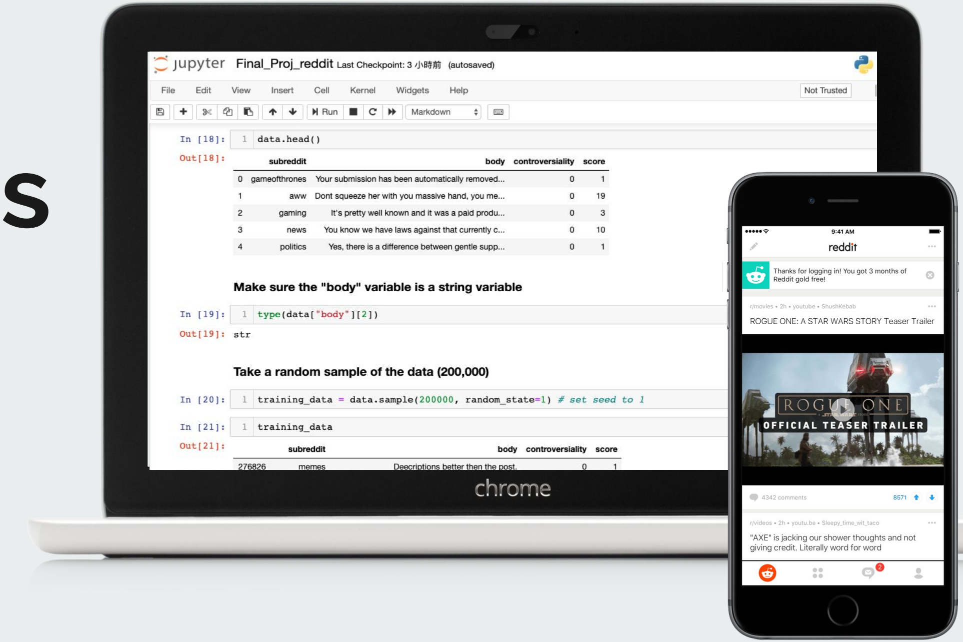
Applied NLP on Reddit Comments

Final Presentation

Alex Mead & Wafer Hsu



```
In [18]: data.head()
```

Out[18]:

	subreddit	body	controversiality	score
0	gameofthrones	Your submission has been automatically removed...	0	1
1	aww	Dont squeeze her with you massive hand, you me...	0	19
2	gaming	It's pretty well known and it was a paid produ...	0	3
3	news	You know we have laws against that currently c...	0	10
4	politics	Yes, there is a difference between gentle supp...	0	1

Make sure the "body" variable is a string variable

```
In [19]: type(data["body"][2])
Out[19]: str
```

Take a random sample of the data (200,000)

```
In [20]: training_data = data.sample(200000, random_state=1) # set seed to 1
In [21]: training_data
```

Out[21]:

	subreddit	body	controversiality	score
276826	memes	Decriptions better then the post.	0	1

Thanks for logging in! You got 3 months of Reddit gold free!

r/movies • 2h • youtube • ShushKebab

ROGUE ONE: A STAR WARS STORY Teaser Trailer

4342 comments 8571

r/videos • 2h • youtu.be • Sleepy_time_wit_taco

"AXE" is jacking our shower thoughts and not giving credit. Literally word for word

American University | April, 23, 2020

Outline

Executive Summary

Goals

Methods

Results

1. Regression
2. Topic Modeling
3. Clustering
4. Classification

Q&A

Executive Summary

Executive Summary

1. Attempted through **multiple threads**: sentiment analysis, linear modelling, topic modelling, and through more traditional machine learning algorithms such as clustering and classification
2. Subsetted original dataset to one subreddit – **r/politics** – for the majority of the analysis
3. Cleaned, normalized, and created variables of interest
4. Created a **net sentiment score variable** which was used in multiple regression models
   - The models we trained proved to be weak at predicting the score of reddit comments
5. Assigned to one of ten clusters based on document similarity
   - Results suggested that the model accurately clustered comments based on their underlying topics
6. Original dataset was re-visited and re-subsetted to see if it was possible to train a classification model to tag comments from similar themed subreddits
   - The results showed that the classification algorithm was able to be trained with remarkable accuracy

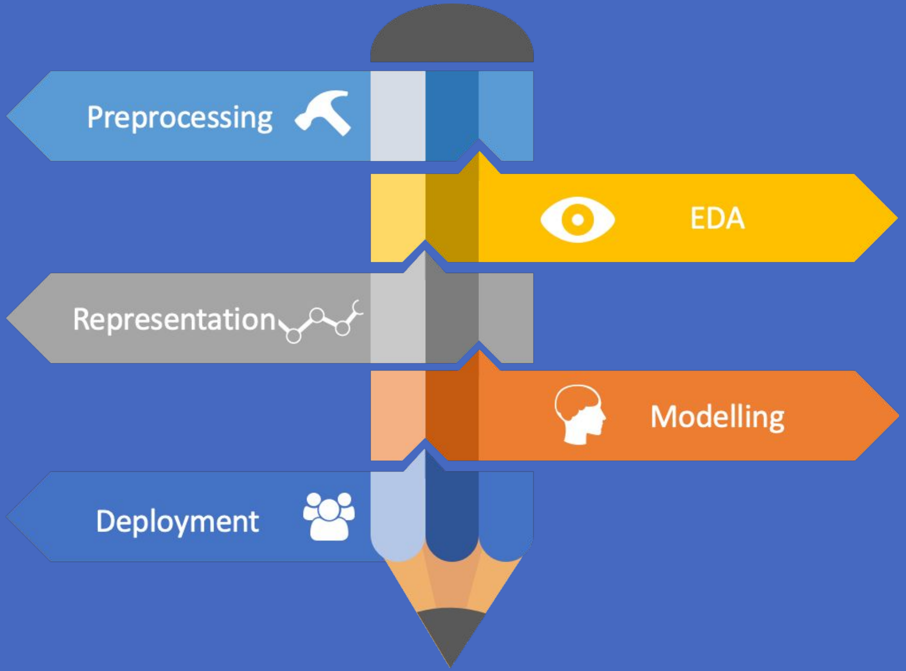
Sentiment Scoring → Regression → Topic Modelling → Clustering → Classification

