South Park Dialogue Analysis

BANA 275: Natural Language Processing

Team 4: Drew Boyd, Minji Jeon, Supriya Shahane, Kathleen Sebastian, and Tatiksha Singh

PRESENTATION AGENDA

01

PROJECT INTRODUCTION

02

EXPLORATORY DATA ANALYSIS

03

MODELING APPROACH & DEMO

04

CHALLENGES, KEY FINDINGS & CONCLUSIONS

01 PROJECT INTRODUCTION

PROJECT OBJECTIVES

Conduct **sentiment analysis** on the South Park script and generate a **character prediction model**

DATASET OVERVIEW

SOURCE: Kaggle

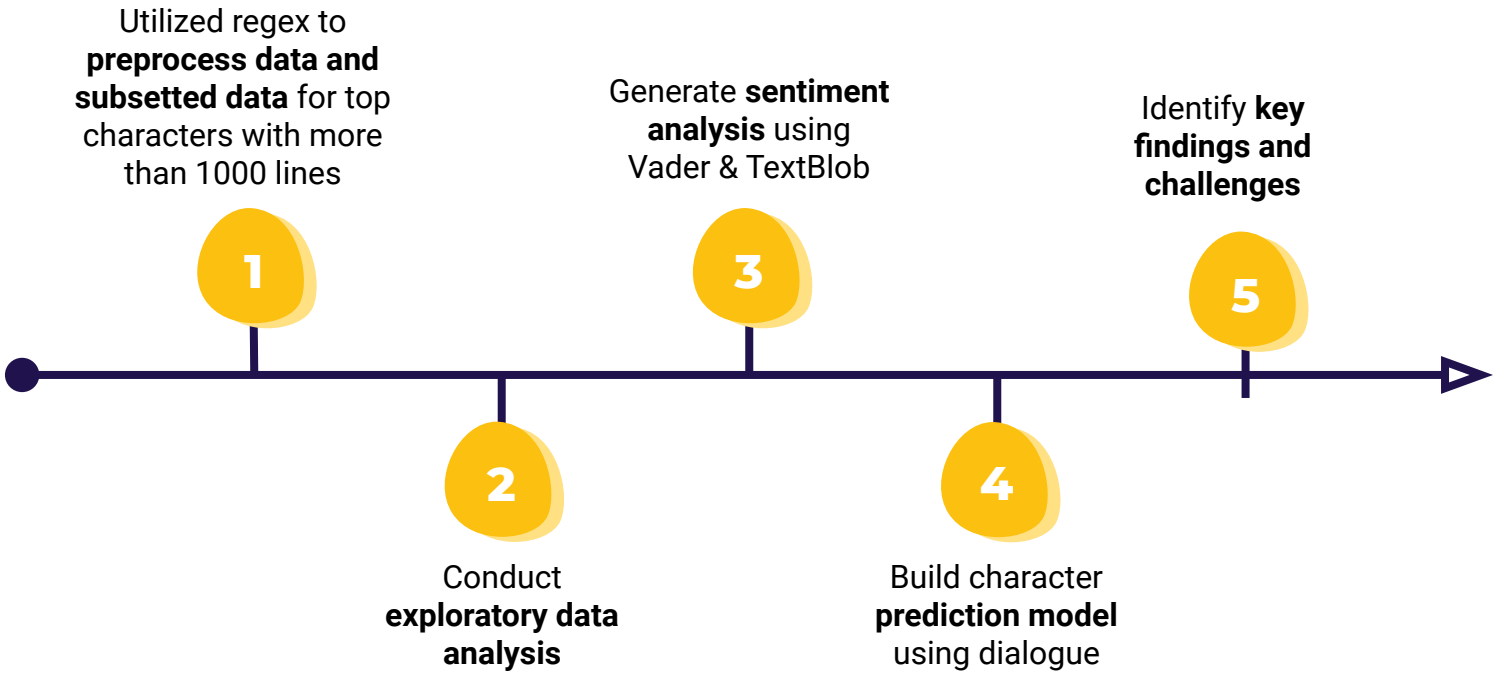
DATASET SIZE: 5.28 MB

CONTEXT: The dataset consists of 70K+ lines from the dialogue/script of Seasons 1-18 from the popular tv show, South Park.

COLUMN NAMES:

- Season
- Episode
- Character
- Line

METHODOLOGY



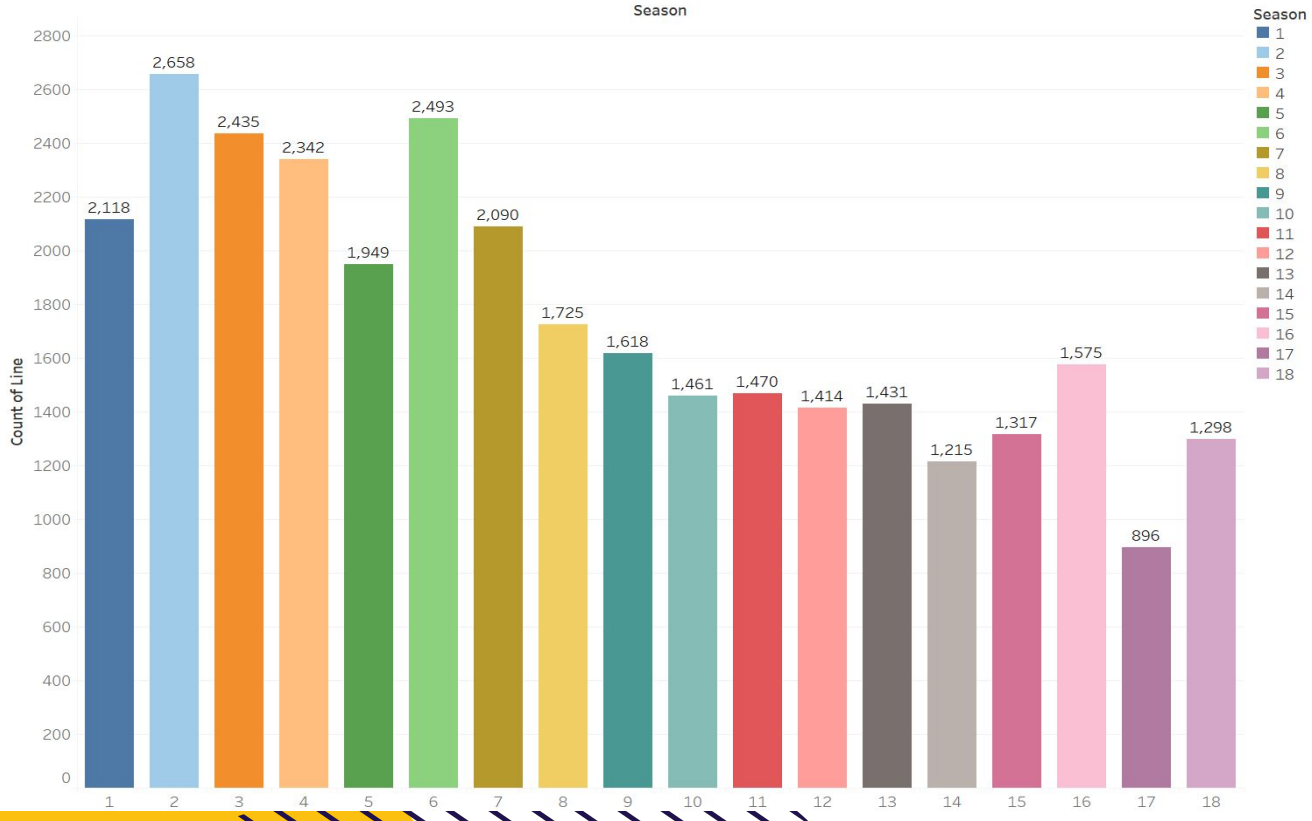
02 EXPLORATORY DATA ANALYSIS

WORDCLOUD: ALL SEASONS

COUNT OF LINES EACH SEASON

South Park Seasons by Count of Lines

Seasons 1-18



CHARACTER SHARE OF WORDS

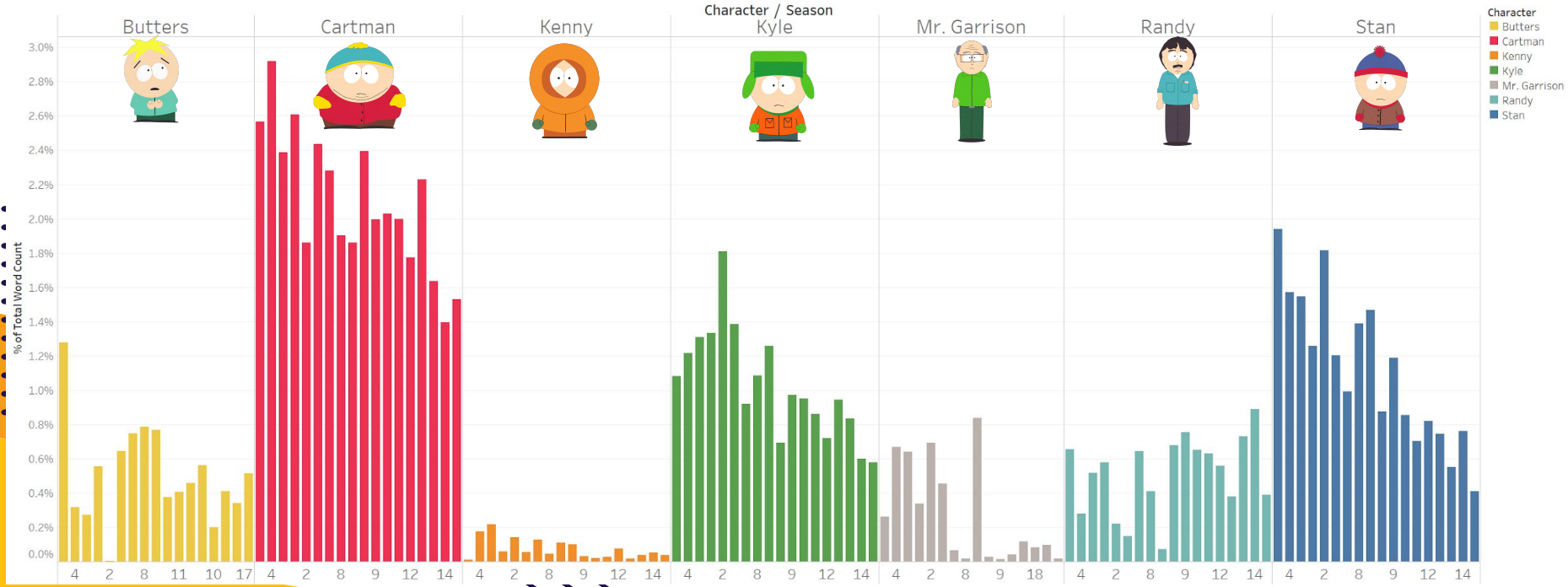
South Park Characters Share of Words

Seasons 1-18

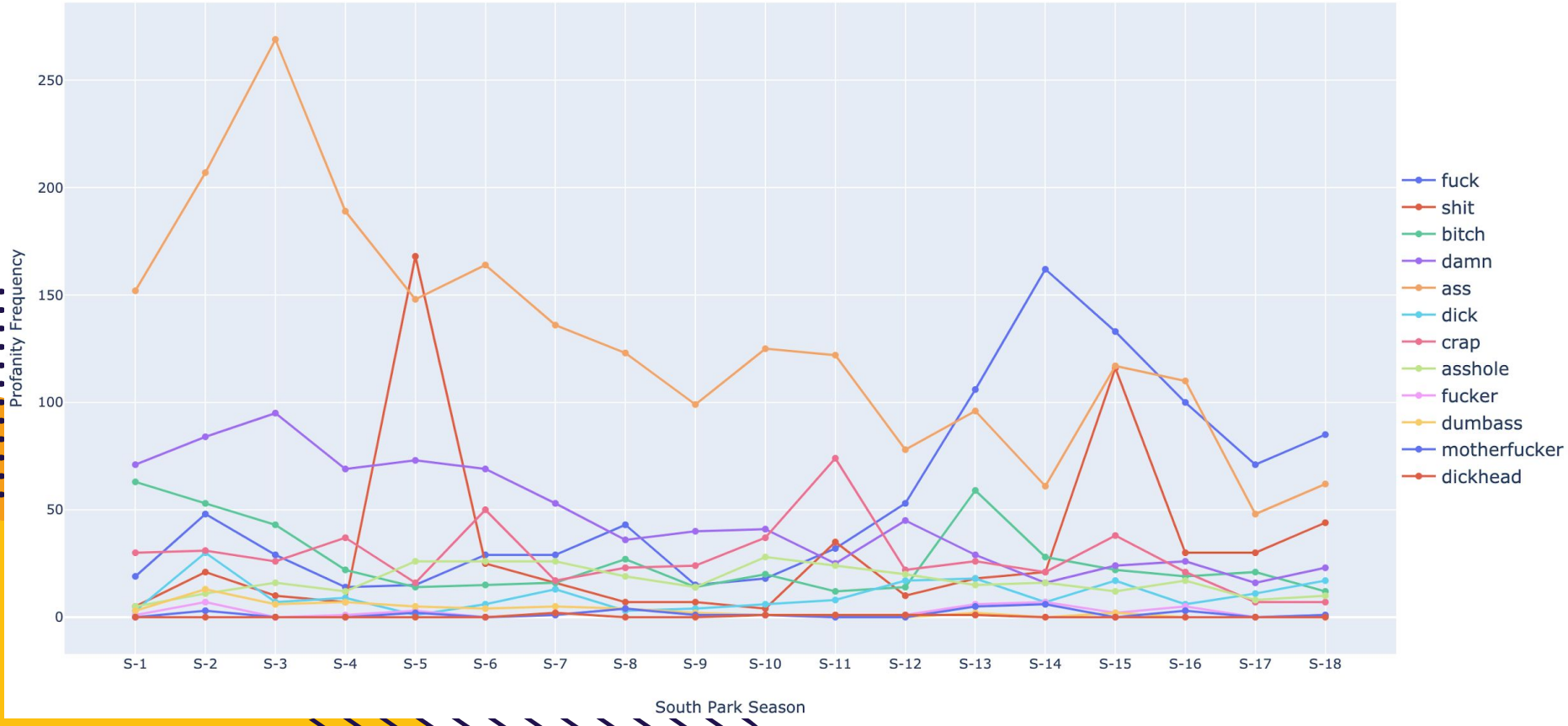
CHARACTER SHARE OF WORDS

South Park Characters Share of Words by Seasons

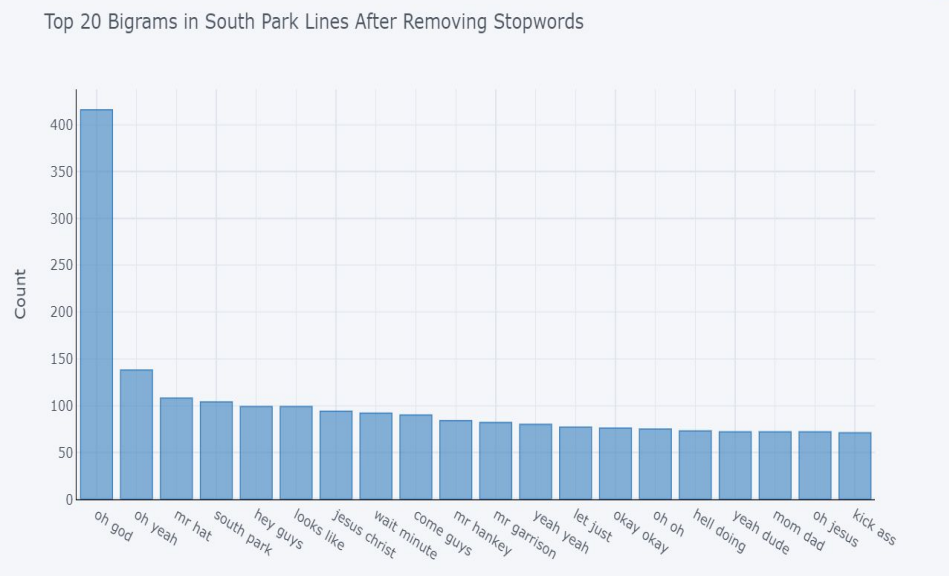
Seasons 1-18

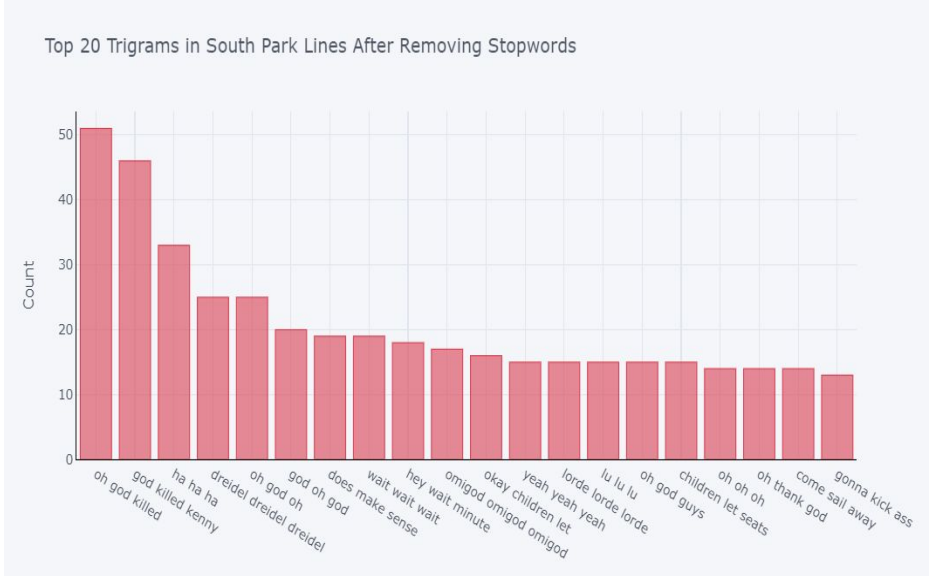


FREQUENCY OF PROFANITY



TOP 20 BIGRAMS & TRIGRAMS FOR ALL SOUTH PARK SEASONS

