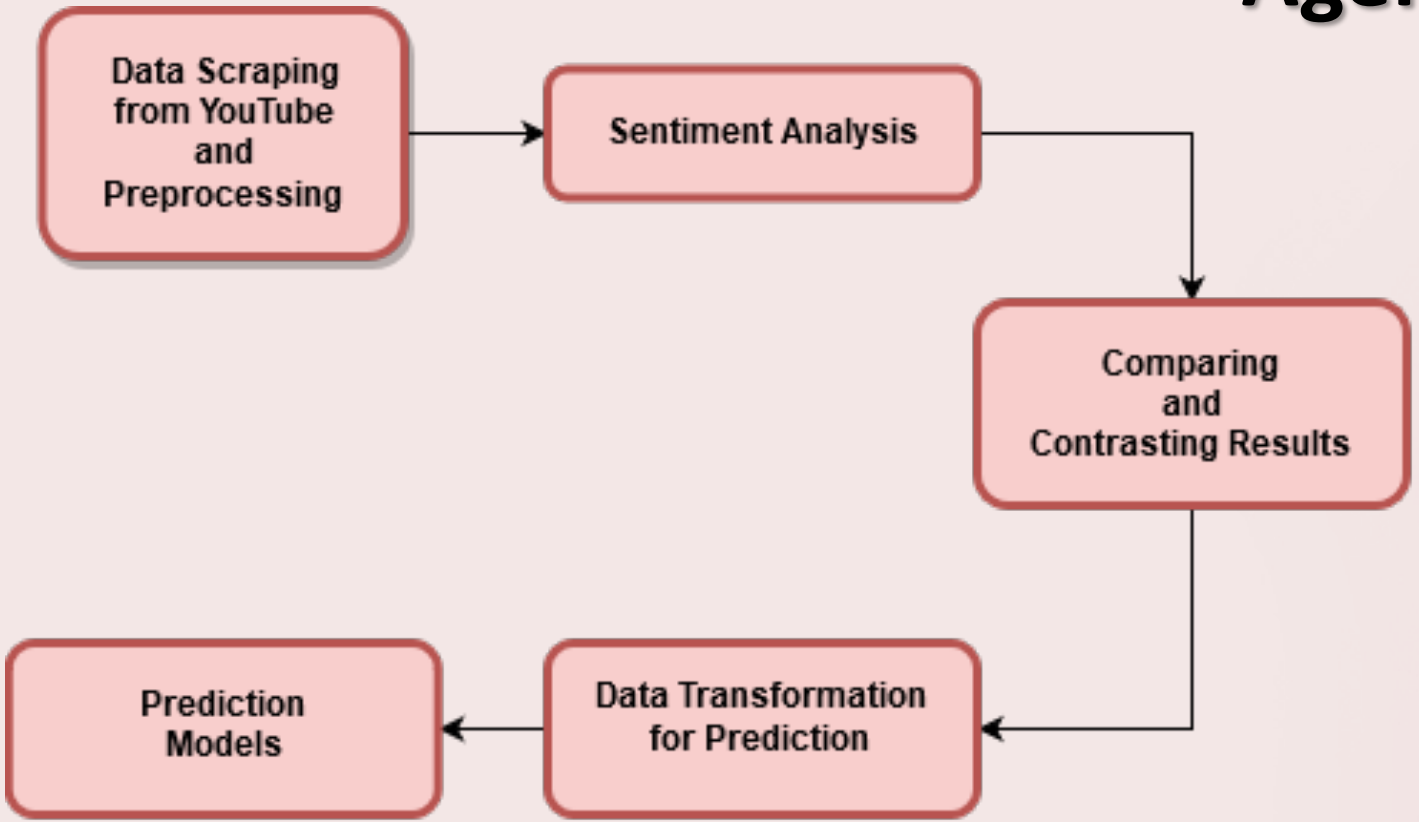
CIS-600 Social Media and Data Mining
Project Presentation

YouTube Video Comment Sentiment Analysis

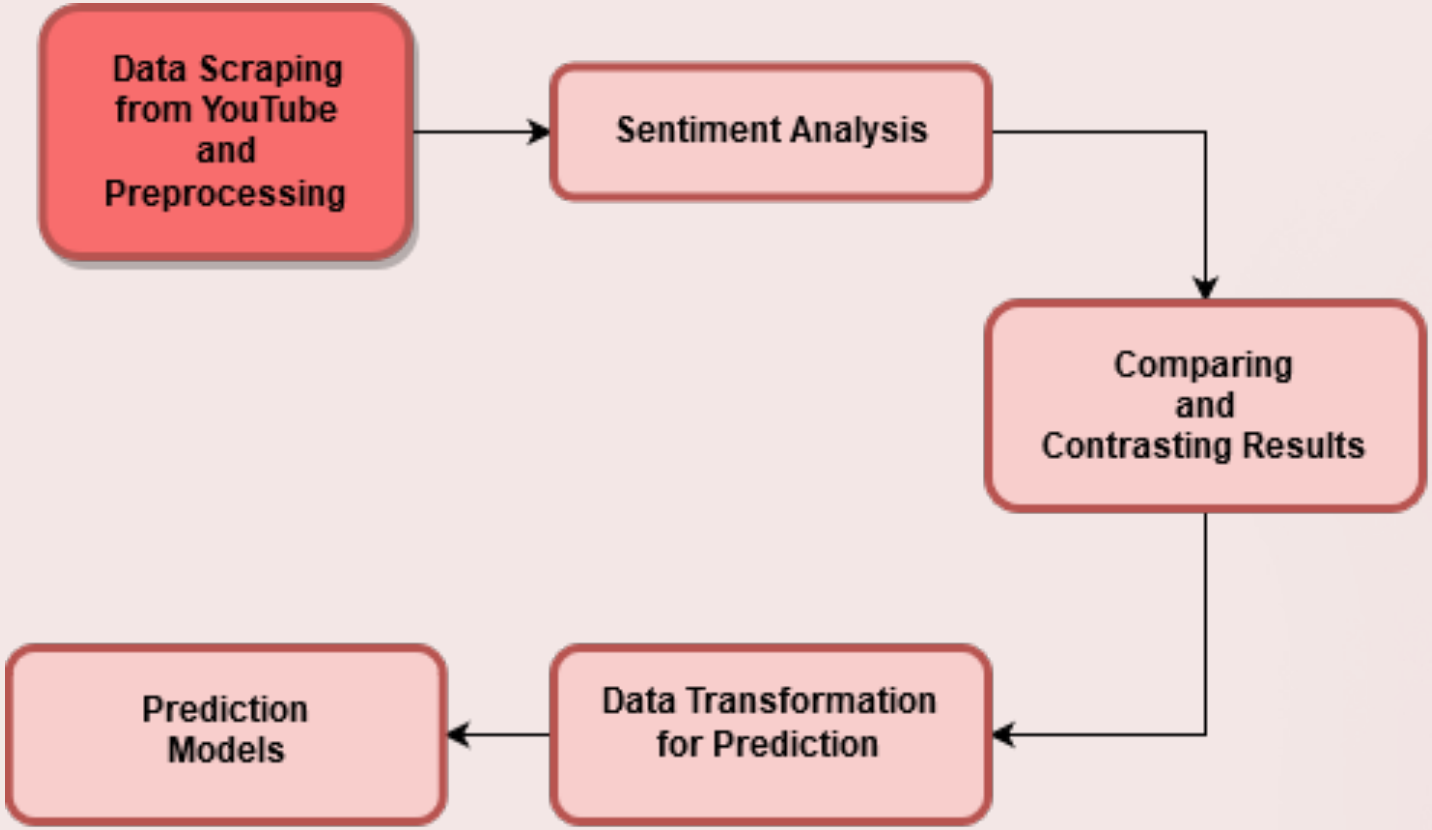
Pranay Kumar Verma
Prateek Sahu
Rohith Pattathil
Sitesh Mishra
Nitesh Nawlani

