Self-Driving Market Assessment

Text Classification & Sentiment Analysis Using NLP

Subreddits: r/Selfdrivingcars vs r/Futurology vs r/Technology

Garrett Bradley
Oct 18, 2019

Problem Statement

Problem Statement

Explore online communities to identify potential self-driving car buyers in order to conduct upcoming marketing campaign, ultimately increase targeted buyers for self-driving cars.

Specifically, what other subreddits besides r/Selfdrivingcars may provide a good customer base for targeted marketing?

Solution

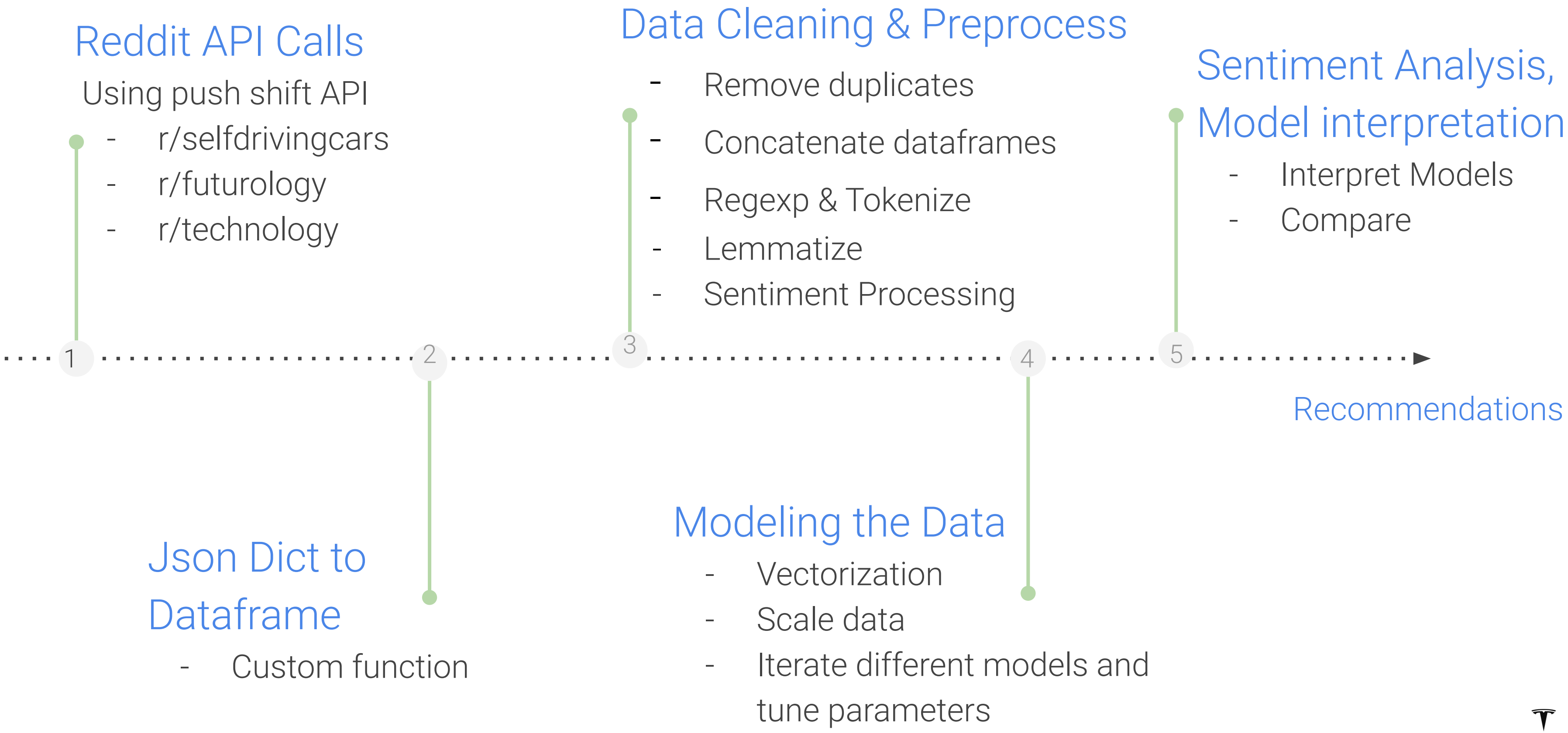
Build a classification model that uses NLP to classify text data from multiple subreddit groups.