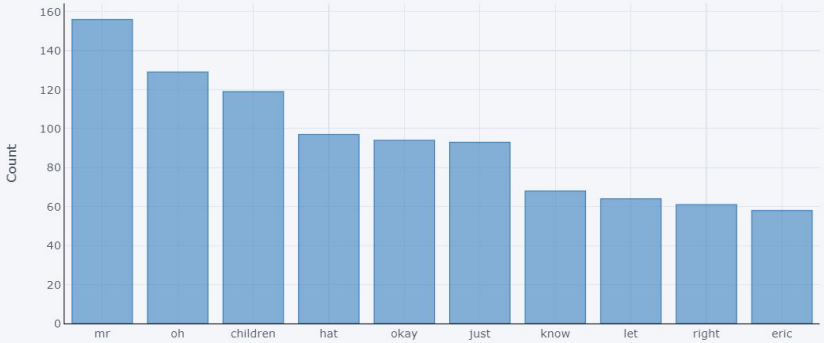




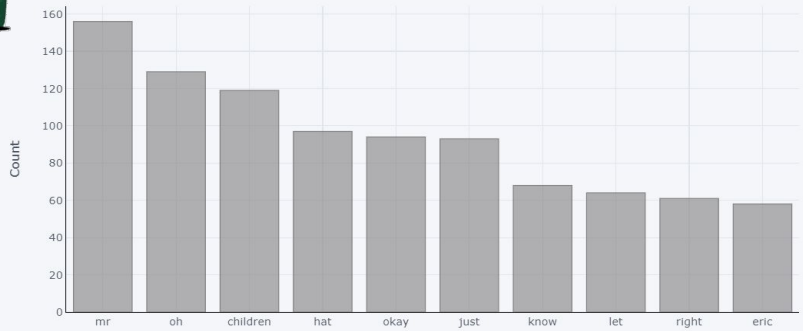
TOP 10 WORDS USED BY MAIN CHARACTERS

TOP 10 WORDS BY CARTMAN

TOP 10 WORDS BY STAN



TOP 10 WORDS BY MR. GARRISON

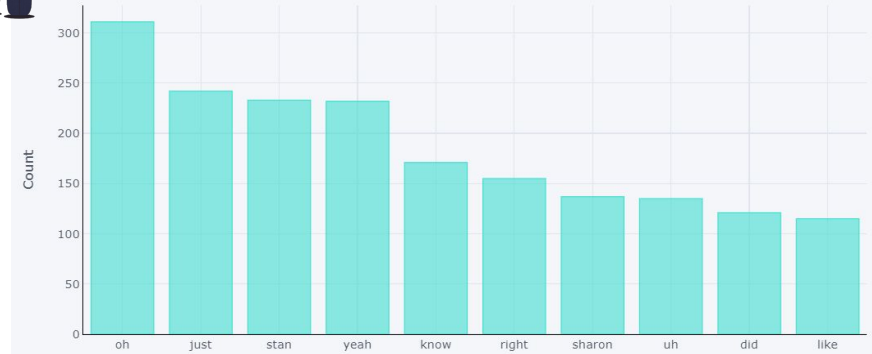


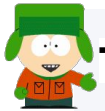
TOP 10 WORDS BY BUTTERS

TOP 10 WORDS USED BY MAIN CHARACTERS

TOP 10 WORDS BY KENNY

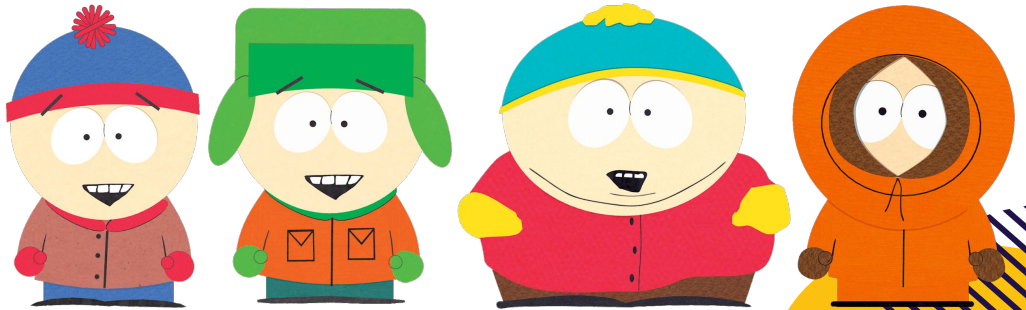
TOP 10 WORDS BY RANDY





TOP 10 WORDS BY KYLE

03 MODELING APPROACH & DEMO



SENTIMENT ANALYSIS: VADER VS. TEXTBLOB

CLEAN & FILTERED DATASET

	Season	Episode	Character	Line
0	10	1	Stan	you guys you guys chef is going away
1	10	1	Kyle	going away for how long
2	10	1	Stan	forever
4	10	1	Stan	chef said he is been bored so he joining a gr...
9	10	1	Cartman	i am gonna miss him i am gonna miss chef and...
...	...	...	...	...
70891	9	14	Stan	i think you are pushing it
70892	9	14	Randy	how about twenty
70893	9	14	Stan	that is not disciprine
70894	9	14	Randy	right right does vodka count
70895	9	14	Stan	dad

31505 rows × 4 columns

VADER

Lexicon and rule-based tool that returns **positive, negative,** and **neutral sentiment** for each line. Appropriate given nature of explicit language and lack of formal language used.