Goals

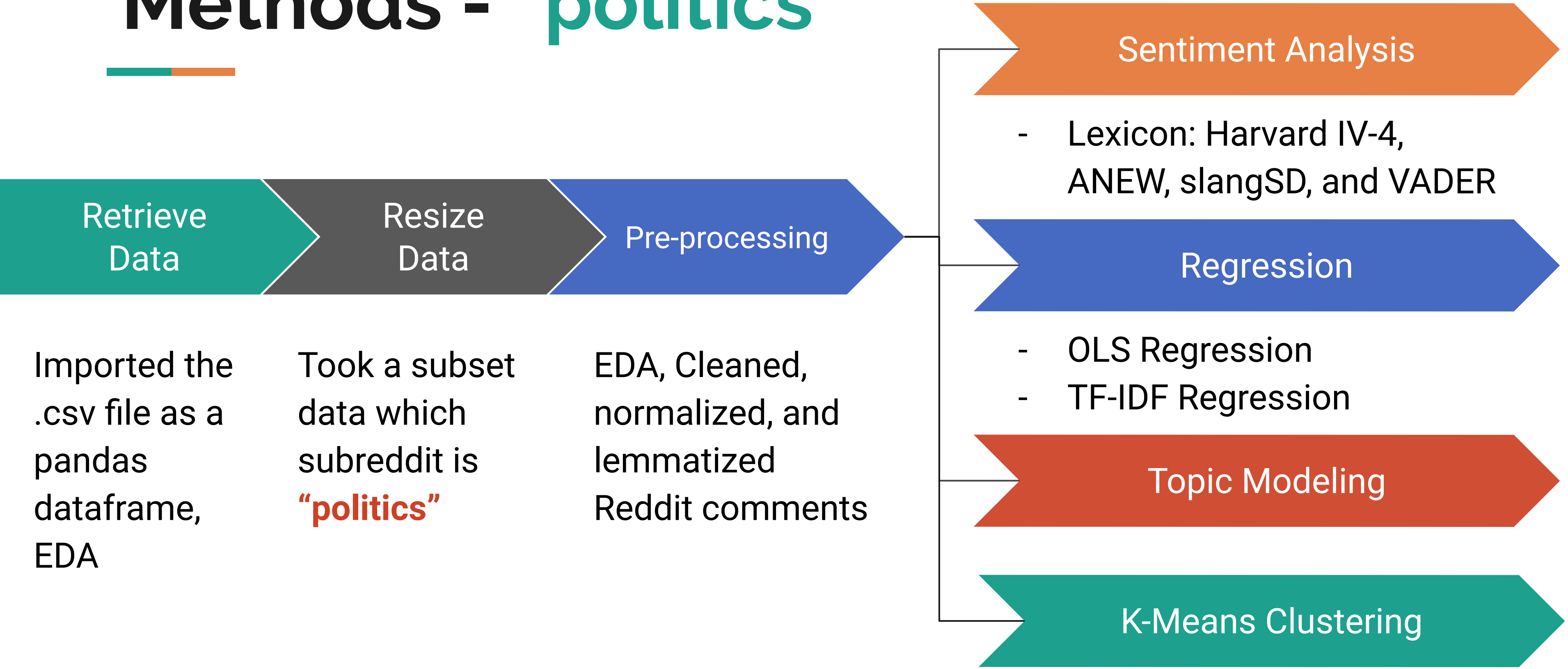
Goals

- See whether comment valence contributes to its score - Regression
    - TF-IDF matrix
    - General linear regression on pandas dataframe
- Analyze the topic distribution of the corpus - Topic Modelling
- Find popular subtopics via document clustering - Clustering
    - What do people like to talk about in this subreddit?
    - What language do they use?
- Train a clustering algorithm to classify subreddit - Classification
    - Can a classification algorithm be trained to predict what subreddit a comment belongs to?

Methods



Methods - “politics”