Conduct Sentiment Analysis on text data to gain insights for targeted marketing campaign

Why is this important?

- Discover insights on consumer sentiment and behavior
- Learn which online communities frequently discuss Self-driving cars
- Understand public sentiment changes over time with respect to self-driving cars
- Learn how different communities perceive self-driving cars
- In addition to Tesla marketing, Public Sentiment can inform policy-making, either conservative or progressive

Workflow from Reddit API
to Classification & Sentiment Analysis



Reddit API Calls

Using push shift API

- r/selfdrivingcars
- r/futurology
- r/technology

Json Dict to Dataframe

- Custom function

Data Cleaning & Preprocess

- Remove duplicates
- Concatenate dataframes
- Regexp & Tokenize
- Lemmatize
- Sentiment Processing

Modeling the Data

- Vectorization
- Scale data
- Iterate different models and tune parameters

Sentiment Analysis, Model interpretation

- Interpret Models
- Compare

Recommendations

Data Collected from three subreddits from 2008 to 2019

- 5000 Data points on Submissions for each of r/Selfdrivingcars, r/Futurology, r/Technology from 2018 to the present
  - Submission data used to train classification model

- 5000 Data points on Comments for each of r/Selfdrivingcars, r/Futurology, r/Technology from 2018 to the present
  - Used for comparing model accuracy in submissions vs comments

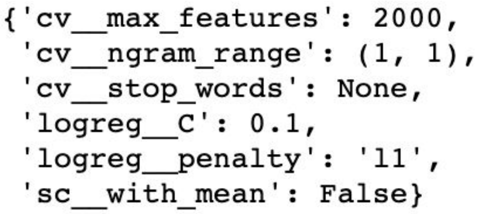
- 500 data points on each of Comments and Submissions for each Subreddit from its respective start year 2008/2011 to present year
  - Used exclusively to conduct Sentiment Analysis

Classification Algorithms Compared
On r/Selfdrivingcars Submissions

All three models performed roughly the same in Accuracy at 89%,
and all were fairly overfit

Logistic Regression

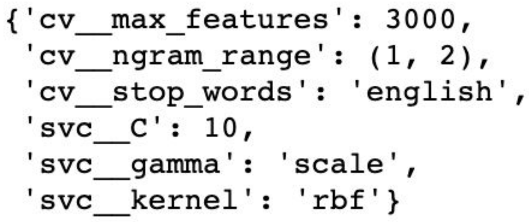
Grid Best: 0.900
Train Score: 0.959
Test Score: 0.892



```
{'cv__max_features': 2000,
 'cv__ngram_range': (1, 1),
 'cv__stop_words': None,
 'logreg__C': 0.1,
 'logreg__penalty': 'l1',
 'sc__with_mean': False}
```

Support Vector Machine

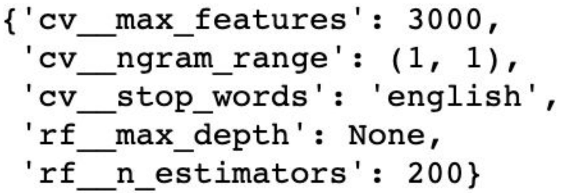
Grid Best: 0.904
Train Score: 0.992
Test Score: 0.898



```
{'cv__max_features': 3000,
 'cv__ngram_range': (1, 2),
 'cv__stop_words': 'english',
 'svc__C': 10,
 'svc__gamma': 'scale',
 'svc__kernel': 'rbf'}
```

