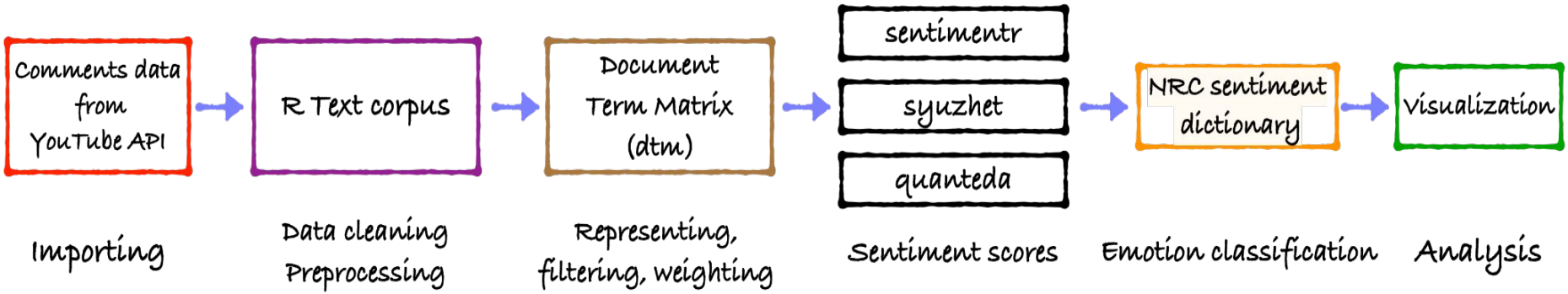
YOUTUBE COMMENTS SENTIMENT ANALYSIS USING R

Susmitha Marripalapu (A20489531)
Swetha Radhakrishnan (A20460652)
Aastha Dhir (A20468022)

INTRODUCTION

- Various social media posts and comments in recent years are used to understand the feedback and opinions of the general people.
- Sentiment analysis is used to identify the opinions of the general public and can be used to take important business decisions.
- Polarity, or positive or negative expressions, facts, or objective expressions, and views, or people's perspectives, are some crucial characteristics that can be retrieved from the study.
- In this project, we'll use data from YouTube video comments to examine polarity and influence created by eight videos using Sentimentr, Syuzhet and Quanteda (Naive Bayes classifier) packages from R.
- The final performance of the model is measured by plotting a confusion matrix to understand the accuracy of each method.

PROJECT OVERVIEW



Order of operations for sentiment analysis

DATA SETS

- YouTube comments data on 8 videos to perform sentiment analysis and emotional classification. This is imported into the R console using the **YouTube Data API and the R 'tuber' package .**
- This data includes the video ID, comments corresponding to each video ID, the author of the comment, the author's channel, the rating, and the comment's published timestamp.
- **Kaggle YouTube statistics data** to evaluate the accuracy and compare different approaches. This data is loaded manually into the R interface as a CSV file and is used to assess accuracy. The dataset that we are using is https://www.kaggle.com/datasets/advaypatil/youtube-statistics

PROPOSED METHODOLOGY

STAGE 1

Preprocess data and perform sentiment analysis and emotional classification on from 8 YouTube videos comments and present the findings.

STAGE 2

Compare the performance of sentiment analysis R packages by using the kaggle labeled youtube statistics data.

SAMPLE YOUTUBE API DATA VIEW

Raw data from youtube API with all the special characters, stopwords and punctuation marks.

videoId	Comment
pZy8115sNXM	supper ❤❤
pZy8115sNXM	Wonderful! Kisses from Brasil!
pZy8115sNXM	Amazing
pZy8115sNXM	3:30. What an epic drop!
pZy8115sNXM	calculation.... perfection... they say u can't attain perfection... right there.. right there... pin point precision ... perfection.. staring right at ur face.
pZy8115sNXM	my fav song ❤🎶
pZy8115sNXM	Love this Any TELUGU fans attendence plzzz ❤️❤️❤️❤️❤️❤️❤️❤️❤️❤️❤️❤️
pZy8115sNXM	3:29 oh girl's you made it.. reaction and energetic! So adora
pZy8115sNXM	How many peaples are watching in 2022
pZy8115sNXM	BRUTAL

SAMPLE KAGGLE DATA VIEW

Below is the picture of the sample Kaggle data with labelled sentiment scores.