Retrieve Data

Imported the .csv file as a pandas dataframe, EDA

Resize Data

Took a subset data which subreddit is **“politics”**

Pre-processing

EDA, Cleaned, normalized, and lemmatized Reddit comments

Sentiment Analysis

- Lexicon: Harvard IV-4, ANEW, slangSD, and VADER

Regression

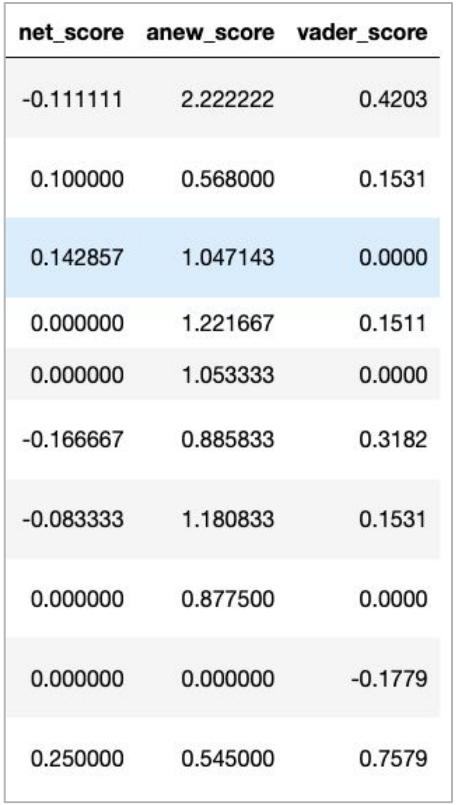
- OLS Regression
- TF-IDF Regression

Topic Modeling

K-Means Clustering

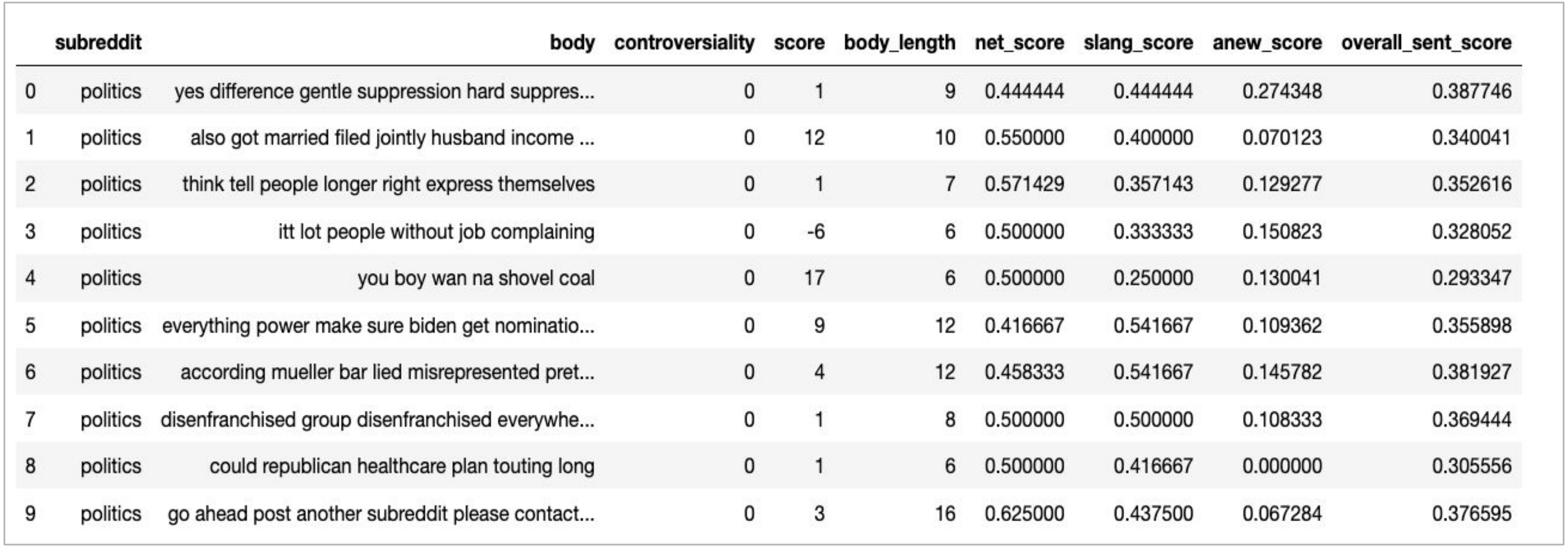
Sentiment Analysis

- Generated sentiment scores using different datasets
  - Harvard IV-4
  - ANEW
  - SlangSD
  - VADER
- Cleaned each dictionary set
- Different scales → Rescale using `MinMaxScaler()`



net_score	anew_score	vader_score
-0.111111	2.222222	0.4203
0.100000	0.568000	0.1531
0.142857	1.047143	0.0000
0.000000	1.221667	0.1511
0.000000	1.053333	0.0000
-0.166667	0.885833	0.3182
-0.083333	1.180833	0.1531
0.000000	0.877500	0.0000
0.000000	0.000000	-0.1779
0.250000	0.545000	0.7579

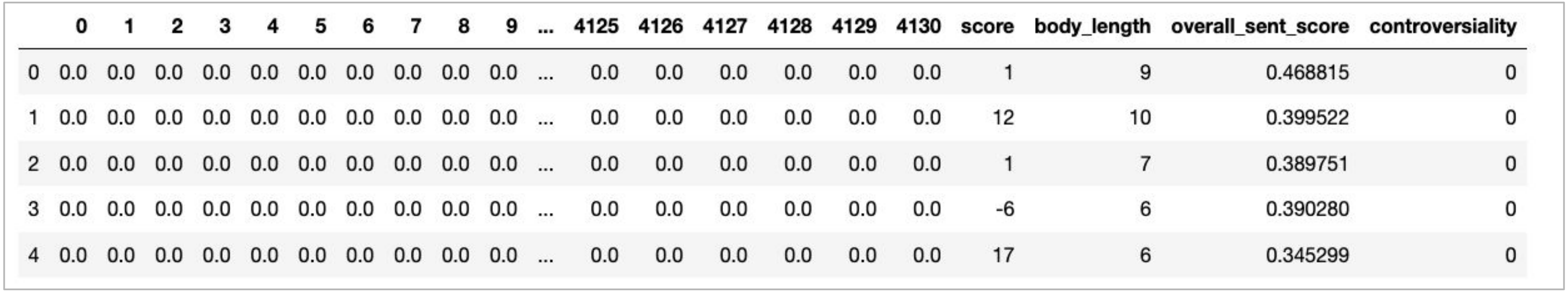
Sentiment Analysis



	subreddit	body	controversiality	score	body_length	net_score	slang_score	anew_score	overall_sent_score
0	politics	yes difference gentle suppression hard suppres...	0	1	9	0.444444	0.444444	0.274348	0.387746
1	politics	also got married filed jointly husband income ...	0	12	10	0.550000	0.400000	0.070123	0.340041
2	politics	think tell people longer right express themselves	0	1	7	0.571429	0.357143	0.129277	0.352616
3	politics	itt lot people without job complaining	0	-6	6	0.500000	0.333333	0.150823	0.328052
4	politics	you boy wan na shovel coal	0	17	6	0.500000	0.250000	0.130041	0.293347
5	politics	everything power make sure biden get nominatio...	0	9	12	0.416667	0.541667	0.109362	0.355898
6	politics	according mueller bar lied misrepresented pret...	0	4	12	0.458333	0.541667	0.145782	0.381927
7	politics	disenfranchised group disenfranchised everywhe...	0	1	8	0.500000	0.500000	0.108333	0.369444
8	politics	could republican healthcare plan touting long	0	1	6	0.500000	0.416667	0.000000	0.305556
9	politics	go ahead post another subreddit please contact...	0	3	16	0.625000	0.437500	0.067284	0.376595

Regression

- TF-IDF Matrix
    - (1, 1) grams, no more frequent than 90% of comments and in at least 10 comments
- Pandas Dataframe
- Train/test → 75% to 25%