Random Forest

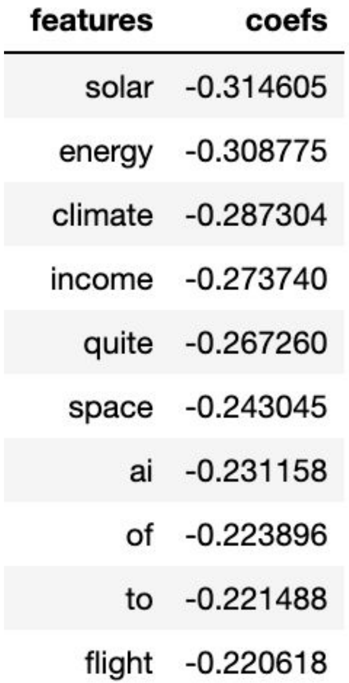
Grid Best: 0.904
Train Score: 0.996
Test Score: 0.896



```
{'cv__max_features': 3000,
 'cv__ngram_range': (1, 1),
 'cv__stop_words': 'english',
 'rf__max_depth': None,
 'rf__n_estimators': 200}
```

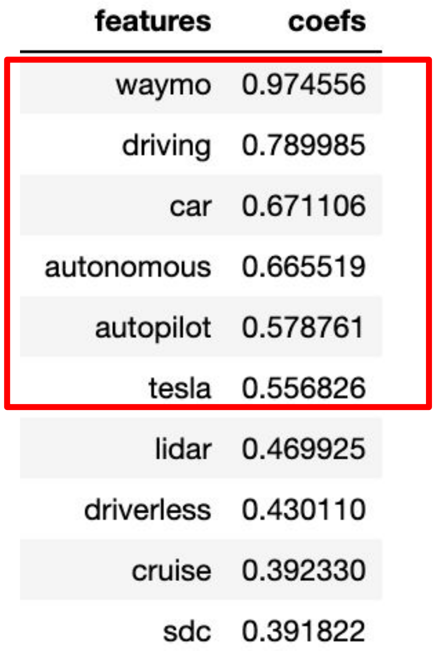
Logistic Regression High and Low influence Coefficients

Low



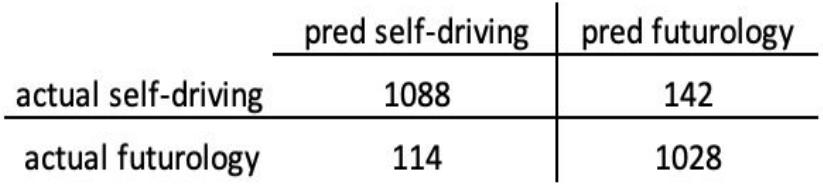
features	coefs
solar	-0.314605
energy	-0.308775
climate	-0.287304
income	-0.273740
quite	-0.267260
space	-0.243045
ai	-0.231158
of	-0.223896
to	-0.221488
flight	-0.220618

High



features	coefs
waymo	0.974556
driving	0.789985
car	0.671106
autonomous	0.665519
autopilot	0.578761
tesla	0.556826
lidar	0.469925
driverless	0.430110
cruise	0.392330
sdc	0.391822