comments

	Video ID	Comment	Likes	Sentiment
0	wAZZ-UWGVHI	Let's not forget that Apple Pay in 2014 required a brand new iPhone in order to use it. A significant portion of Apple's user base wasn't able to use it even if they wanted to. As each successive iPhone incorporated the technology and older iPhones were replaced the number of people who could use the technology increased.	95.0	1.0
1	wAZZ-UWGVHI	Here in NZ 50% of retailers don't even have contactless credit card machines like pay-wave which support Apple Pay. They don't like the high fees that come with these.	19.0	0.0
2	wAZZ-UWGVHI	I will forever acknowledge this channel with the help of your lessons and ideas explanations, Now It's quite helpful while you'll just sit at your comfort and monitor your account Growth.	161.0	2.0

DATA CLEANING

- Removing spurious characters, duplicates and special characters from the data.
- Removing common words which are not informative in nature helps in reducing the size of the data and also increases the computational efficiency.
- Removing whitespaces and punctuations.
- ”tm” package in R is used for data cleaning

DOCUMENT TERM MATRIX

- Document term matrix represents how often different words are used in a document.
- Use a word cloud to visualize and analyze qualitative data. The size of each word indicates how often it was used in the comments.
- A word cloud with different colors representing different emotions is shown in the next slide. The different colors represent different emotions, such as happy, sad, angry, and scared.

	word <chr>	freq <dbl>
peopl	peopl	1282
need	need	1272
student	student	1075
good	good	1054
make	make	1022
use	use	1017
work	work	982
love	love	861
tech	tech	834
much	much	833

WORD CLOUD

Word cloud with Afinn lexicon

-> The bing lexicon categorizes words in a **binary fashion** into positive and negative categories.

<- The AFINN lexicon assigns words with a score that runs between **-5 and 5**, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

Word cloud with bing lexicon

POLARITY OF VIDEOS

-> Scaling polarity scores of all lexicons.

-> Finding mean polarity of a video.

-> Sentiment scores are classified based on below:

Less than zero as **negative**
Equal to zero as **neutral**
Greater than zero as **positive**

-> Packages used:

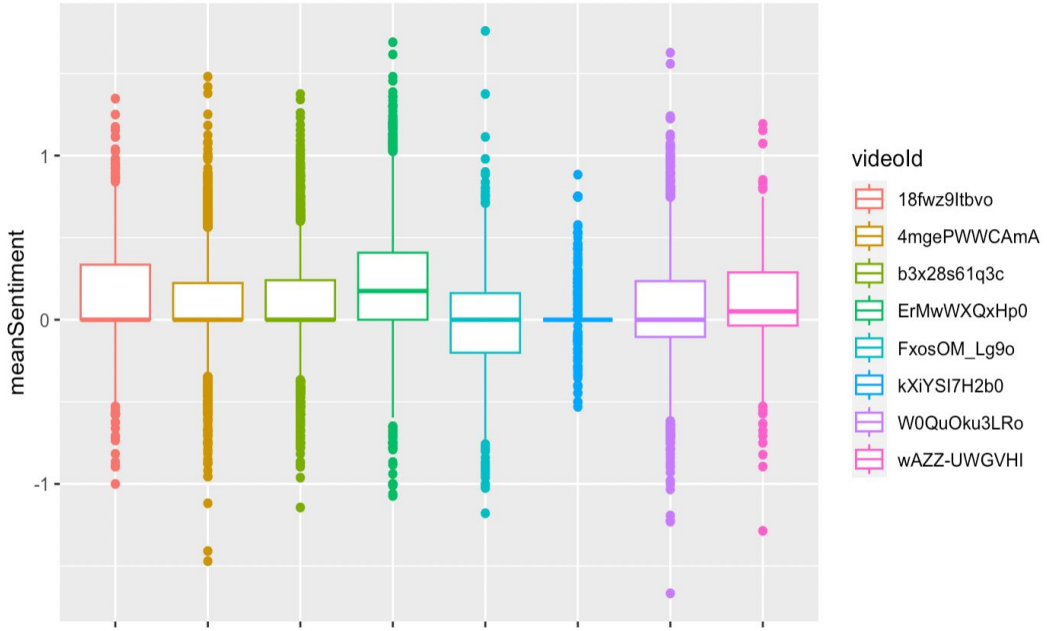
Sentimentr,
Syuzhet and
Quanteda (Naive Bayes classifier)