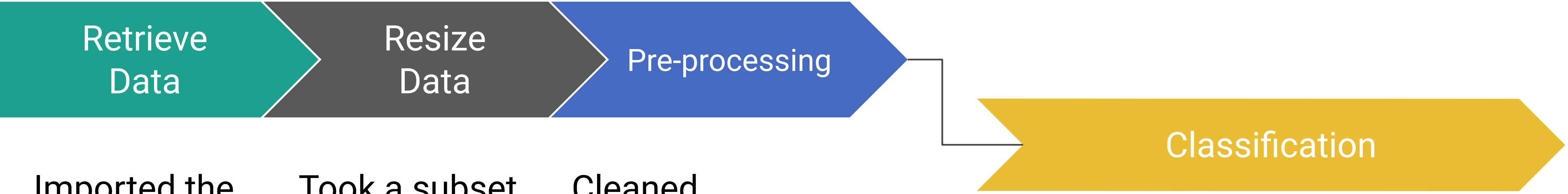


	0	1	2	3	4	5	6	7	8	9	...	4125	4126	4127	4128	4129	4130	score	body_length	overall_sent_score	controversiality
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1	9	0.468815	0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	12	10	0.399522	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1	7	0.389751	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	-6	6	0.390280	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	17	6	0.345299	0

Topic Modelling/Clustering

- TF-IDF vectorizing → NMF
- Print top topics and words
- K Means clustering using TF-IDF
    - 10 clusters
    - Hyperparameters (1,2) grams, minimum 10 comments, maximum 90% of comments

Methods - “sports”



Retrieve Data

Resize Data

Pre-processing

Classification

Imported the .csv file as a pandas dataframe, EDA

Took a subset data which subreddits are **“sports”**

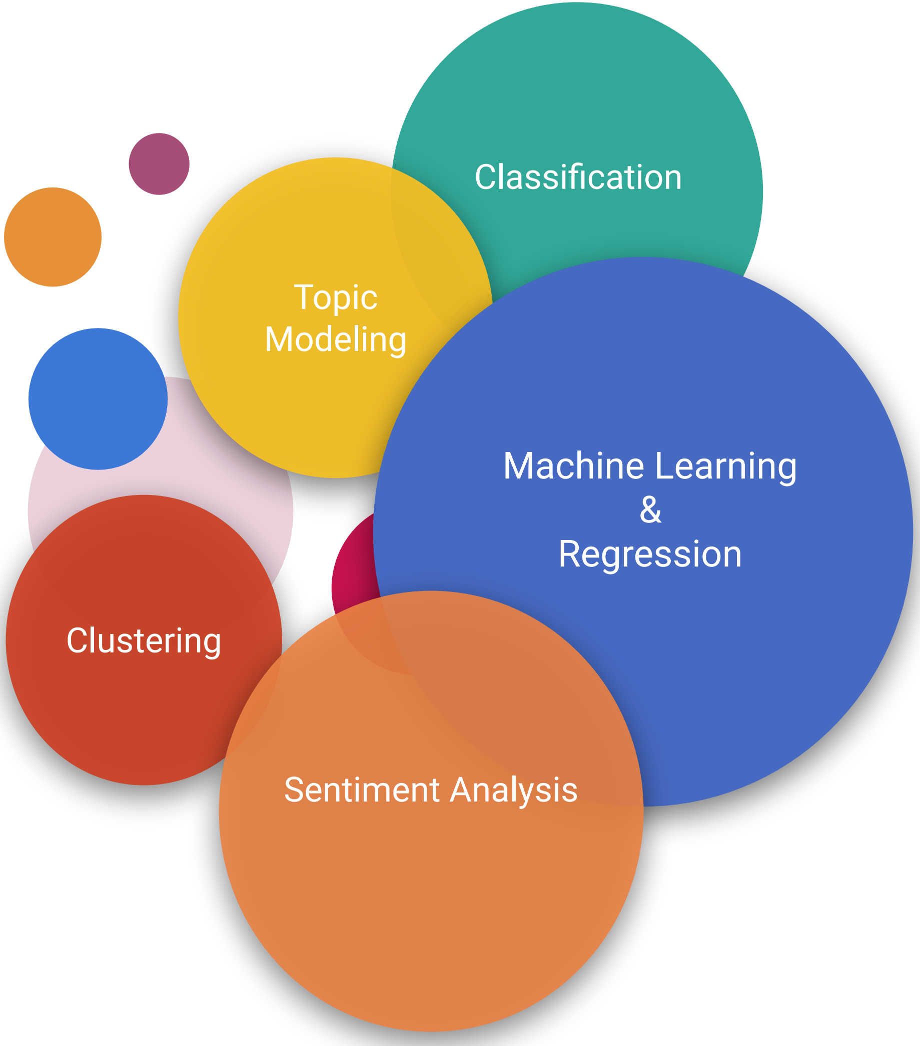
Cleaned, normalized, lemmatized Reddit comments, and create a new variable **“category”**

- 2 Classes
- 4 Classes

Classification

- Trained two classification models
- Re-defined our initial dataset
- Cleaned the body variable
- Subsetted to four subreddits (NBA, NFL, Hockey, Soccer)
  - Subreddits that were similar to each other but had subtle differences
  - 80/20 train/test split
- Classified by two subreddits and by four subreddits