Confusion Matrix

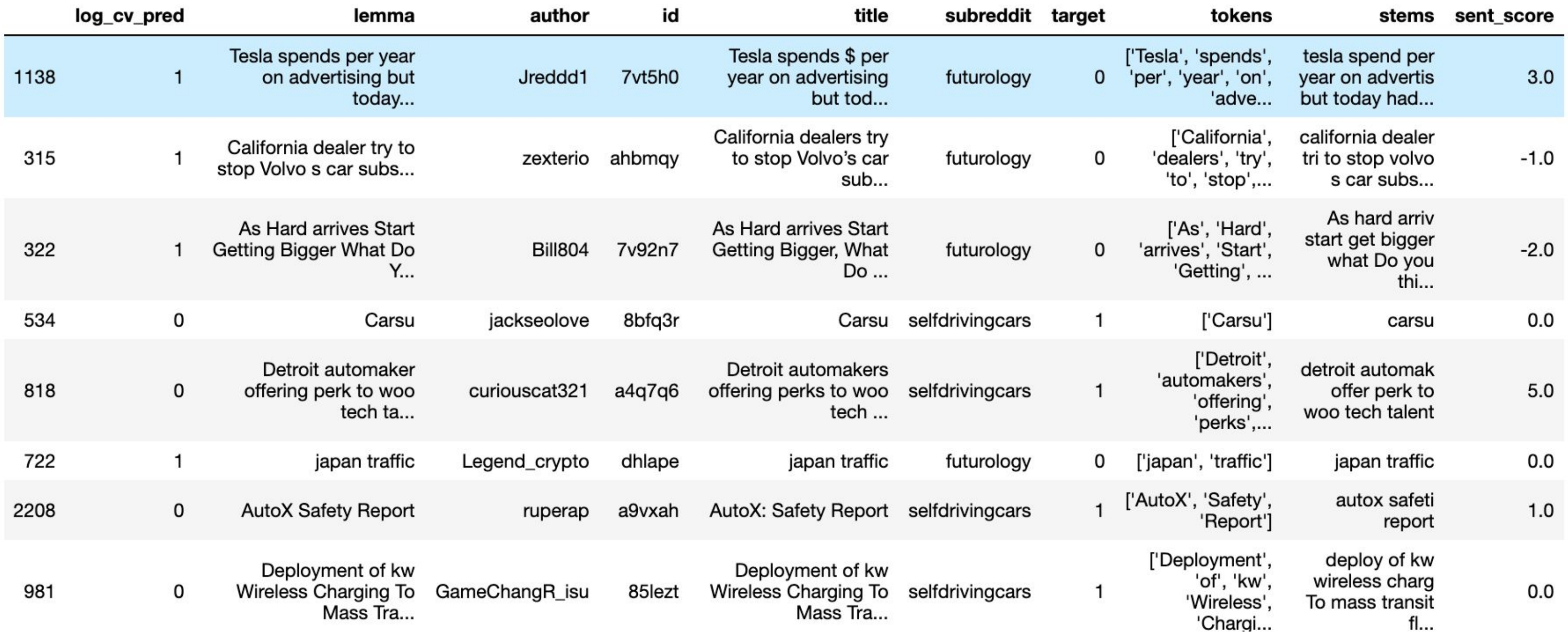


	pred self-driving	pred futurology
actual self-driving	1088	142
actual futurology	114	1028