Agenda


Data Scraping from YouTube and Preprocessing
Sentiment Analysis
Comparing and Contrasting Results
Data Transformation for Prediction
Prediction Models

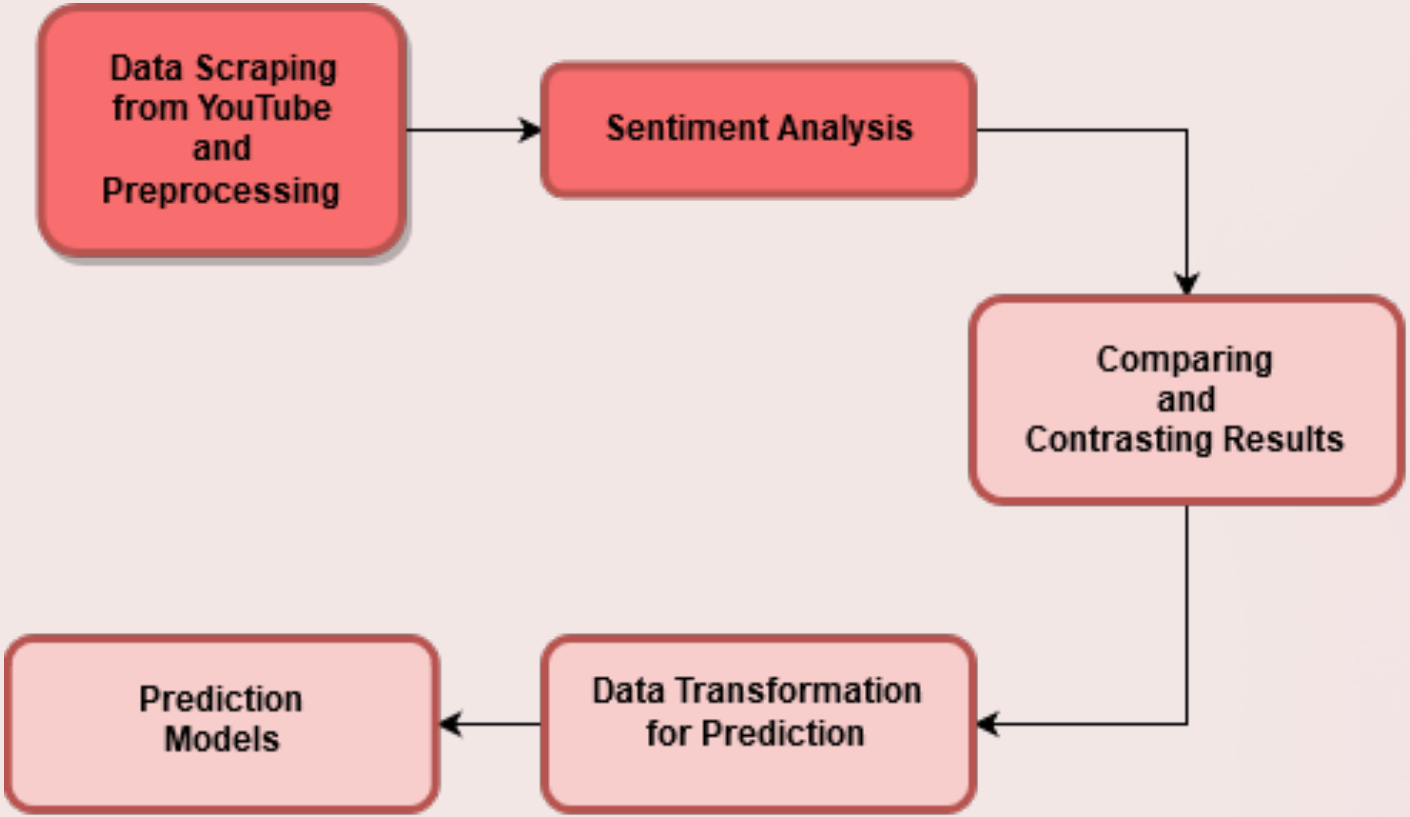
Introduction

- Founded in 2005, YouTube is the most popular platform for sharing and watching videos. The content until February 2020 amounts up to 6.5 million hours viewing (750 years).
- With the onset of the COVID-19 pandemic and restrictions in place forcing people to stay at home, YouTube witnessed a huge spike in viewing traffic, so much so that it had to reduce the default streaming quality.
- Such a huge viewing time generates exponential ad revenue. Hence, motivating many people to pursue YouTube as a career option.


Data Scraping from YouTube and Preprocessing
Sentiment Analysis
Comparing and Contrasting Results
Data Transformation for Prediction
Prediction Models

All About Data

- Dataset is extracted using the YouTube Data API v3.
- The baseline data was made by extracting 1200 comments from multiple YouTube channels and manually classifying them into positive, negative and neutral comments.
- Extracted 2500 most recent comments for each of the 200 videos of a YouTube channel based on user input.
- Feature selection: Video ID, Title, Comments, their last update date, etc.
- Preprocessing the dataset includes removing punctuations, special characters, stop words, tokenizing and stemming.


Data Scraping from YouTube and Preprocessing
Sentiment Analysis
Comparing and Contrasting Results
Data Transformation for Prediction
Prediction Models

Sentiment Analysis

Models used

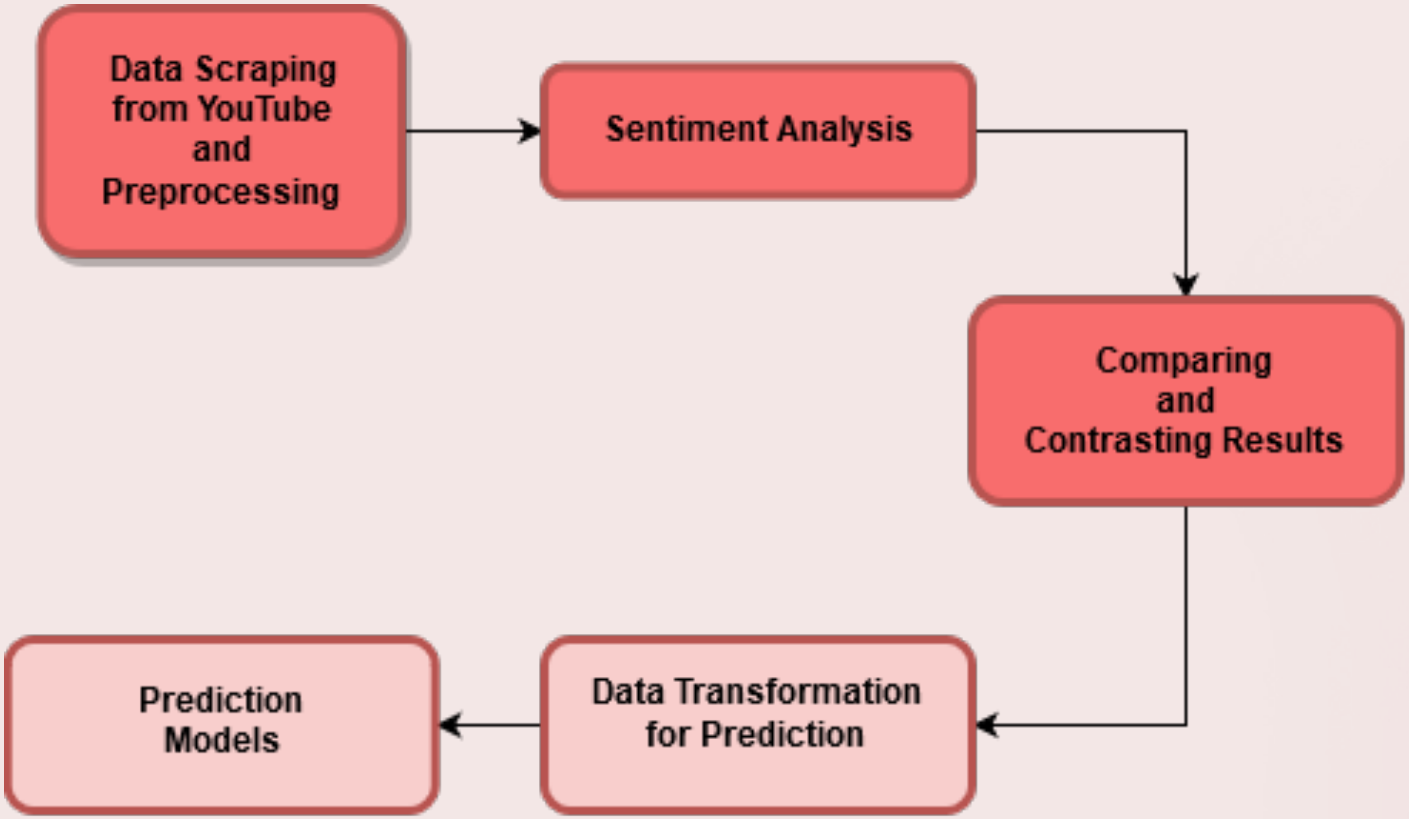
- **Vader** - Lexicon and rule-based sentiment analysis which is specifically attuned to sentiments expressed in social media. It works on a complete sentence analyzing text in addition to punctuations and emojis.
- **Afinn** - Builds up a score for a sentence word by word after tokenizing, removing the stop words and stemming the text.
- **NRC Lexicon** - Lexicon based approach used for its ability to grasp actual emotion rather than positive or negative.