TEXTBLOB

Returns both the **polarity** and **subjectivity** of the input text

WORDCOUNT & PARTS OF SPEECH

numerical representations of new lines in each of the five aspects

NEW DATASET WITH **NEW FEATURES** OBTAINED FROM SENTIMENT ANALYSIS

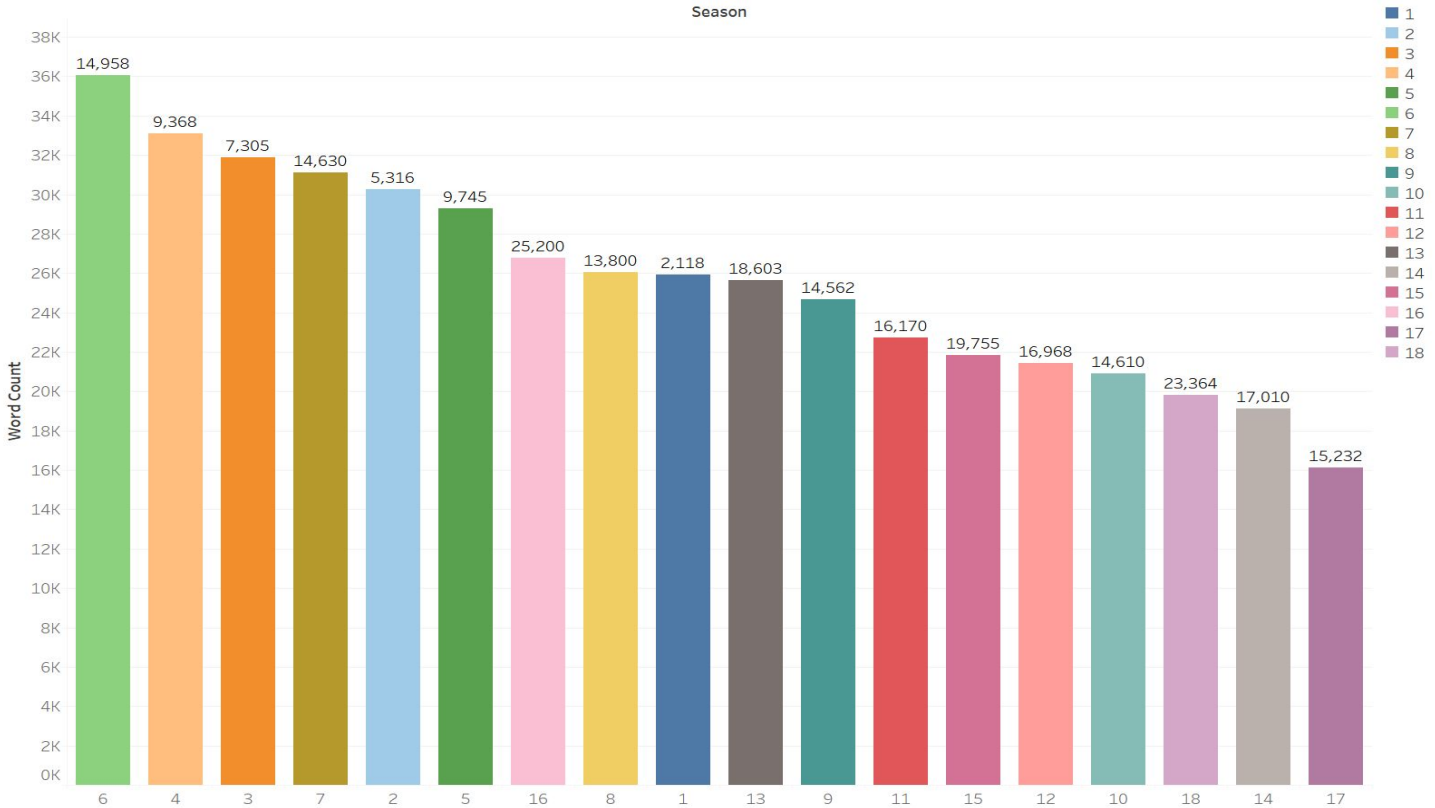
```
FinalData.head()
```

son	Episode	Character	Line	tb_polarity	tb_subjectivity	Tags	Vader_Pos	Vader_Neg	Vader_Neu	Word_Count	Vader_Net	Part_of_Speech	#_PoS
10	1	Stan	you guys you guys chef is going away	0.000000	0.000000	[(you, PRP), (guys, VBP), (you, PRP), (guys, V...	0.000	0.000	1.000	12	0.000	{'PRP': 2, 'VBP': 2, 'NN': 1, 'VBZ': 1, 'VBG':...	6
10	1	Kyle	going away for how long	-0.050000	0.400000	[(going, VBG), (away, RB), (for, IN), (how, WR...	0.000	0.000	1.000	7	0.000	{'VBG': 1, 'RB': 1, 'IN': 1, 'WRB': 1, 'JJ': 1}	5
10	1	Stan	forever	0.000000	0.000000	[(forever, RB)]	0.000	0.000	1.000	2	0.000	{'RB': 1}	1
10	1	Stan	chef said he is been bored so he joining a gr...	-0.083333	0.833333	[(chef, NN), (said, VBD), (he, PRP), (is, VBZ)...	0.291	0.099	0.610	19	0.192	{'NN': 5, 'VBD': 2, 'PRP': 2, 'VBZ': 1, 'VBN':...	8
10	1	Cartman	i am gonna miss him i am gonna miss chef and...	0.000000	0.000000	[(i, NN), (am, VBP), (gon, VBG), (na, TO), (mi...	0.000	0.144	0.856	27	-0.144	{'NN': 4, 'VBP': 3, 'VBG': 2, 'TO': 3, 'VB': 5...	9