Misclassified Posts require further investigation

Some are clearly due to posts of very few words that provide little context

Others are less clear and require further investigation



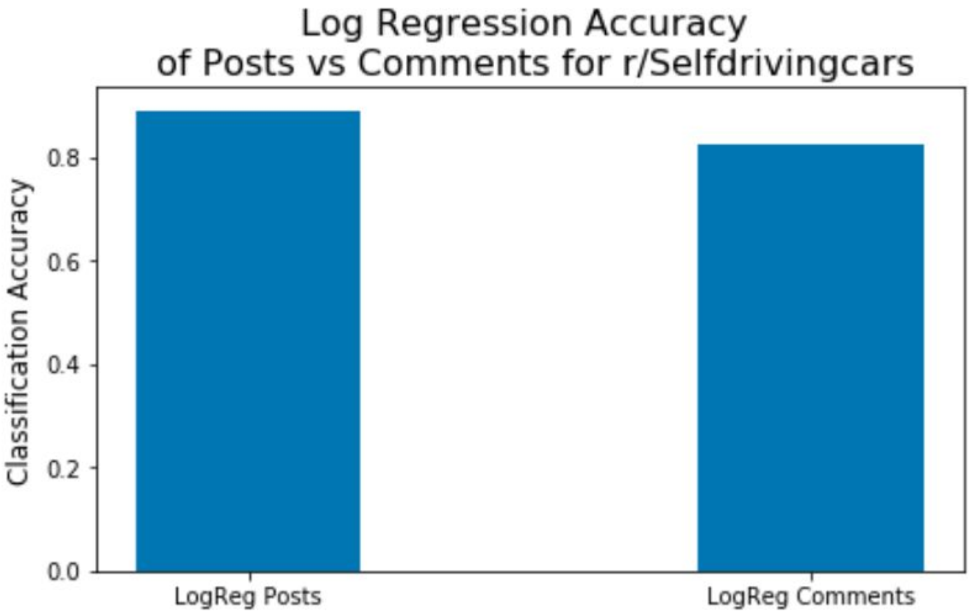
	log_cv_pred	lemma	author	id	title	subreddit	target	tokens	stems	sent_score
1138	1	Tesla spends per year on advertising but today...	Jreddd1	7vt5h0	Tesla spends $ per year on advertising but tod...	futurology	0	['Tesla', 'spends', 'per', 'year', 'on', 'adve...	tesla spend per year on advertis but today had...	3.0
315	1	California dealer try to stop Volvo s car subs...	zexterio	ahbmqy	California dealers try to stop Volvo's car sub...	futurology	0	['California', 'dealers', 'try', 'to', 'stop',...	california dealer tri to stop volvo s car subs...	-1.0
322	1	As Hard arrives Start Getting Bigger What Do Y...	Bill804	7v92n7	As Hard arrives Start Getting Bigger, What Do ...	futurology	0	['As', 'Hard', 'arrives', 'Start', 'Getting', ...	As hard arriv start get bigger what Do you thi...	-2.0
534	0	Carsu	jackseolove	8bfq3r	Carsu	selfdrivingcars	1	['Carsu']	carsu	0.0
818	0	Detroit automaker offering perk to woo tech ta...	curiouscat321	a4q7q6	Detroit automakers offering perks to woo tech ...	selfdrivingcars	1	['Detroit', 'automakers', 'offering', 'perks',...	detroit automak offer perk to woo tech talent	5.0
722	1	japan traffic	Legend_crypto	dhlape	japan traffic	futurology	0	['japan', 'traffic']	japan traffic	0.0
2208	0	AutoX Safety Report	ruperap	a9vxah	AutoX: Safety Report	selfdrivingcars	1	['AutoX', 'Safety', 'Report']	autox safeti report	1.0
981	0	Deployment of kw Wireless Charging To Mass Tra...	GameChangR_isu	85lezt	Deployment of kw Wireless Charging To Mass Tra...	selfdrivingcars	1	['Deployment', 'of', 'kw', 'Wireless', 'Chargi...	deploy of kw wireless charg To mass transit fl...	0.0

Comments provide a more challenging classification problem than posts

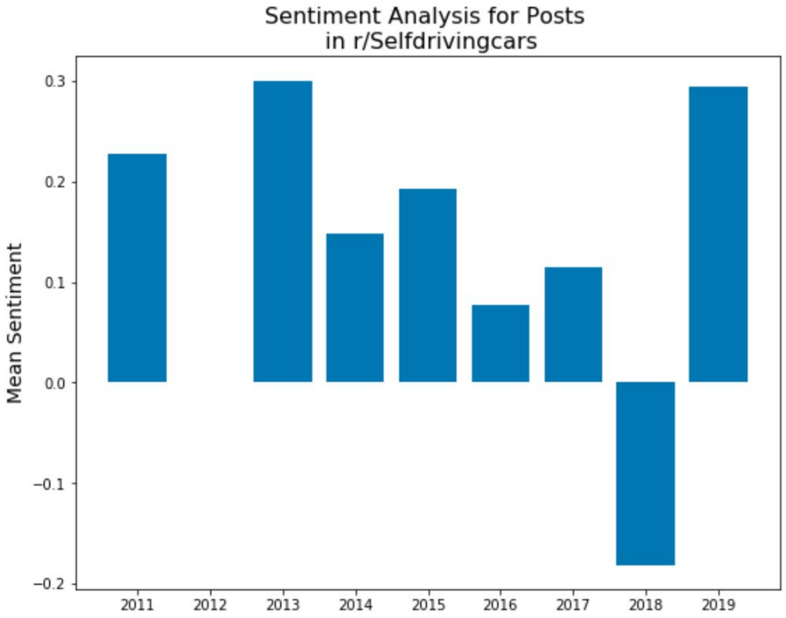
This is to be expected with increased sentence length and more variation in dialogue content