videoId <chr>	videoSentimentr <dbl>	videoSyuzhetSentiment <dbl>	videoBingSentiment <dbl>	videoAfinnSentiment <dbl>
18fwz9ltbvo	0.13421490	0.35465612	0.26475959	1.0596470
4mgePWWCAmA	0.06982809	0.20234662	0.20497597	0.5660164
b3x28s61q3c	0.07618748	0.22649945	0.14525627	0.6538713
ErMwWXQxHp0	0.20409775	0.63102410	0.74763339	1.8836059
FxosOM_Lg9o	-0.01101395	-0.03920530	-0.24105960	-0.2450331
kXiYSI7H2b0	0.01711570	0.03434191	0.01210287	0.1338880
W0QuOku3LRo	0.05838271	0.25737327	-0.08433180	0.1400922
wAZZ-UWGVHI	0.10478785	0.34442289	0.19592629	0.4820563

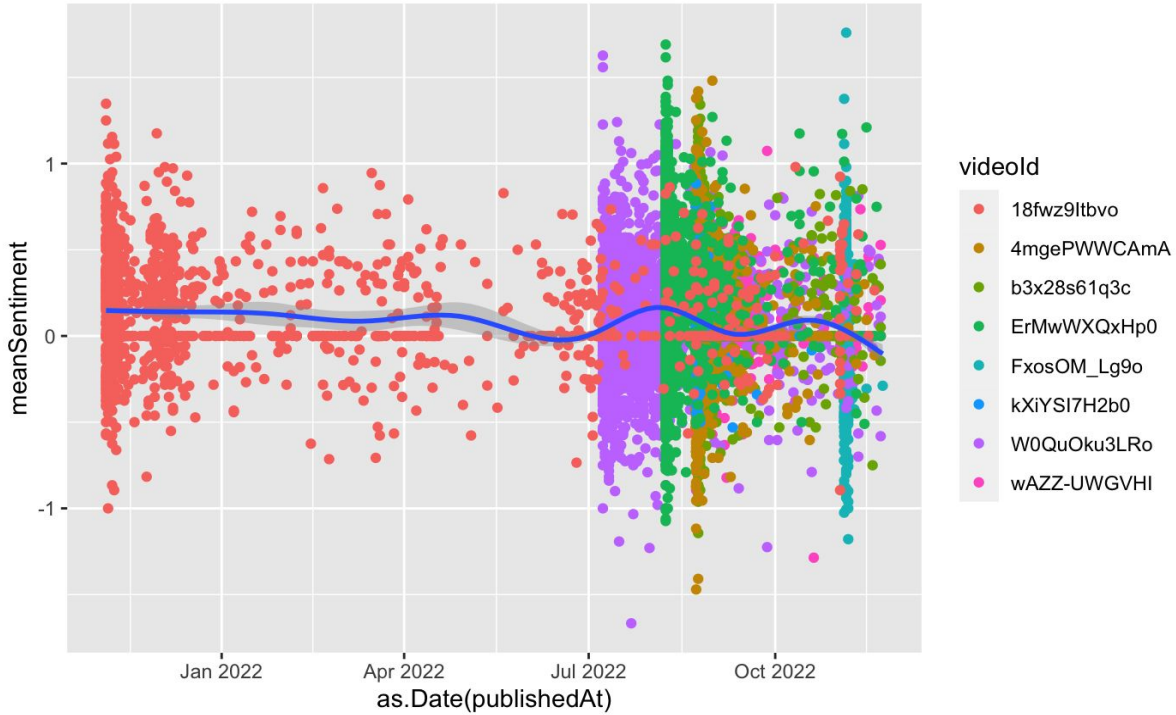
BOX PLOT FOR VIDEO SENTIMENT SCORES

Observations:

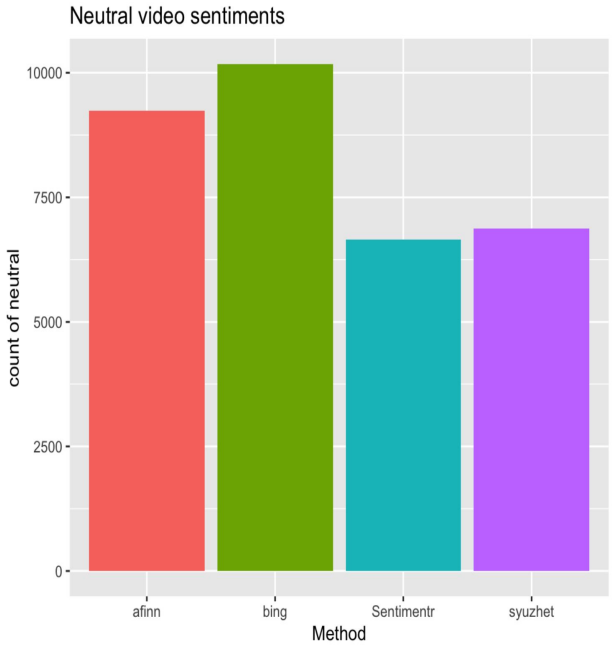
- Majority of videos, have neutral or positive sentiments.
- One out of 8 videos has negative sentiment (like "FxosOM Lg9o")
- This box plot is using sentimentr scores.

