaming platforms to implement a new filter option to view the comments. This new filter would have categories like for/against/neutral. The implementation of this filter will help people find others with the same viewpoints as themselves without having to scroll through countless irrelevant comments.

Key learning

Our team had the opportunity to focus on the detailed process of building an NLP model. From choosing a video, scraping the comments, labelling (while keeping an objective perspective), cleaning the text data, building the model all the way to the exploratory data analytics and hypothesis testing was a first time experience for a few and a great revision for the rest. We were exposed to the various packages available for different aspects of the project along with learning the nature of algorithms we were using. Hypothesis testing was a truly testing section as we thought long and hard regarding what kind of hypothesis would be suitable for an NLP project. We're truly grateful to have received this opportunity to work as a team on this project as it has truly enhanced our knowledge and skills in critical areas required for our future as data scientists.

Link to code and dataset (GitHub link)

https://github.com/yobahBertrandYonkou/sentiment-analysis.git

References

1 Preprocessing: https://ijettcs.org/Volume1Issue2/IJETTCS-2012-08-14-047.pdf

2 Straight vs Curly Quotes affecting search: https://discuss.elastic.co/t/straight-vs-curly-quotes-affecting-search/92911/3

3 Dealing with contractions in NLP: https://medium.com/@lukei_3514/dealing-with-contractions-in-nlp-d6174300876b

4 A review - preprocessing techniques and data augmentation for sentiment analysis: https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-020-00080-x

5 Extra spaces and non printable characters: https://www.linkedin.com/pulse/extra-spaces-non-printable-characters-problems-abdul-rahman-sherzad/

6 A beginners guide to EDA: https://www.analyticsvidhya.com/blog/2020/04/beginners-guide-exploratory-data-analysis-text-data/

7 Polarity and subjectivity in sentiment analysis: https://www.quora.com/What-is-polarity-and-subjectivity-in-sentiment-analysis https://getthematic.com/sentiment-analysis/#:~:text=Sentiment%20Scoring&text=Polarity%20refers%20to%20the%20overall,with%200%20representing%20neutral%20sentiment

8 Checking subjectivity scores for the sentiment categories: https://towardsdatascience.com/twitter-sentiment-analysis-in-python-1bafebe0b566

9 Reference for how the null hypothesis could be formulated: https://link.springer.com/chapter/10.1007/978-981-19-2600-6_7

10 Checking the dates of the release of season 3: https://www.deseret.com/faith/2022/11/16/23452209/the-chosen-theatrical-release-s eason-3

11 Checking the different approaches to perform sentiment analysis: https://www.researchgate.net/publication/351351202_YouTube_COMMENTS_SENTIMENT_ANALYSIS

12 Classification of youtube data based on sentiment analysis https://www.irjet.net/archives/V7/i5/IRJET-V7I5875.pdf

13 Sentiment analysis for youtube videos with user comments. https://sci-hub.se/10.1109/ICAIS50930.2021.9396049

14 Hypothesis testing of tweet text using NLP https://link.springer.com/chapter/10.1007/978-981-19-2600-6_7

Start Date: *April 21st, 2023*.

End Date: *April 26th, 2023*.