Posts: 89%

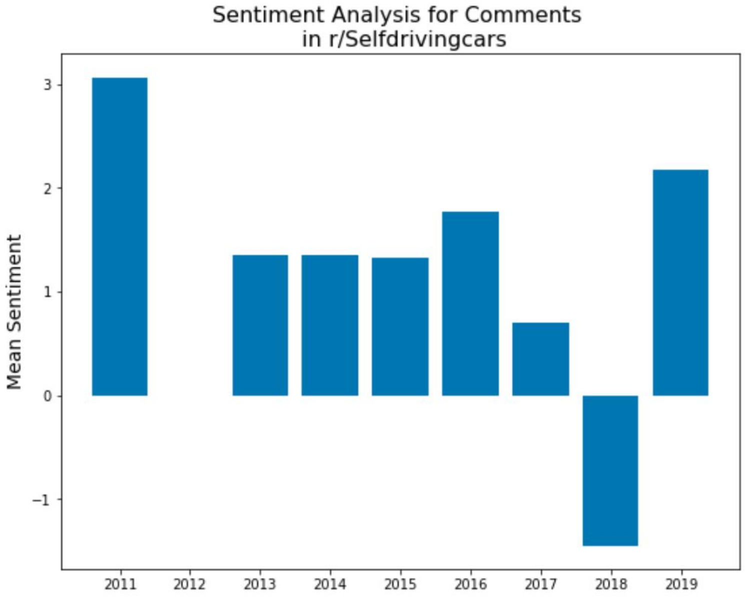
Comments: 83%



Sentiment Analysis for r/Selfdrivingcars



Sentiment Analysis for Posts in r/Selfdrivingcars



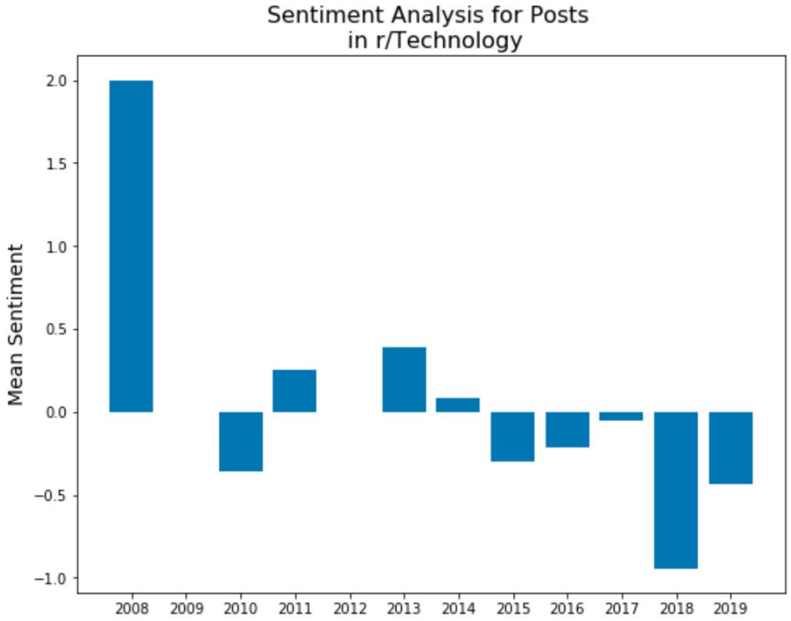
Sentiment Analysis for Comments in r/Selfdrivingcars

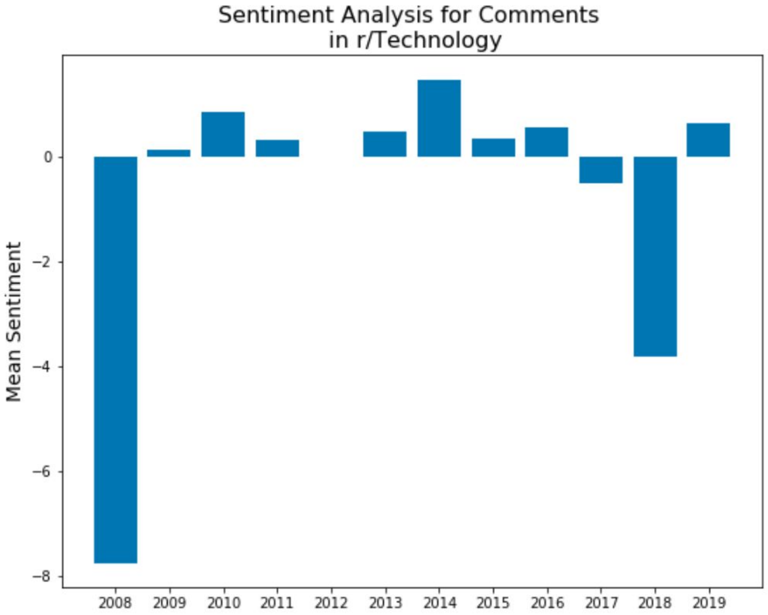
High sentiment across the entire lifespan with the exception of 2018, due to the death of pedestrian in Arizona after Uber self-driving accident