Results



Classification

Topic Modeling

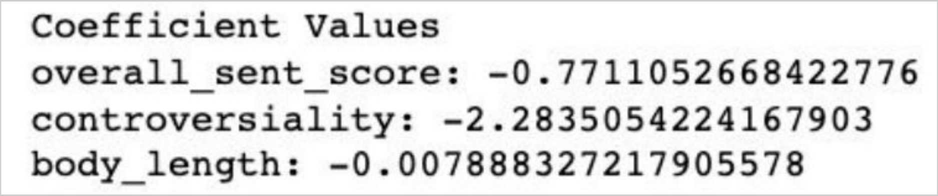
Machine Learning & Regression

Clustering

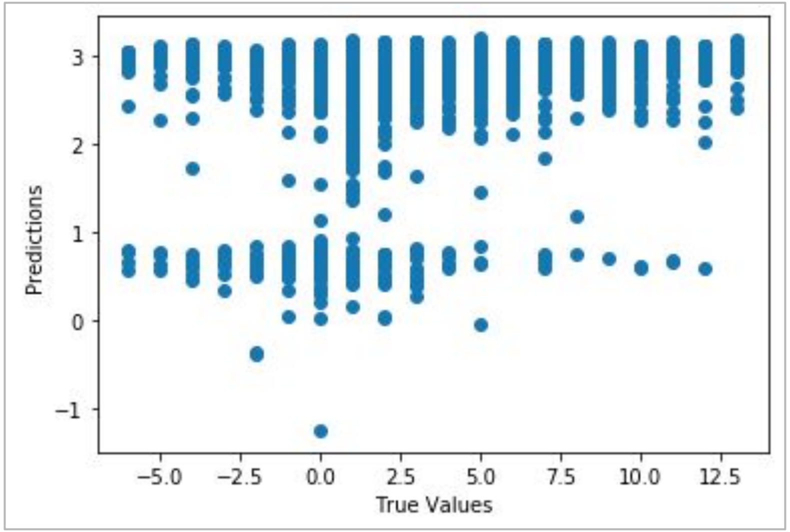
Sentiment Analysis

OLS - Regression

- **Multiple Regression**
  - Independent Variable (x): "overall_sent_score", "controversiality", "body_length"
  - Dependent Variable (y): "score"
  - R-squared = **0.0246**, RMSE = **3.19**

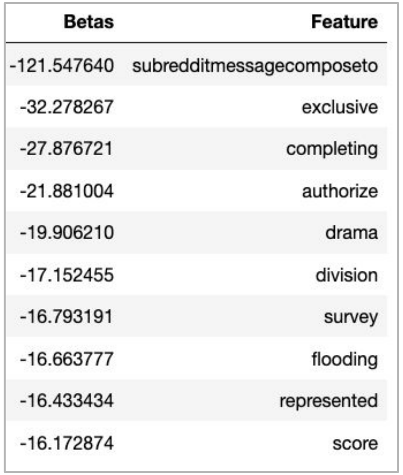


```
Coefficient Values
overall_sent_score: -0.7711052668422776
controversiality: -2.2835054224167903
body_length: -0.007888327217905578
```

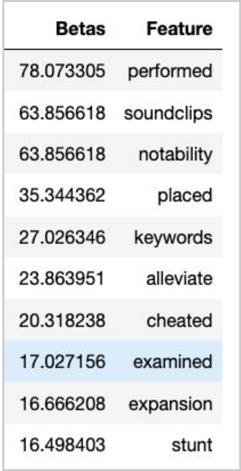


TFIDF - Regression

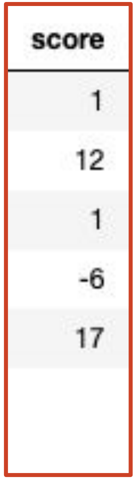
- Independent Variable (x): BoW transform matrix + "body_length" + "controversiality" + "overall_sent_score"
- Dependent Variable (y): score



Betas	Feature
-121.547640	subredditmessagecomposeto
-32.278267	exclusive
-27.876721	completing
-21.881004	authorize
-19.906210	drama
-17.152455	division
-16.793191	survey
-16.663777	flooding
-16.433434	represented
-16.172874	score



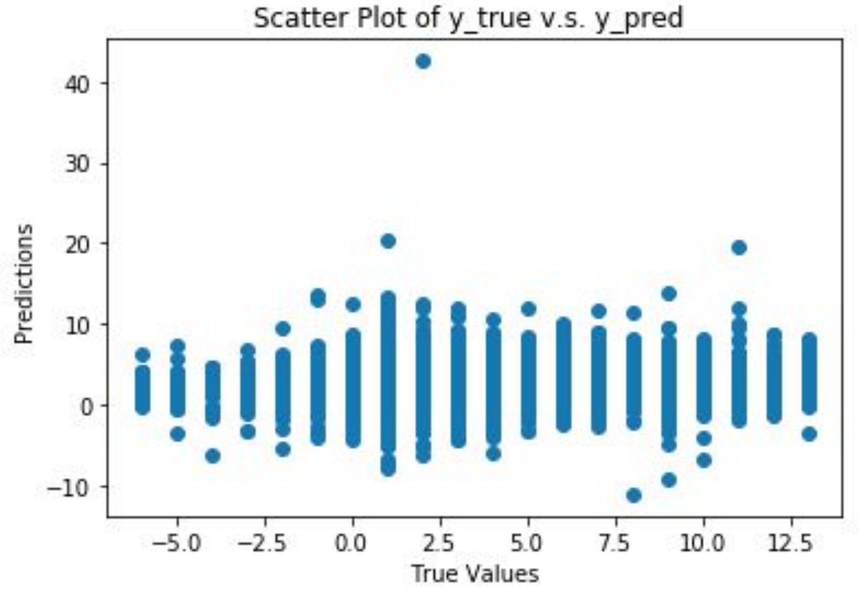
Betas	Feature
78.073305	performed
63.856618	soundclips
63.856618	notability
35.344362	placed
27.026346	keywords
23.863951	alleviate
20.318238	cheated
17.027156	examined
16.666208	expansion
16.498403	stunt