Sentiment Analysis

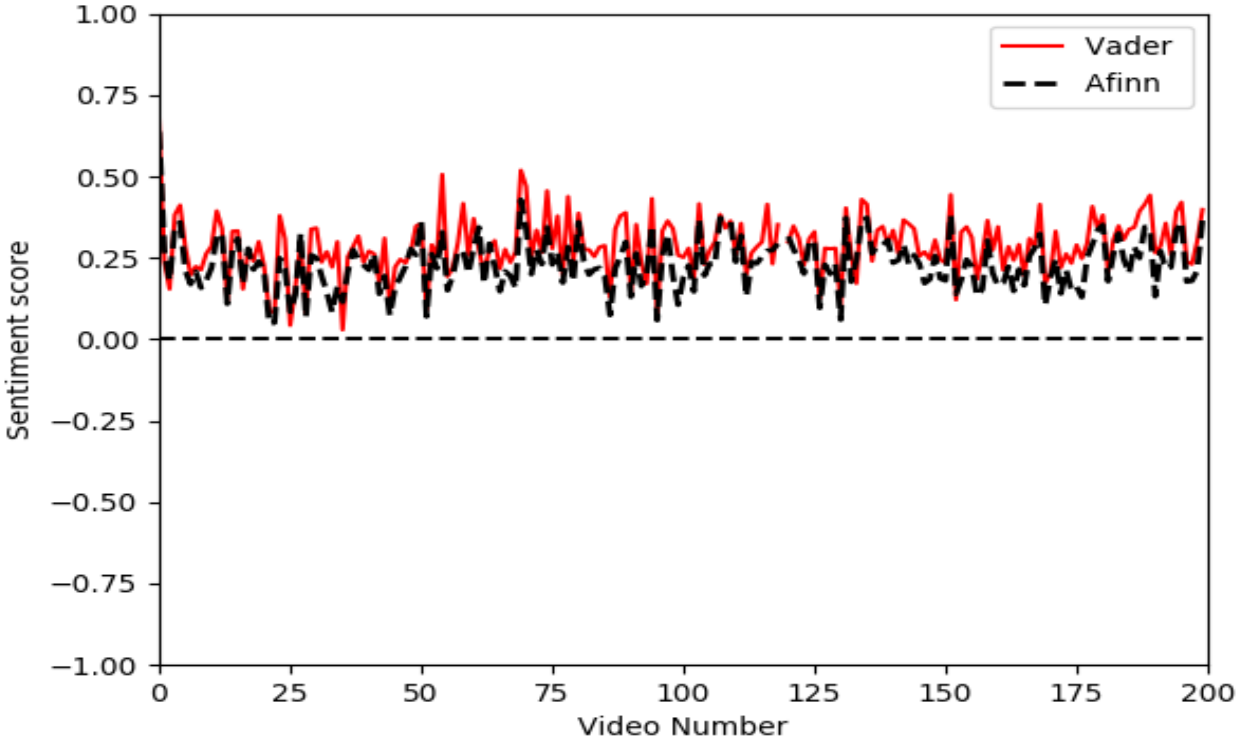
Baseline accuracy:

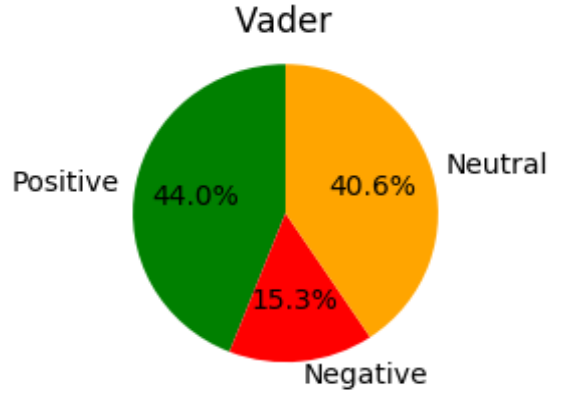
- We selected various sentiment analysis models such as Vader and Afinn and ran them on baseline data for validation of performance.
- On analysis, Vader showed more promising results in comparison to Afinn.

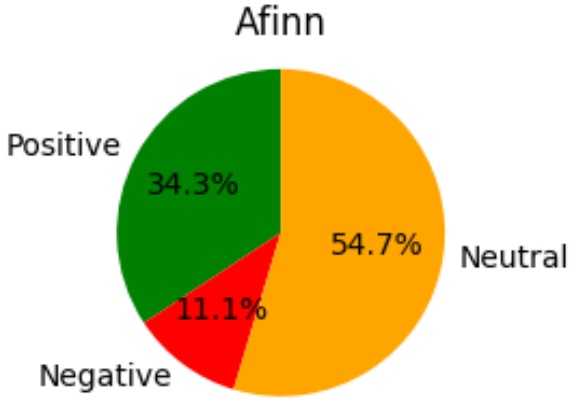
Model	Accuracy
Vader	78.79%
Afinn	70.88%


Data Scraping from YouTube and Preprocessing
Sentiment Analysis
Comparing and Contrasting Results
Data Transformation for Prediction
Prediction Models