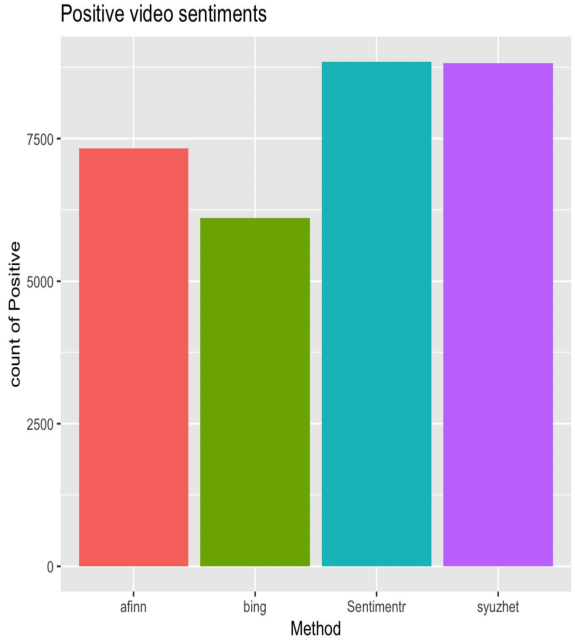


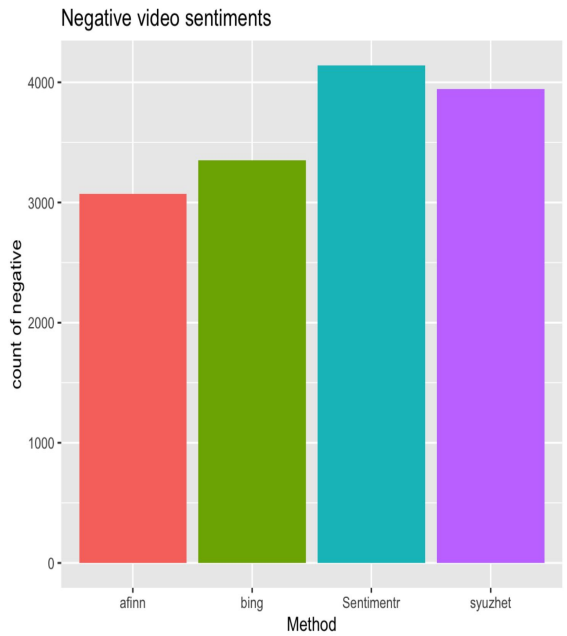
SENTIMENT OF COMMENT OVERTIME



SENTIMENT CLASSIFICATION COMPARISON







EMOTION CLASSIFICATION OF NEGATIVE SENTIMENT VIDEO

The video “FxosOM Lg9o” has overall negative sentiment.

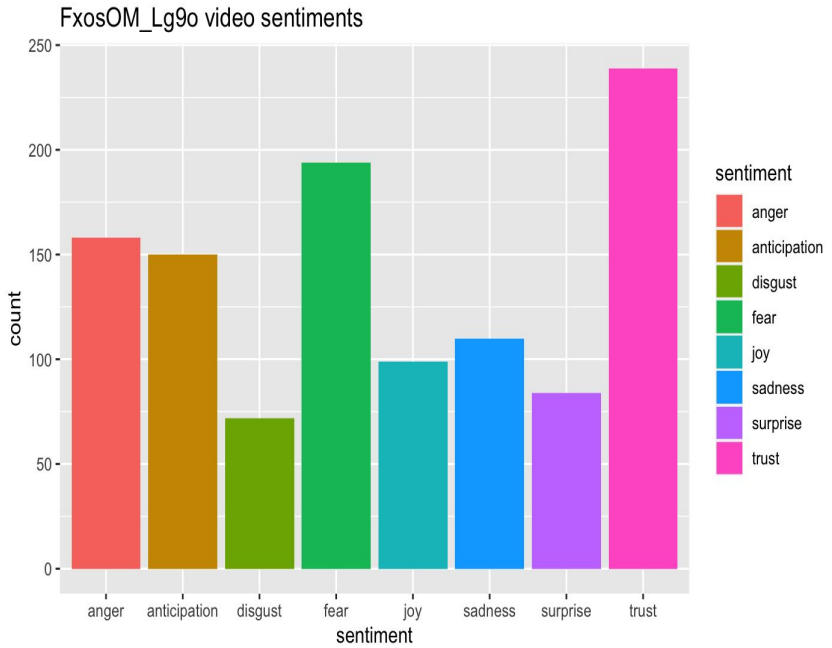
Majority of negative emotions include angry and disappointed, sadness and frustration.