score
1
12
1
-6
17

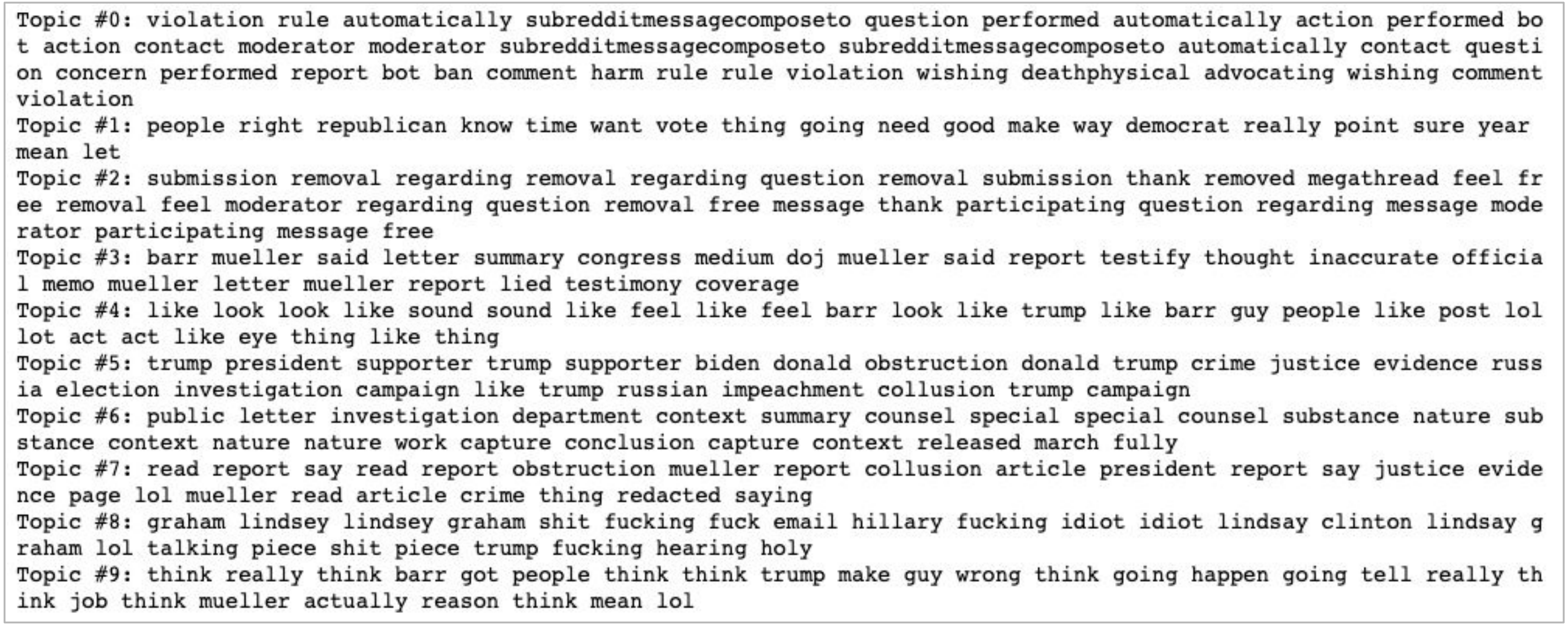
TFIDF - Regression

- R-squared = **-0.3442**, RMSE = **3.74565**



Topic Modeling

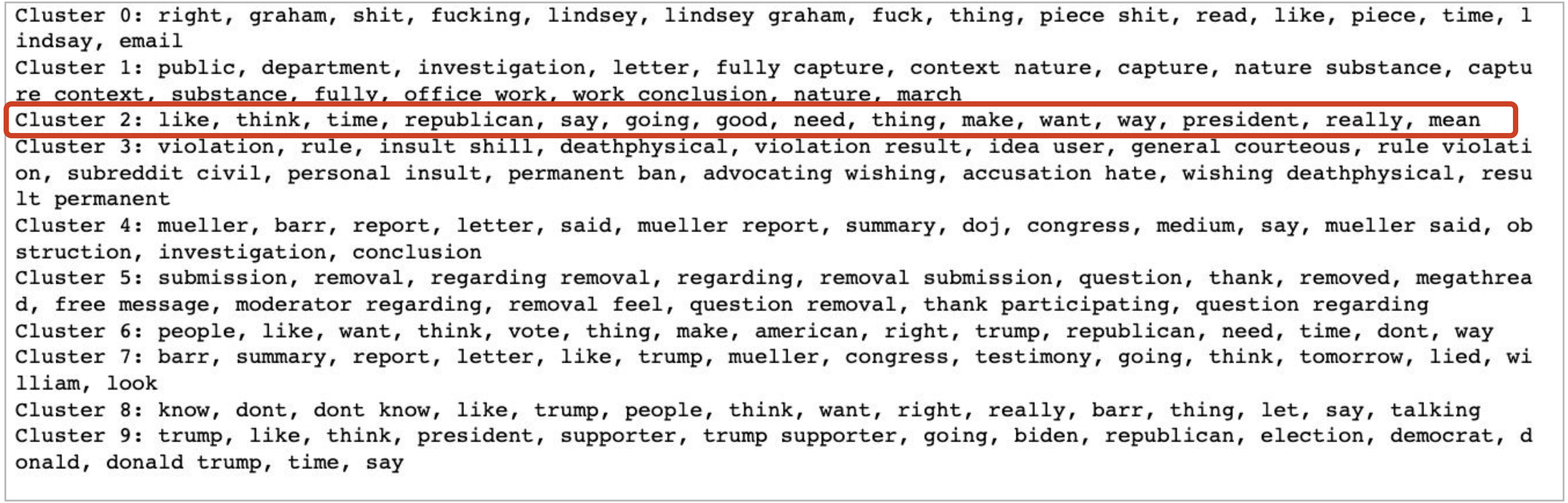
- Retrieve **top 20 words** from each topic



```
Topic #0: violation rule automatically subredditmessagecomposeto question performed automatically action performed bo
t action contact moderator moderator subredditmessagecomposeto subredditmessagecomposeto automatically contact questi
on concern performed report bot ban comment harm rule rule violation wishing deathphysical advocating wishing comment
violation
Topic #1: people right republican know time want vote thing going need good make way democrat really point sure year
mean let
Topic #2: submission removal regarding removal regarding question removal submission thank removed megathread feel fr
ee removal feel moderator regarding question removal free message thank participating question regarding message mode
rator participating message free
Topic #3: barr mueller said letter summary congress medium doj mueller said report testify thought inaccurate officia
l memo mueller letter mueller report lied testimony coverage
Topic #4: like look look like sound sound like feel like feel barr look like trump like barr guy people like post lol
lot act act like eye thing like thing
Topic #5: trump president supporter trump supporter biden donald obstruction donald trump crime justice evidence russ
ia election investigation campaign like trump russian impeachment collusion trump campaign
Topic #6: public letter investigation department context summary counsel special special counsel substance nature sub
stance context nature nature work capture conclusion capture context released march fully
Topic #7: read report say read report obstruction mueller report collusion article president report say justice evide
nce page lol mueller read article crime thing redacted saying
Topic #8: graham lindsey lindsey graham shit fucking fuck email hillary fucking idiot idiot lindsay clinton lindsay g
raham lol talking piece shit piece trump fucking hearing holy
Topic #9: think really think barr got people think think trump make guy wrong think going happen going tell really th
ink job think mueller actually reason think mean lol
```