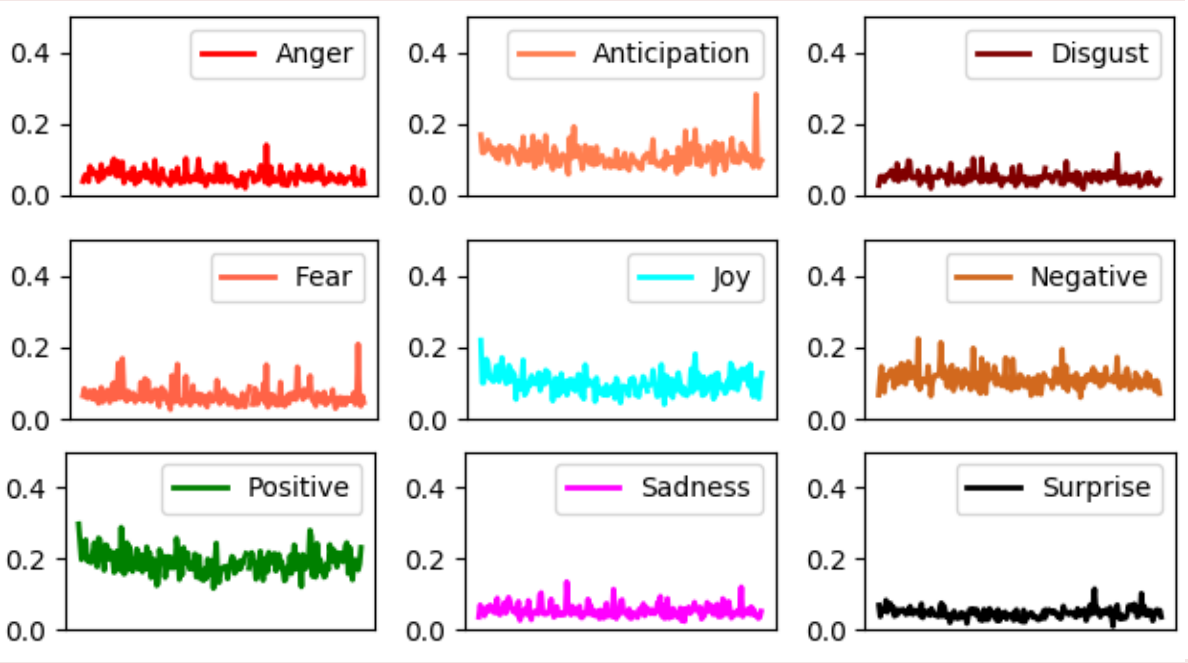
Vader V/S Afinn

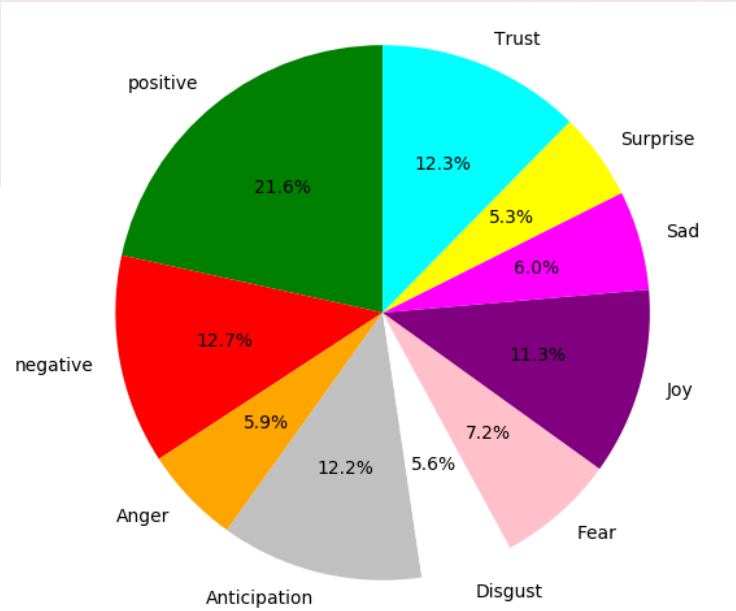






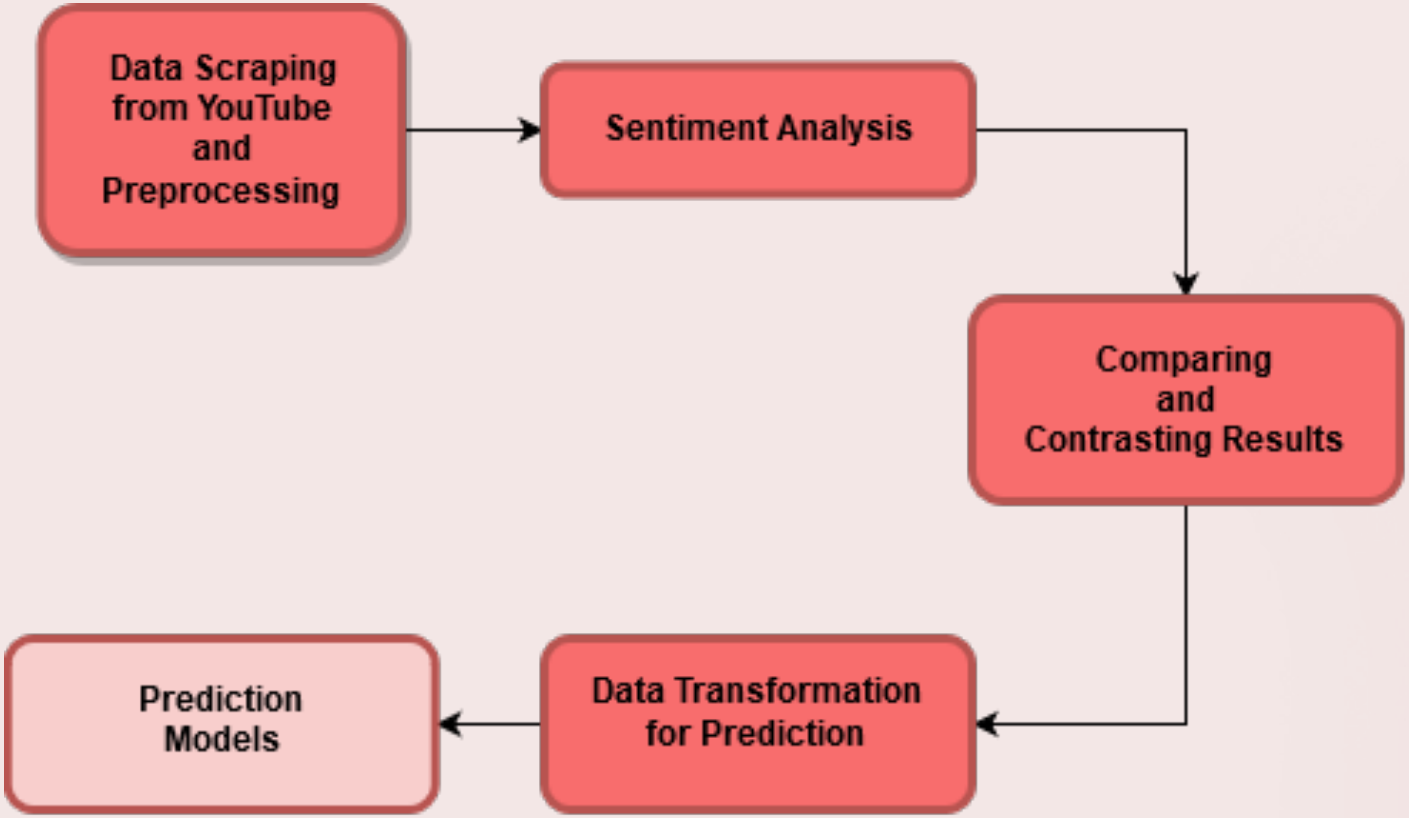
NRC Lexicon


Anger
Anticipation
Disgust
Fear
Joy
Negative
Positive
Sadness
Surprise
Trust
0.0
0.2
0.4