Afinn Sentiment Analysis Package

https://github.com/fnielsen/afinn

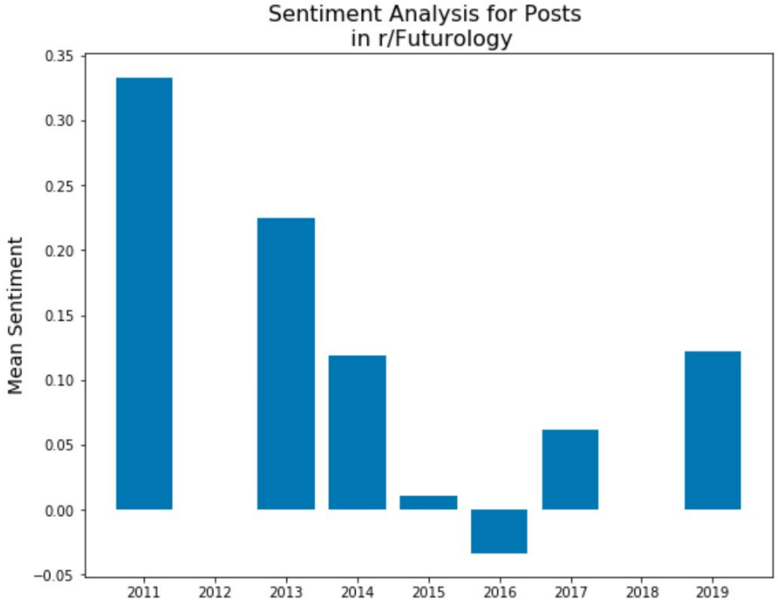
Sentiment Analysis for r/Technology

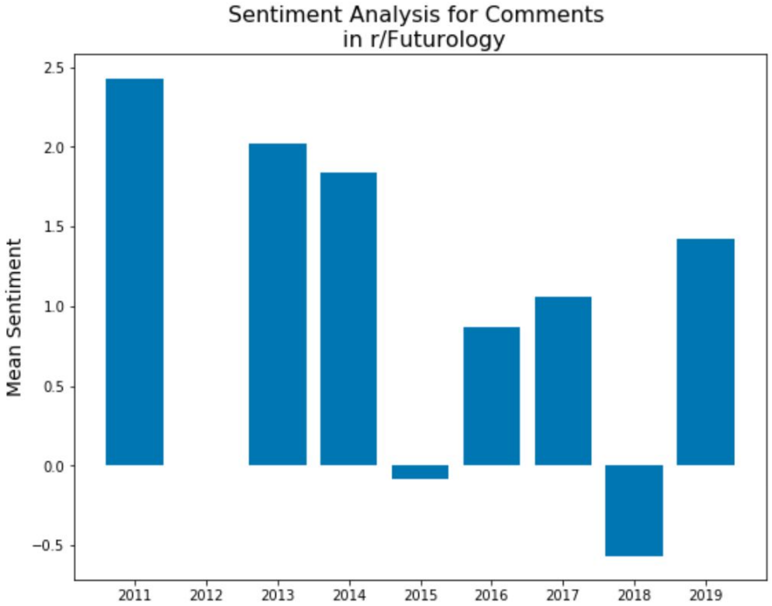




Technology has the lowest
sentiment of the three subreddits

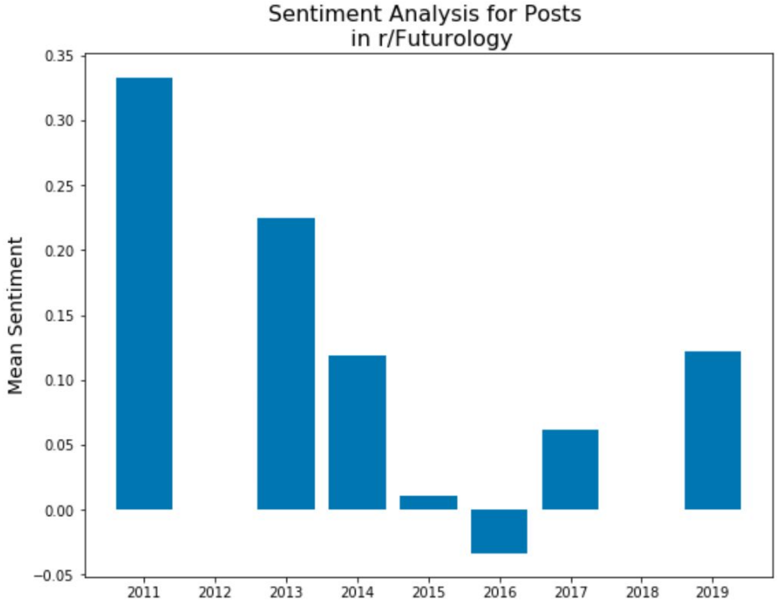
Sentiment Analysis for r/Futurology

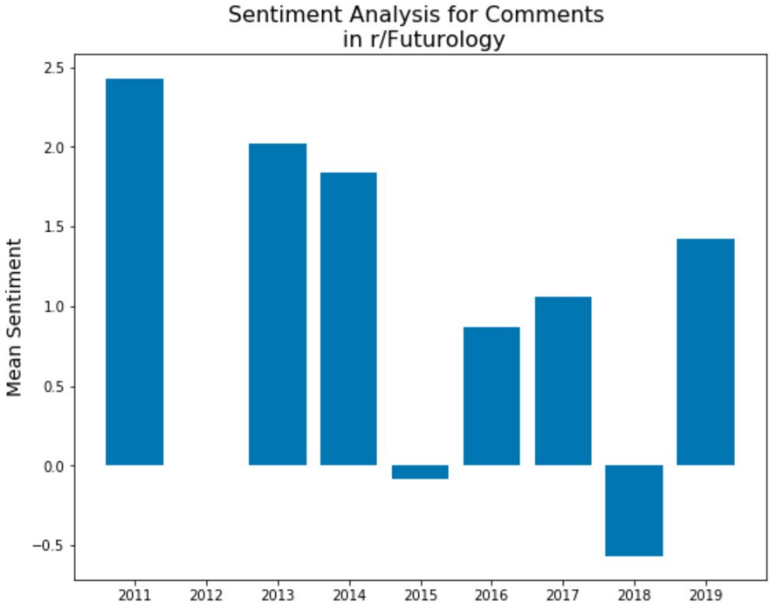




Higher Sentiment overall than r/Technology. Decline from 2011 - 2016, but trending upwards in past three years

r/Futurology contains relatively high Sentiment





Higher over Sentiment overall than r/Technology. Decline from 2011 - 2016, but on the way back up in past three years

Conclusions & Improvements

- Conduct more analysis and model tuning with increased features, utilizing more of the data collected. Specifically training on the comments.

- Incorporate Sentiment Scores into Classification Model

- Conduct Multi-class Classification among all three subreddits

- Investigate Posts and Comments that were misclassified, and along with analyzing their sentiment

- Investigate potential connection of news events to drops or spikes in sentiment

Improvements in Modeling and Analysis

- Conduct more analysis and model tuning with increased features, utilizing more of the data collected. Specifically training on the comments.
- Incorporate Sentiment Scores into Classification Model
- Conduct Multi-class Classification among all three subreddits
- Investigate Posts and Comments that were misclassified, and along with analyzing their sentiment
- Investigate potential connection of news events to drops or spikes in sentiment