Clustering

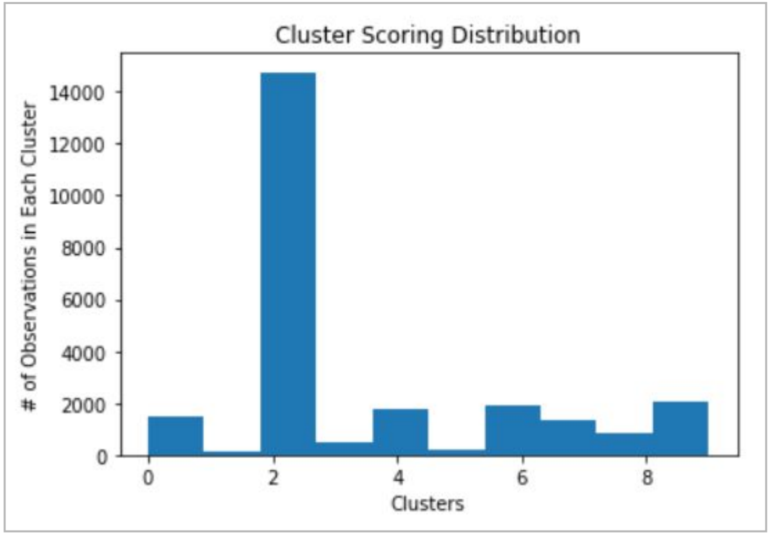
- Retrieve **top 10 words** which are nearing to the each centroids



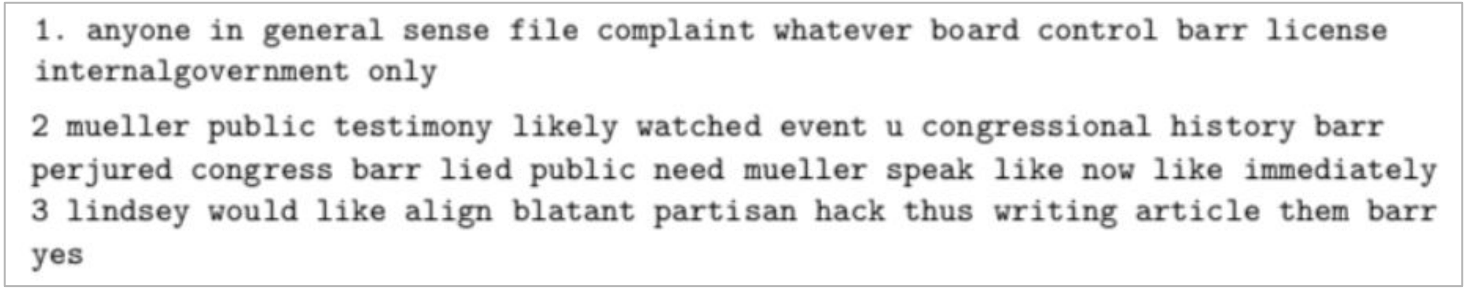
```
Cluster 0: right, graham, shit, fucking, lindsey, lindsey graham, fuck, thing, piece shit, read, like, piece, time, l
indsay, email
Cluster 1: public, department, investigation, letter, fully capture, context nature, capture, nature substance, captu
re context, substance, fully, office work, work conclusion, nature, march
Cluster 2: like, think, time, republican, say, going, good, need, thing, make, want, way, president, really, mean
Cluster 3: violation, rule, insult shill, deathphysical, violation result, idea user, general courteous, rule violati
on, subreddit civil, personal insult, permanent ban, advocating wishing, accusation hate, wishing deathphysical, resu
lt permanent
Cluster 4: mueller, barr, report, letter, said, mueller report, summary, doj, congress, medium, say, mueller said, ob
struction, investigation, conclusion
Cluster 5: submission, removal, regarding removal, regarding, removal submission, question, thank, removed, megathrea
d, free message, moderator regarding, removal feel, question removal, thank participating, question regarding
Cluster 6: people, like, want, think, vote, thing, make, american, right, trump, republican, need, time, dont, way
Cluster 7: barr, summary, report, letter, like, trump, mueller, congress, testimony, going, think, tomorrow, lied, wi
lliam, look
Cluster 8: know, dont, dont know, like, trump, people, think, want, right, really, barr, thing, let, say, talking
Cluster 9: trump, like, think, president, supporter, trump supporter, going, biden, republican, election, democrat, d
onald, donald trump, time, say
```

Clustering

- Counts of each clusters
    - Unbalanced distribution
- View the text from each cluster
    - Find the index of comments from each cluster to see if the contents are similar



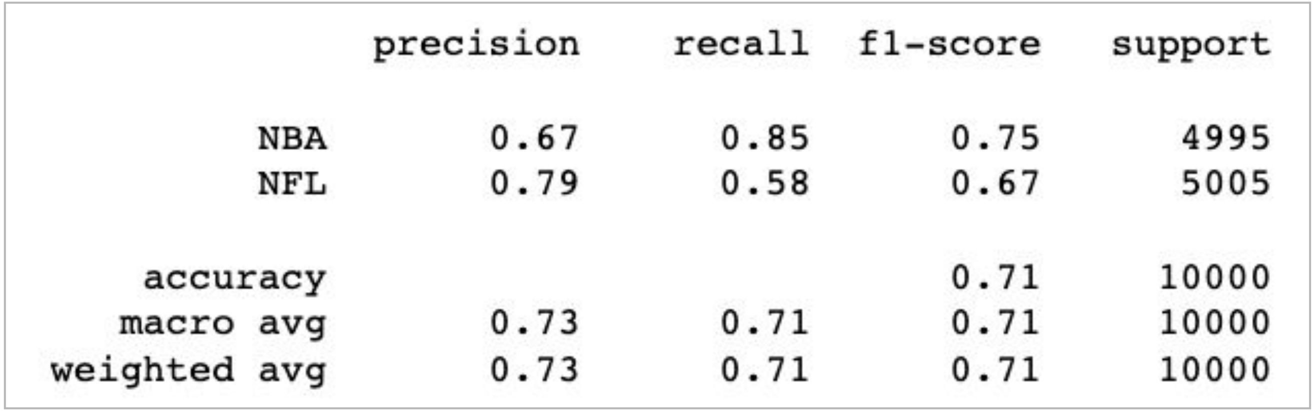
Cluster = 7
Index = 16, 209, 24956



```
1. anyone in general sense file complaint whatever board control barr license
internalgovernment only
2 mueller public testimony likely watched event u congressional history barr
perjured congress barr lied public need mueller speak like now like immediately
3 lindsey would like align blatant partisan hack thus writing article them barr
yes
```