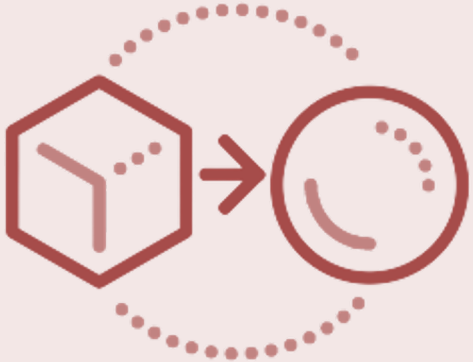

positive
Trust
Surprise
Sad
Joy
Fear
Disgust
Anticipation
Anger
negative
21.6%
12.3%
5.3%
6.0%
11.3%
7.2%
5.6%
12.2%
5.9%
12.7%

WordCloud

even
lambo
look
channel
amazing
made
Great video
shit
new
actually
awesome
way
man
something
great
content
Ferrari
time
say
McLaren
engine
please
sound
Supra
comment
make
guy
buy
house
sick
let
see
take
watch
love
really
never
said
cool
come
one
lot
better
thing
lol
well
day
sure
drive
best
E63
BMW
fun
much
keep
video
know
P1
Rolls Royce
bad
going
Damn
Next video
car
already
people
lmao
feel
alway
think
year
600LT
got
gon na
DDE
good
give
thought
mean
still
nice
bro
driving
back
Lamborghini
need
go
Parker
review
put
maybe
want
first
dude
right
seem
Thank
Wow
use
getting

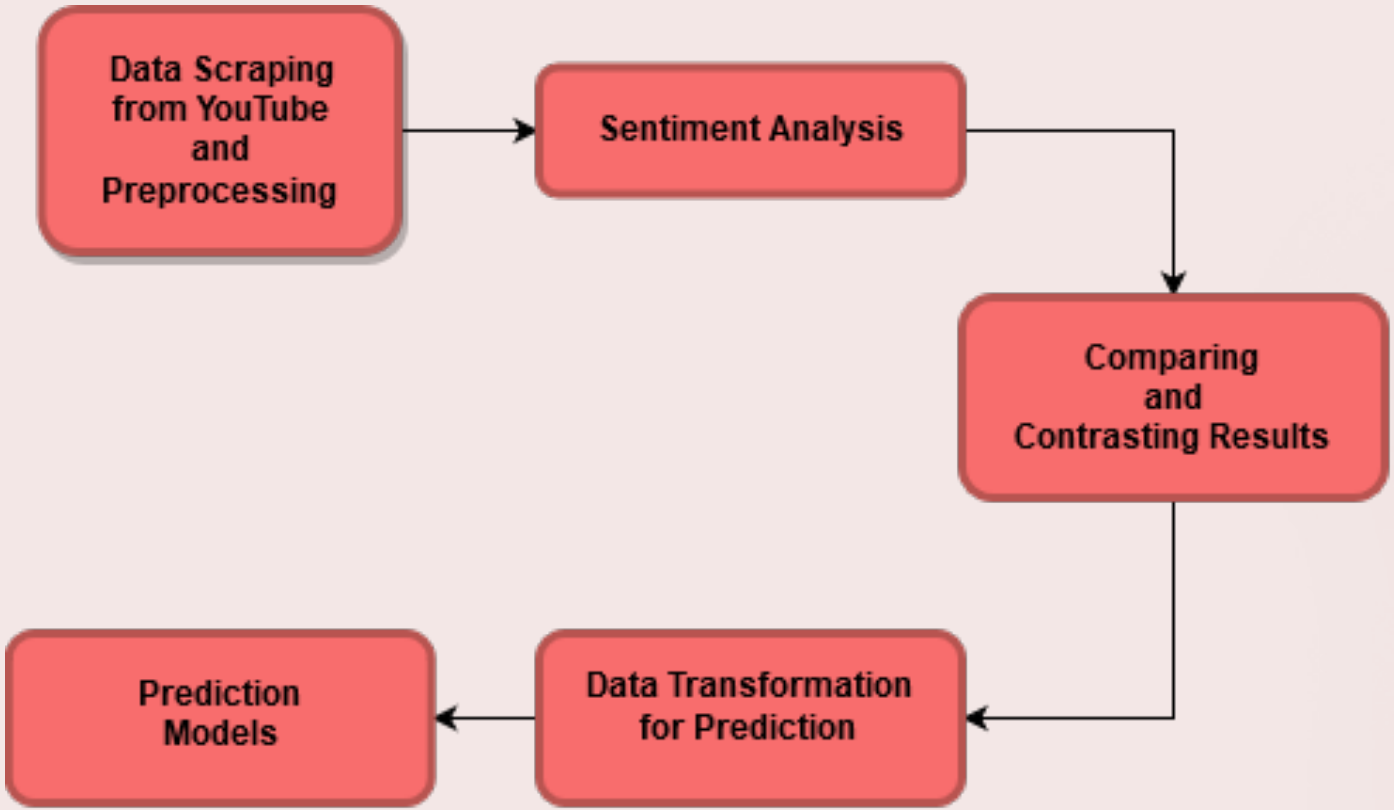

Data Scraping from YouTube and Preprocessing
Sentiment Analysis
Comparing and Contrasting Results
Data Transformation for Prediction
Prediction Models

Data Transformation



Transformation to Time-Series dataset

- The comments as well as sentiment analysis scores are now transformed by grouping based on dates for creating a time-series dataset.
- This dataset is now used for performing time series prediction using simple Machine Learning algorithms as well as Neural Networks.


Data Scraping from YouTube and Preprocessing
Sentiment Analysis
Comparing and Contrasting Results
Data Transformation for Prediction
Prediction Models

Prediction Models

What we aim to achieve?

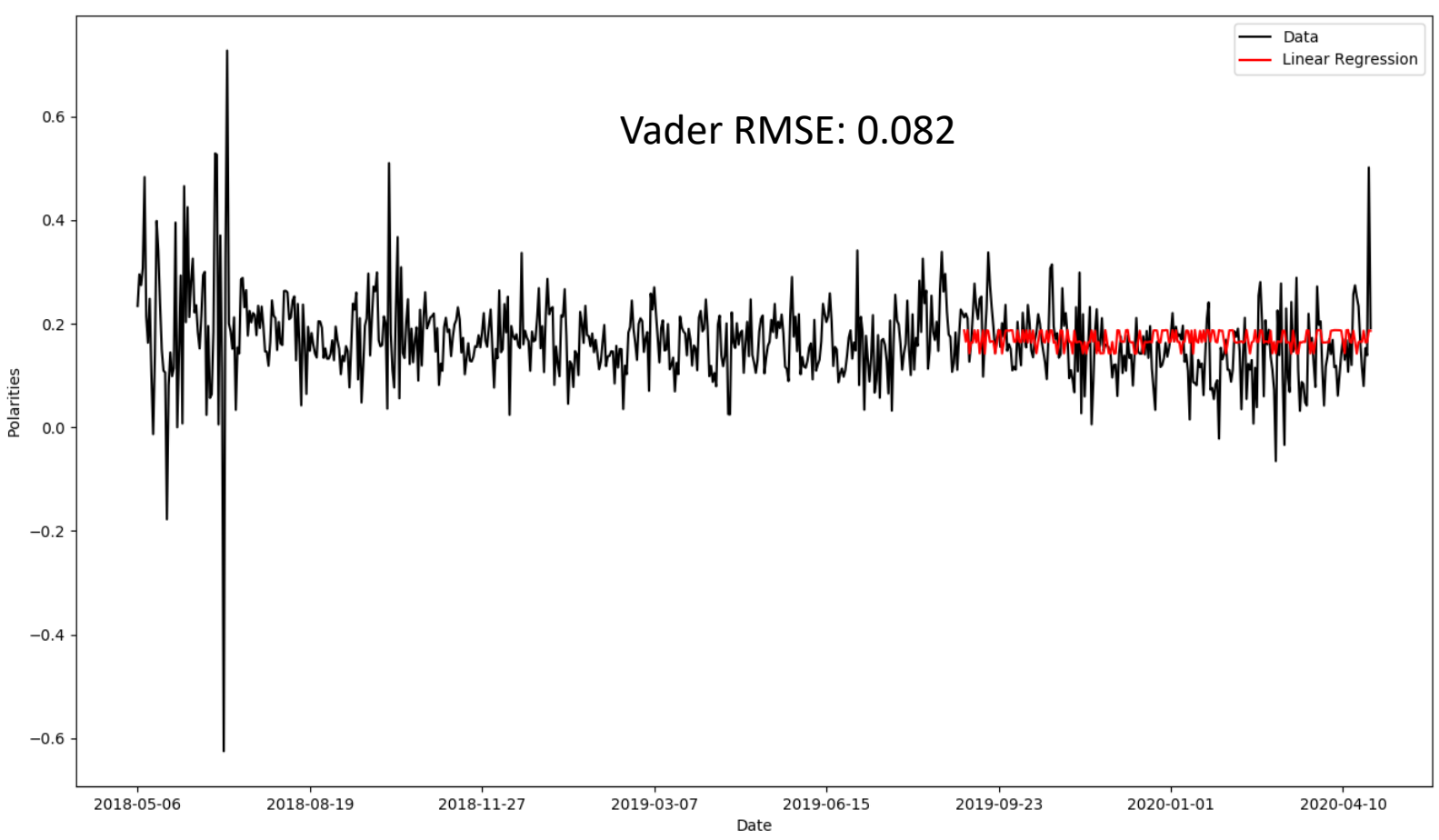
- The goal is to identify how the comments posted on a specific day vary in sentiment.
- Observing the past trends, we aim to make a reasonable prediction on the sentiment in the upcoming days and months.
- This in turn will help the channel owners to alter their content accordingly to keep the viewers happy.