WORD COUNT BY SEASON

Word Count by South Park Seasons

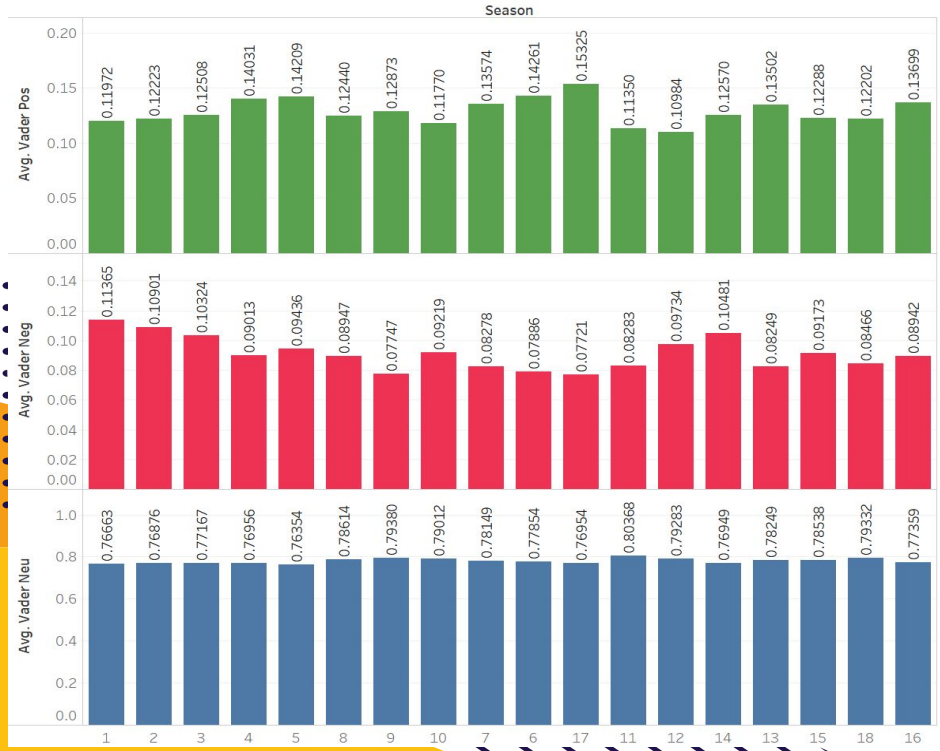
Seasons 1-18



VADER SENTIMENT BY SEASONS & CHARACTERS

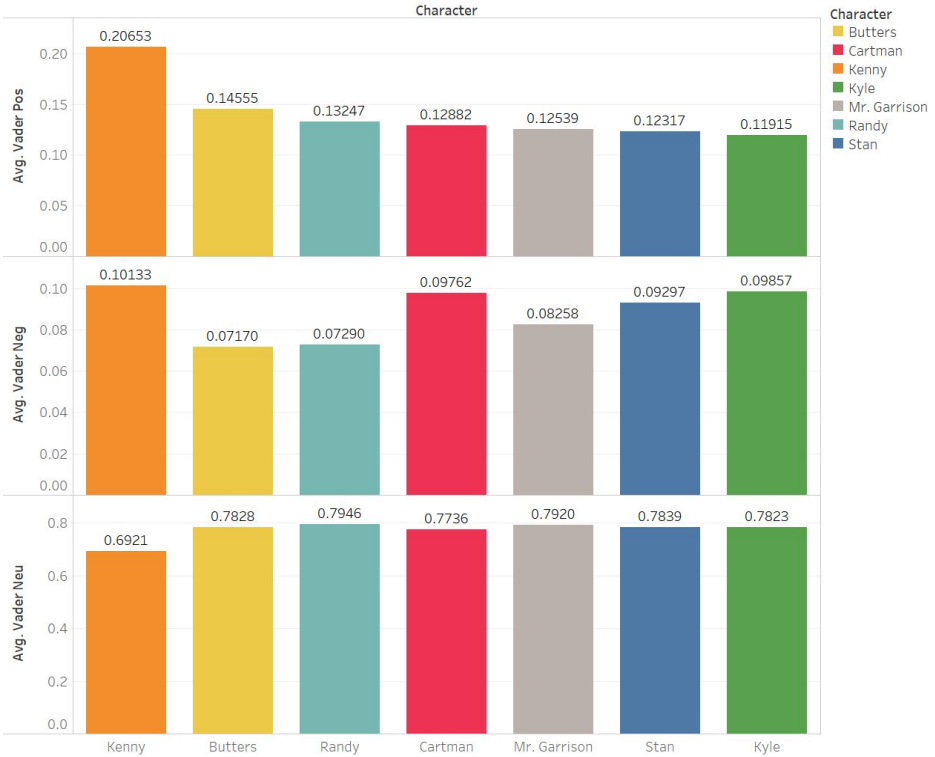
Average Vader Sentiment by South Park Seasons