Classification

- Pick subreddits that were similar enough to be tough to train a model on but also distinct enough to tell differences
- Sports themes subreddits seemed to fit the mold
- **2 Classes - "nba", "nfl"**
  - F1-score 0.71, Contribute to a better result



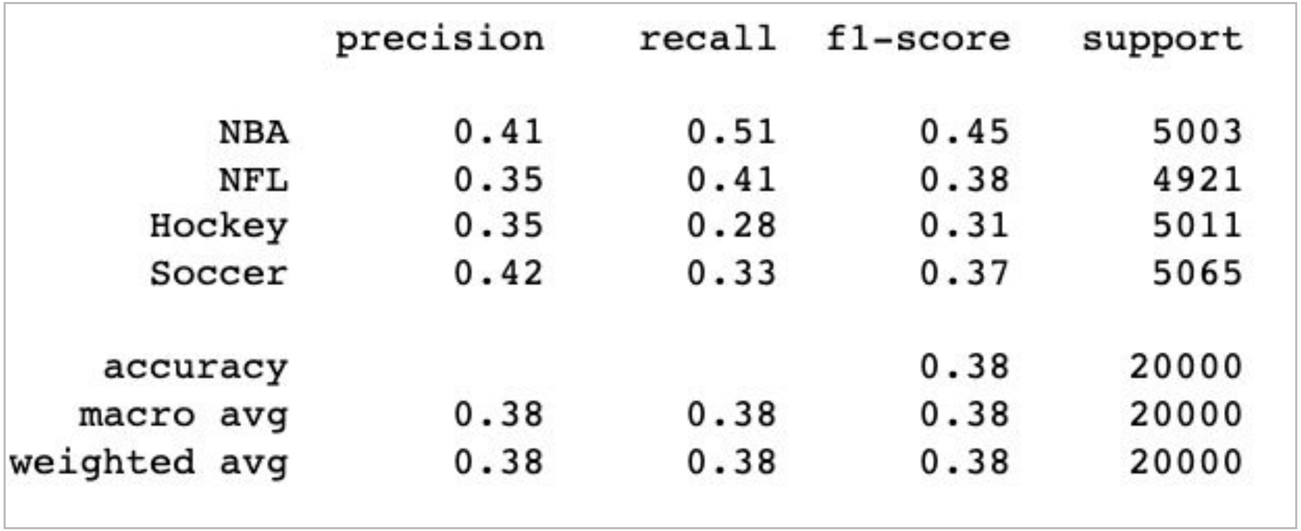
```
              precision    recall  f1-score   support

         NBA       0.67      0.85      0.75      4995
         NFL       0.79      0.58      0.67      5005

    accuracy                           0.71     10000
   macro avg       0.73      0.71      0.71     10000
weighted avg       0.73      0.71      0.71     10000
```

Classification

- 4 Classes - "nba", "nfl", "hockey", "soccer"
    - Tougher but not bad
    - F1-score 0.39 → Less accurate and precise

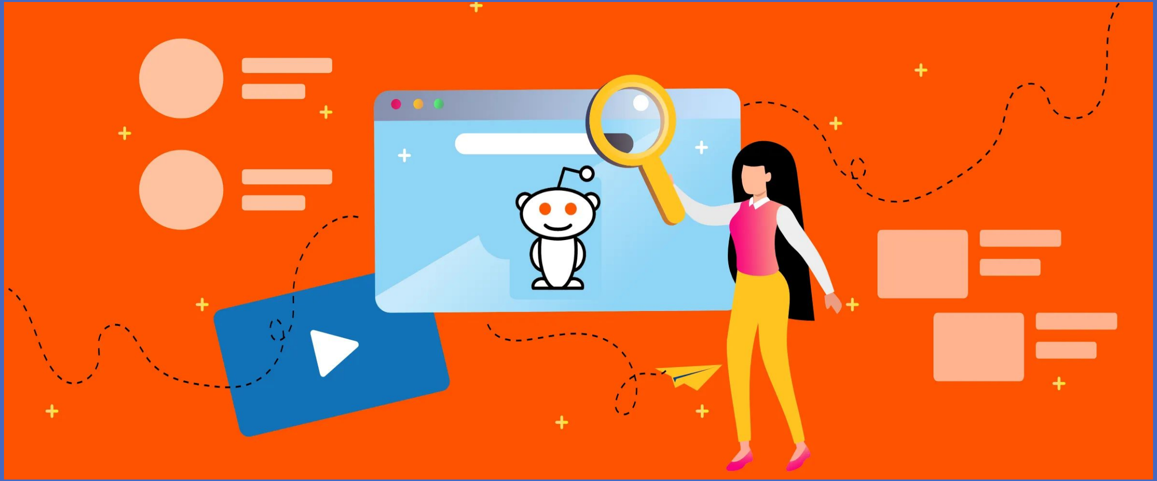


```
              precision    recall  f1-score   support

         NBA       0.41      0.51      0.45      5003
         NFL       0.35      0.41      0.38      4921
      Hockey       0.35      0.28      0.31      5011
      Soccer       0.42      0.33      0.37      5065

    accuracy                           0.38     20000
   macro avg       0.38      0.38      0.38     20000
weighted avg       0.38      0.38      0.38     20000
```

Questions



References

Dataset retrieved from Kaggle

1 million reddit comments on Kaggle

Email

Alex Mead: am7306a@student.american.edu

Wafer Hsu: wh0225a@student.american.edu