Prediction Models (Cont.)

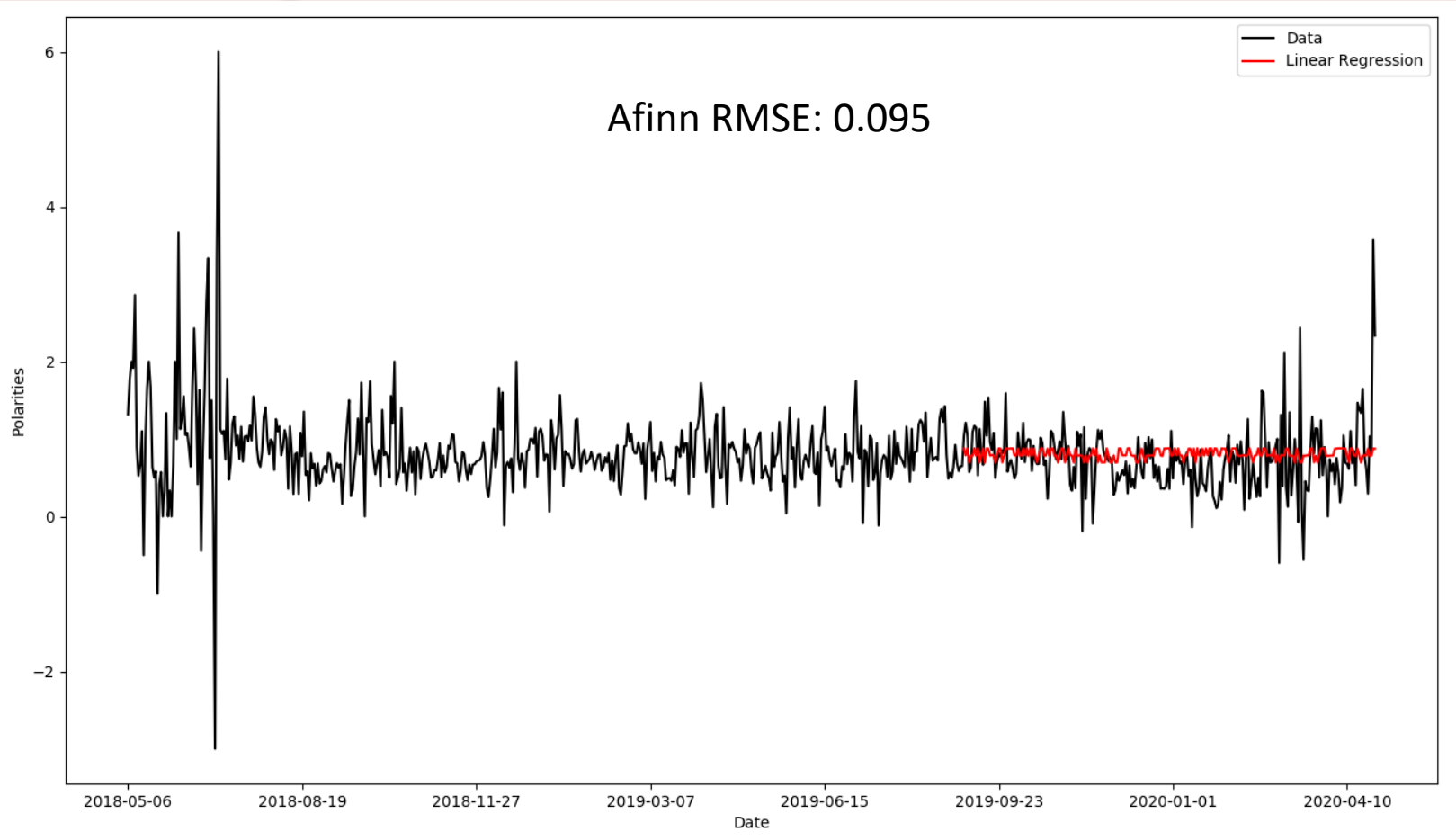
Models Used

- **Linear Regression** - One of the most basic Machine Learning Algorithm, used to fit linear relations
- **Polynomial Regression** - Fits a nonlinear relationship.
- **LSTM** (Long Short Term Memory) - An advanced part of Artificial Neural Networks, which is also recurring, the previous state is preserved.