Seasons 1-18



Average Vader Sentiment by South Park Characters

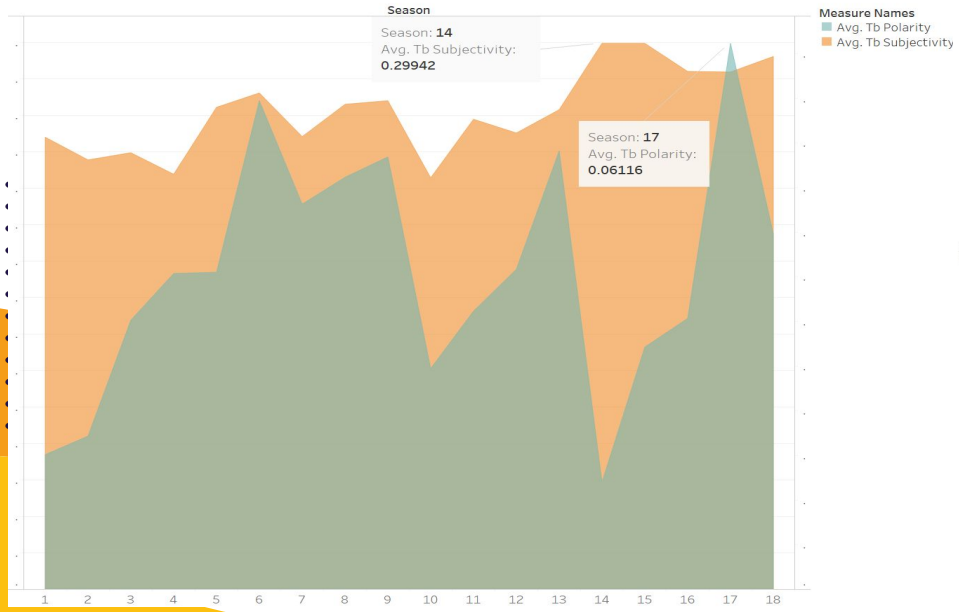
Seasons 1-18



POLARITY VS. SUBJECTIVITY BY SEASONS & CHARACTERS

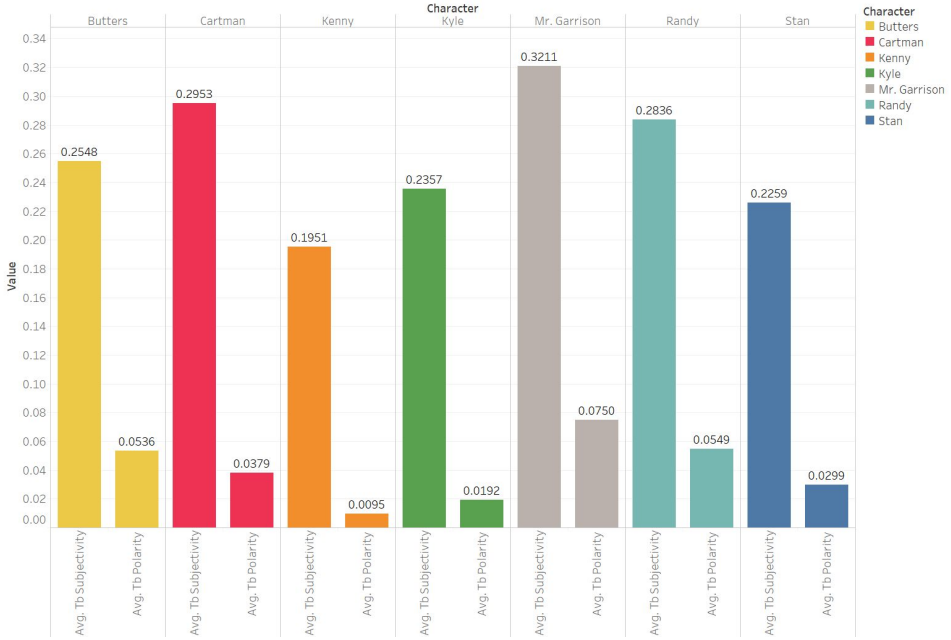
TextBlob Polarity vs. Subjectivity by South Park Seasons