NRC sentiment function from the syuzhet package is used to get the emotion classification.

anger <dbl>	anticipation <dbl>	disgust <dbl>	fear <dbl>	joy <dbl>	sadness <dbl>	surprise <dbl>	trust <dbl>	negative <dbl>	positive <dbl>
0	1	0	1	0	0	0	0	1	0
1	1	0	3	1	1	1	1	4	3
4	1	2	2	2	1	1	2	3	3
0	0	0	0	0	0	0	0	0	0
2	1	2	5	1	1	3	1	5	2
1	0	0	1	0	0	0	0	1	0
2	0	0	3	0	0	0	1	3	0
0	3	0	2	0	1	1	2	3	0
0	0	1	0	0	0	0	1	1	1
1	0	1	1	0	0	1	0	1	0

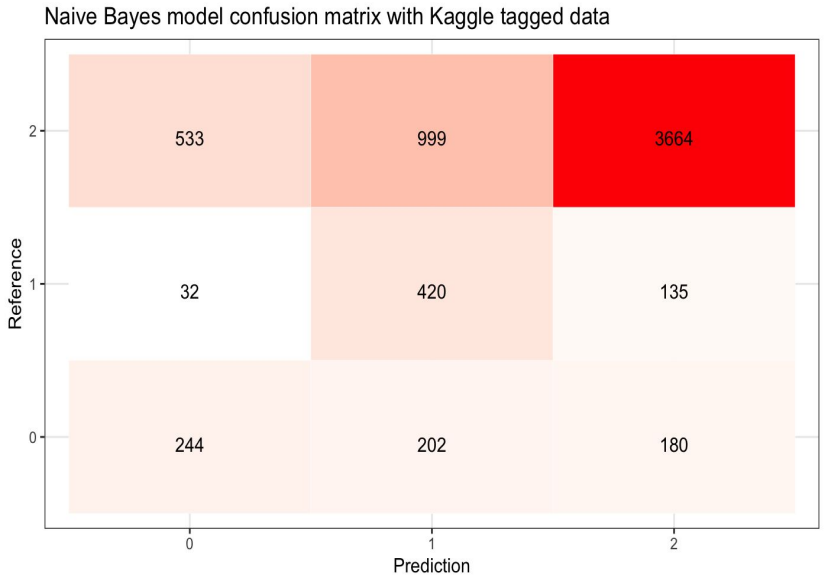
BAR GRAPH OF COUNT OF EMOTIONS

The graph shows that the sum of negative emotions (like anger, disgust, fear, and sadness) is higher than the sum of positive emotions (like trust, happiness, joy, and love).