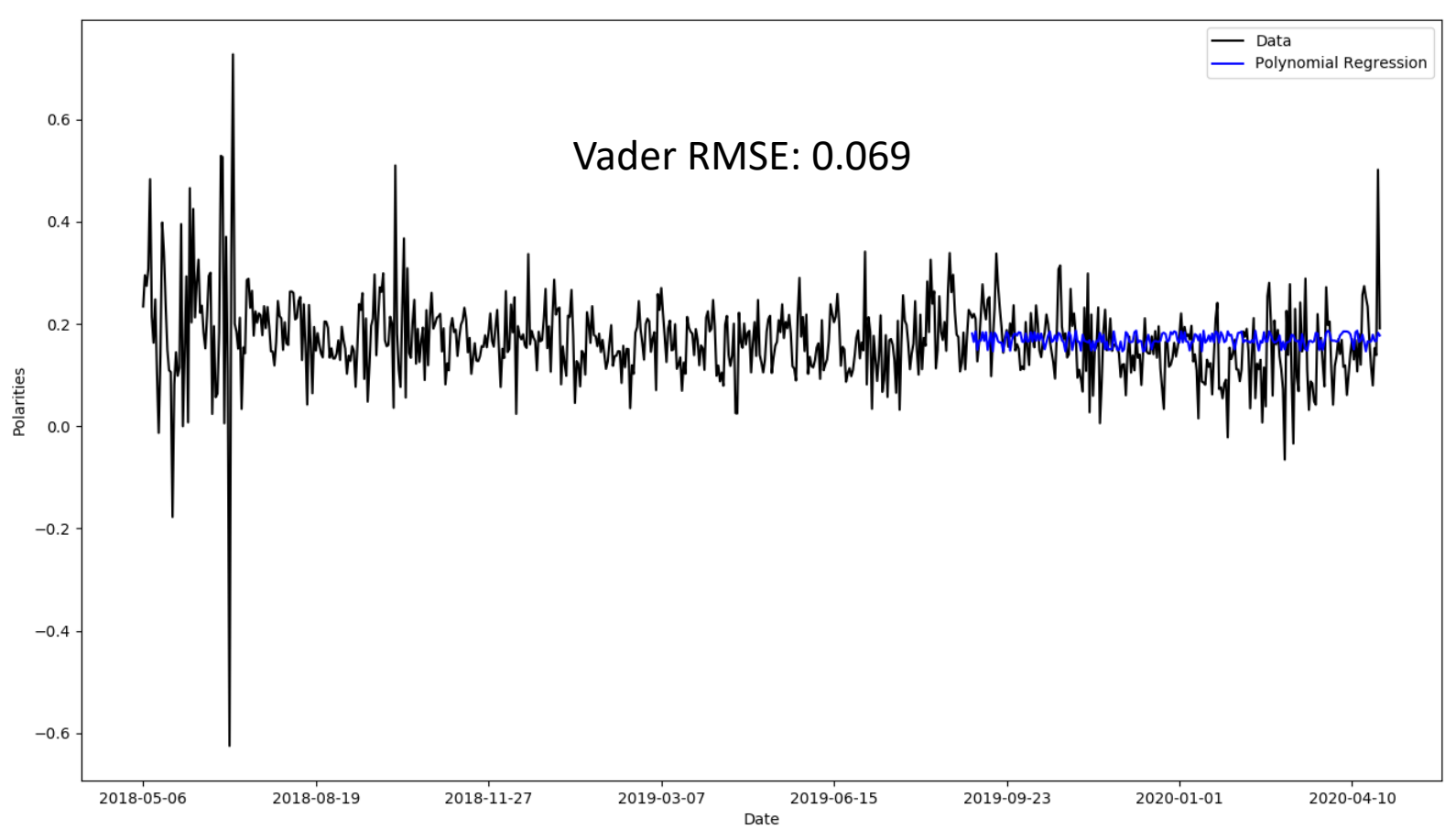
Linear Regression Vader



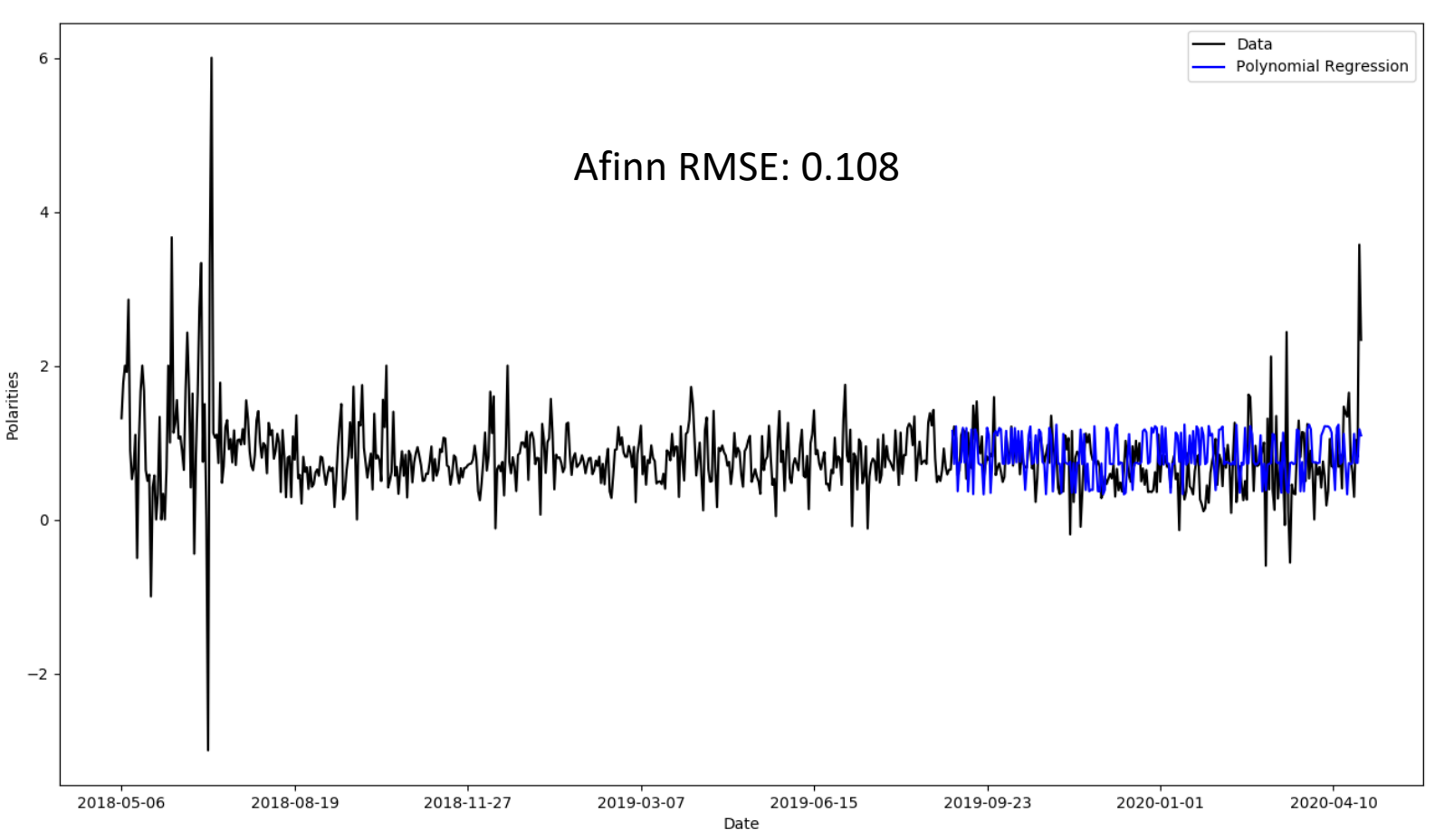
Linear Regression Afinn



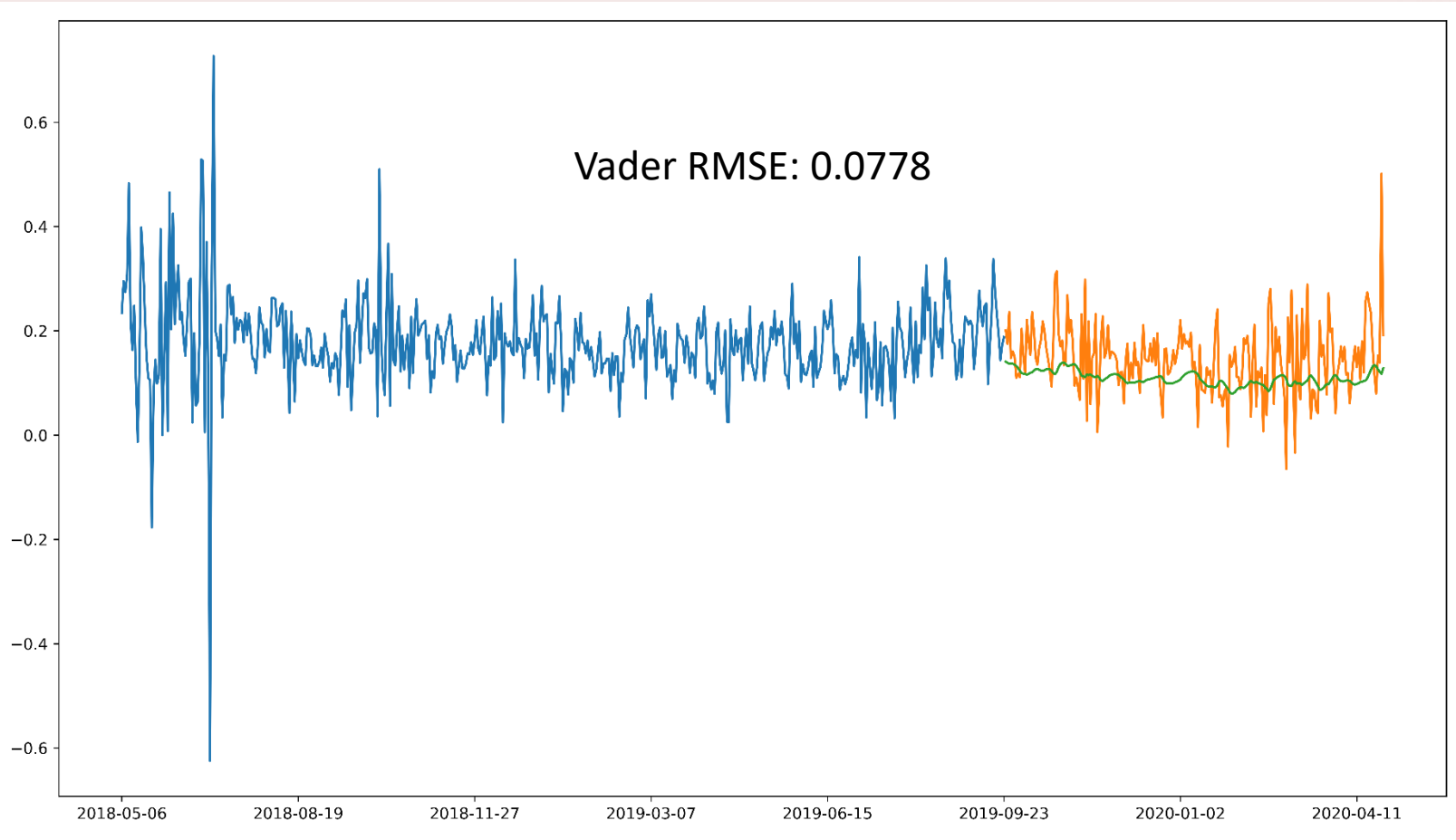
Polynomial Regression Vader



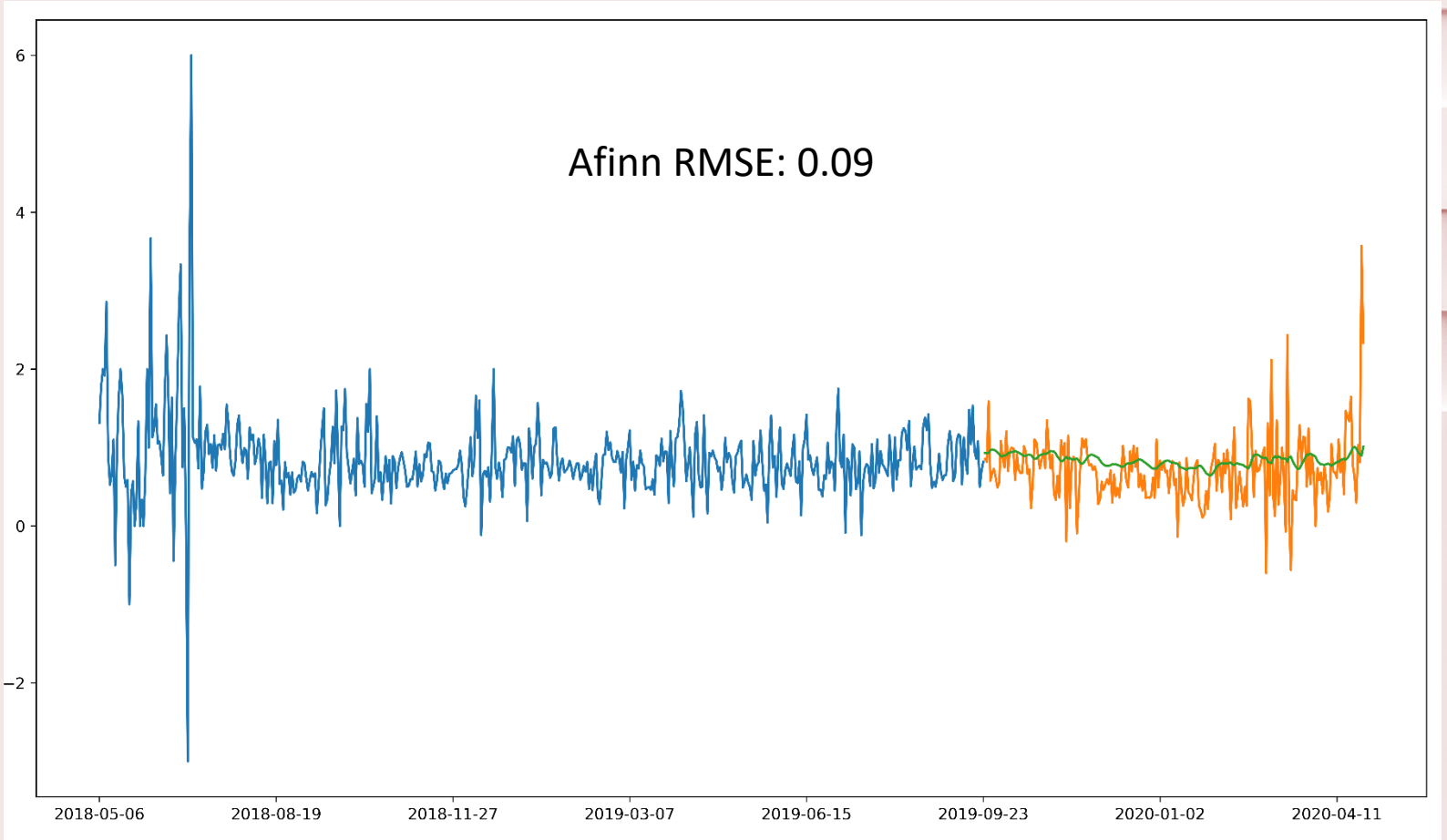
Polynomial Regression Afinn



LSTM Vader


Vader RMSE: 0.0778
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
2018-05-06
2018-08-19
2018-11-27
2019-03-07
2019-06-15
2019-09-23
2020-01-02
2020-04-11

LSTM Afinn


Afinn RMSE: 0.09
6
4
2
0
−2
2018-05-06
2018-08-19
2018-11-27
2019-03-07
2019-06-15
2019-09-23
2020-01-02
2020-04-11

Which Performs Best?

- As per our baseline analysis as well as the predictions performed, Vader constantly outperforms Afinn.
- The results of the model rank from best to worst are:
  1. Polynomial Regression (approx. 6.9% off the mark)
  2. LSTM (approx. 7.7% off the mark)
  3. Linear Regression (approx. 8.2% off the mark)

Additional Analysis

- The common belief is that the like/dislike ratio can project how the sentiment of the comments is.
- But is this really the case?
- We used the sentiment analysis in conjunction with the view count and comment count of a video to try and predict the likes/dislikes ratio.
- What we found was interesting, in that the comments are not necessarily in line with the likes and dislikes.

Future Scope

- Obtain access to YouTube Analytics API and get historical data like subscriber count, like/dislike count and average playtime which is currently not possible with data API.
- Using this, we can predict how fast the channel will grow in future.
- With average play time at our disposal we can also predict the if the channel will have monetary gains.

Questions?

References

- https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- https://developers.google.com/youtube/v3
- https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f
- https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/