Seasons 1-18



TextBlob Polarity vs. Subjectivity by South Park Characters

Seasons 1-18

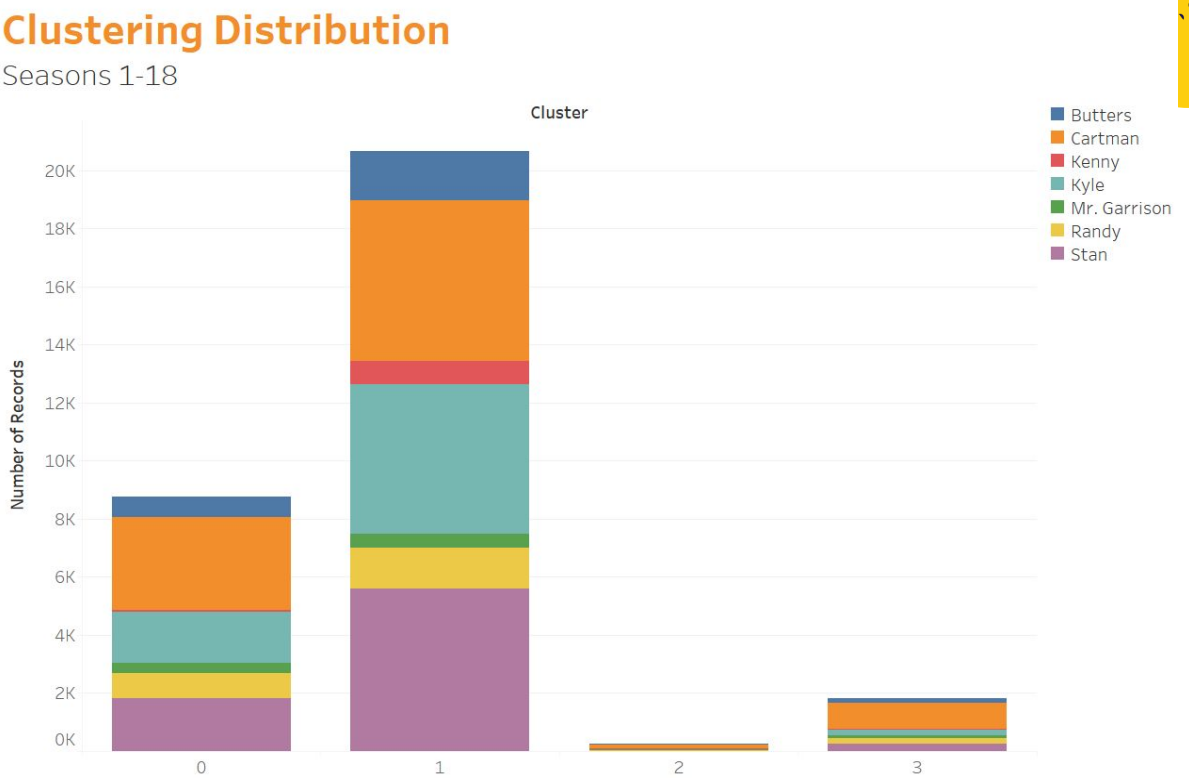


CLUSTERING

USING **SENTIMENT ANALYSIS DATASET** WITH NEW FEATURES

Clustered on:

- **VADER**
  - Neg
  - Pos
  - Neu
- **TextBlob**
  - Polarity
  - Subjectivity
- **Word Count**
- **# of Parts of Speech**
- **TfidfVectorizer**



CHARACTER PREDICTION

CARTMAN/NOT CARTMAN

Less Preprocessed

- Convert words to lowercase (Hello → hello)
- Filter out stopwords (and, the)
- Remove punctuation (e.g. .,!/?)

More Preprocessed

- Break down contractions (don't → do not)
- Lemmatize words (eating/ate → eat)

ALL CHARACTERS vs. MAIN CHARACTERS

COMPARE MODEL ACCURACY

CHARACTER PREDICTION

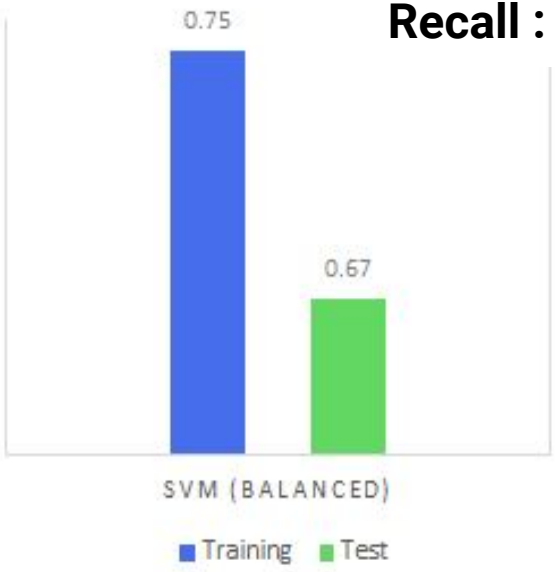
CARTMAN/NOT CARTMAN

LESS PROCESSED

Recall : 0.63

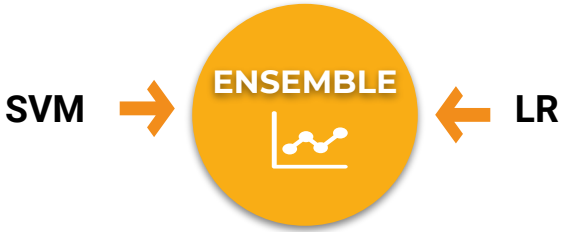
MORE PROCESSED

Recall : 0.63



CHARACTER PREDICTION CARTMAN/NOT CARTMAN

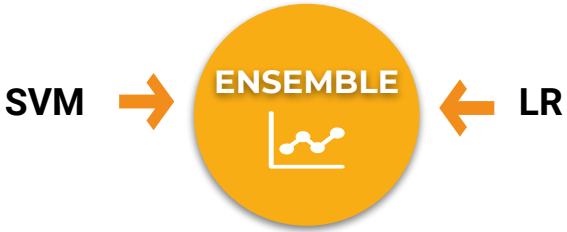
ALL CHARACTERS
(less processed)



Accuracy: 81%
Recall: 78%

>

MAIN CHARACTERS
(more processed)



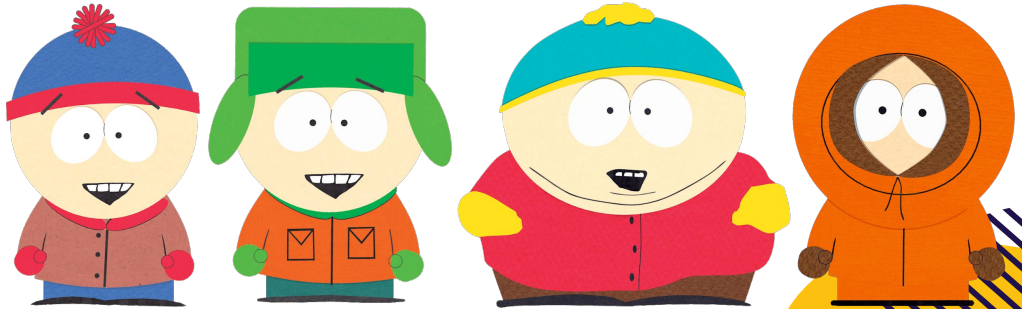
Accuracy: 77%
Recall: 76%

CHARACTER PREDICTION

CARTMAN/NOT CARTMAN

DEMO

04 CHALLENGES & IMPROVEMENTS



CHALLENGES

1. Finding a balance between precision, recall, & accuracy when selecting a model

2. There is a lot of overlap in dialogue for main characters

3. The packages with pre-defined sentiment is not reflecting the context of SP

   e.g) FUCK YEAH, THAT'S SICK, HELL YEAH

THANKS!

Do you have any questions?

05 APPENDIX

FUTURE IMPROVEMENTS

- **Character Prediction Model**
  - Add a new column that calculates line length
    - E.g. set threshold with sentences > 3 words
  - Create custom features for vectorizers