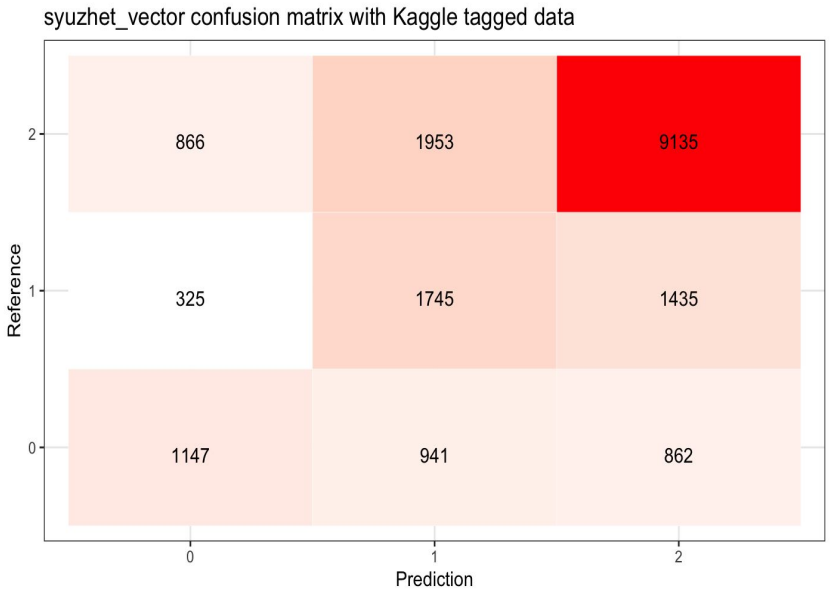


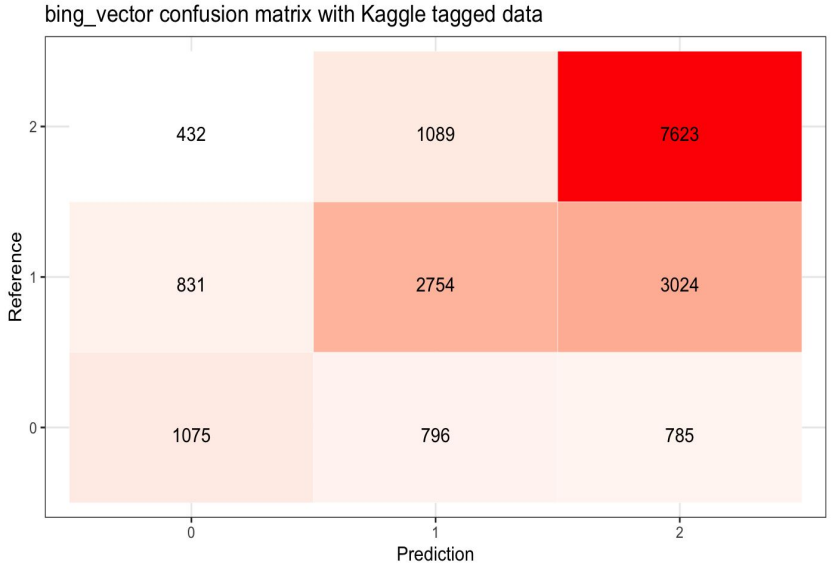
ACCURACY COMPARISON

- In confusion matrix plot,
  negative sentiments are represented as 0,
  Neutral sentiments are represented as 1,
  Positive sentiments are represented as 2.
- The red color is the most intense when maximum sentiment scores of a specific category are matched.
- We found that bing and afinn perform better than kaggle data in accuracy.
- The naive Bayes model is the most accurate in predicting the results of a test when used with data. It is able to achieve 67% accuracy.
- Train test split taken was 60:40

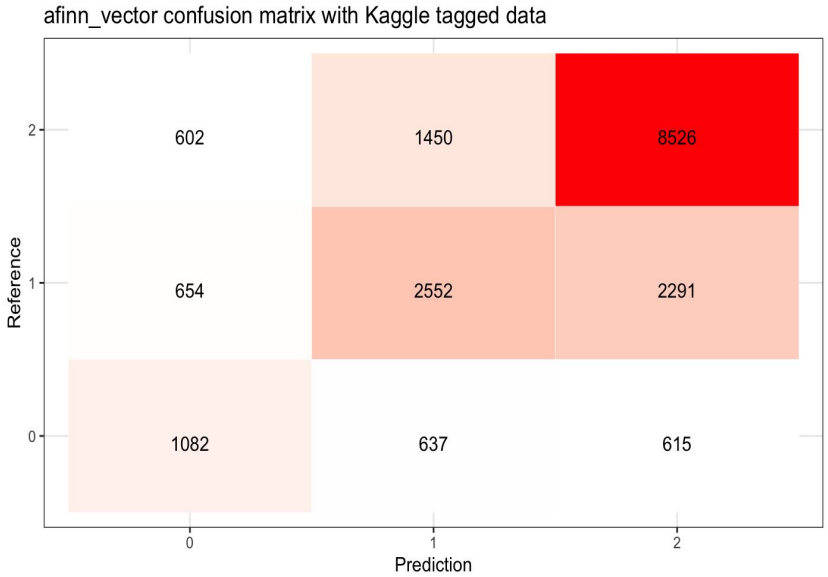


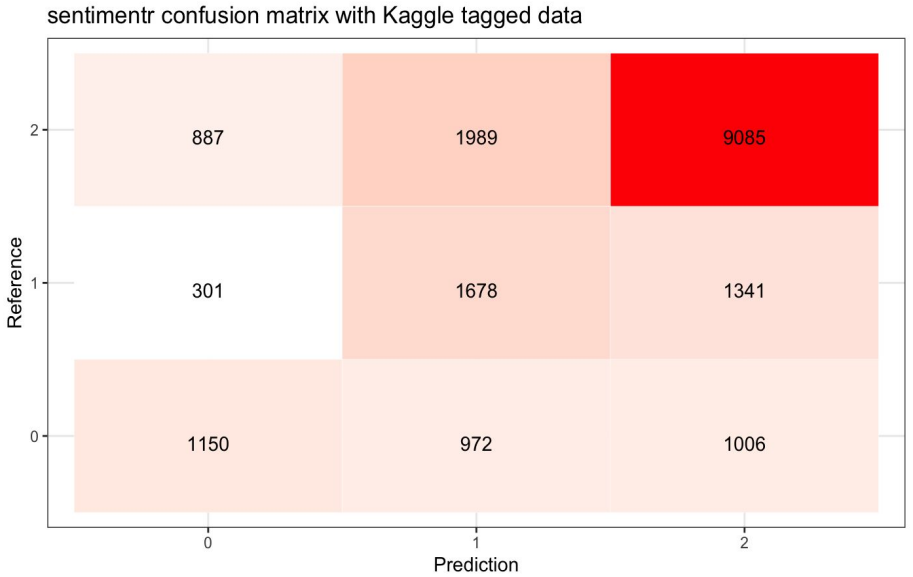
SYUZHET_VECTOR AND BING_VECTOR CONFUSION MATRIX





AFFIN_VECTOR AND SENTIMENTR CONFUSION MATRIX





CONCLUSION

- Text and sentiment analysis is a new field of research that has a lot of potential in data science. There are many other programs like Meanr and Stanford that can be used for text analysis, but they are not the focus of this project.
- In this project, we have reduced number of dimensions in the text data and normalized it so it's in a consistent form. With the help of sentiment scores and emotion classification, we were able to analyze video comments and visualize the results.
- The model that achieved the highest accuracy was a naive Bayes model with 67 percent. The other models had accuracies between 62 and 66%. By removing unnecessary words from data, we can improve the model accuracy.
- The Kaggle dataset has majority of the comments with positive polarity scores.. This could have lead to the model being biased in its predictions.
- This project gives fundamental implementation steps for text analysis and emotion classification using R and can also be implemented on other texts like Twitter data.

REFERENCES

- Text mining with r
-https://www.red-gate.com/simple-talk/databases/sql-server/bisql-server/text-mining-and-sentiment-analysis-with-r/
- Naldi, Maurizio. "A review of sentiment computation methods with R packages." arXiv preprint arXiv:1901.08319 (2019)
- Wankhade, M., Rao, A.C.S., Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1
- Brady D. Lund (2020) Assessing library topics using sentiment analysis in R: a discussion and code sample, Public Services Quarterly, 16:2, 112-123, DOI: 10.1080/15228959.2020.1731402
- Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, Volume 5, Issue 4, 2014, Pages 1093-1113, ISSN 2090-4479, https://doi.org/10.1016/j.asej.2014.04.011.
- NRC Emotion lexicon – http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- Syuzhet - https://github.com/mjockers/syuzhet
- Lexicon - https://github.com/trinker/lexicon
- Import youtube data- https://medium.com/@sohamsp1995/sentiment-analysis-on-youtubedata-using-r-df4021193d1e
- Kaggle dataset - https://www.kaggle.com/datasets/advaypatil/youtube-statistics
- R references for sentiment analysis - https://www.kaggle.com/code/rtatman/tutorialsentiment-analysis-in-r/notebook
- sentimentr - https://github.com/trinker/sentimentr
- https://console.cloud.google.com/apis/library?pli=1project=weather-pwa-c9eb8
- https://www.researchgate.net/profile/Vinod-Shukla/publication/334754287 Sentiment Analysis on Twitter Data using R/links/5e9720c692851c2f52a41b2a/SentimentAnalysis-on-Twitter-Data-using-R.pdf

Thank You