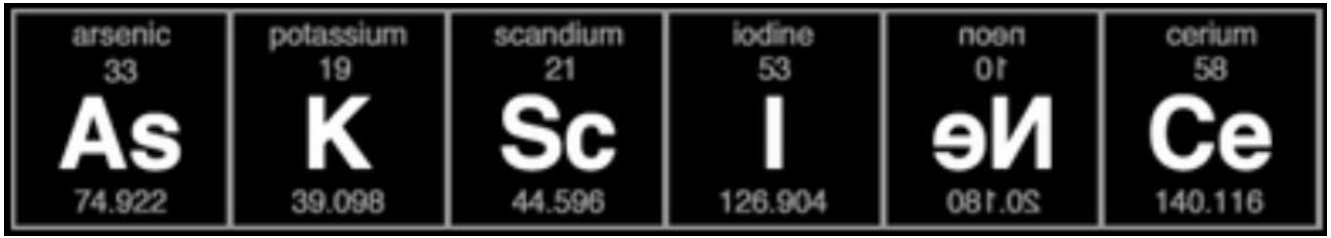
Ikhwan Jerome Yuzheng

Project 3:
Web APIs & Text Classification



Business Problem

How can we **accurately** and **consistently** **auto-filter out troll posts** that don't belong to the r/askscience sub-reddit, to **save the moderators' time**?

r/jokes

r/askscience

Technical Overview

- Data Collection
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Modelling
- Evaluation
- Conclusion/ Recommendations

Data Collection

- Web scrape, prevent blockage by reddit
  - Not by ‘sleep’ duration, but by **randomizing the User-agent EVERY scrape. No waiting time needed**!

```
ua = str(random.randint(1,100))
res = requests.get(current_url, headers={'User-agent': ua})
```

- Explore text structure

Data Cleaning

- Duplicated posts from scrapping

title	selftext
A woman on Vacation rings home	She asks her husband, "How's my cat doing?"\n\...
A woman on Vacation rings home	She asks her husband, "How's my cat doing?"\n\...

title	selftext	subreddit	text
Knock knock	Who's there? My name is My name is who? My ...	1	Knock knock Who's there? My name is My name ...
Knock knock	Who's there? Dishes. Dishes, who? This is S...	1	Knock knock Who's there? Dishes. Dishes, who...
Knock knock	"Who's there?" "There." "There who?" "Yoda,...	1	Knock knock "Who's there?" "There." "There w...
Knock knock	So, my nephew comes running up to me super exc...	1	Knock knock So, my nephew comes running up to ...
Knock knock	Who's there? Interrupting cow Interrupting c...	1	Knock knock Who's there? Interrupting cow In...

Data Cleaning

- Formatting

Michael and Jerry are two third graders in the same school. One day, Michael told Jerry: “I just learned a neat trick that made me twenty bucks yesterday.” “Really? What’s the trick?” Asked Jerry.\n\n“It’s easy.” Michael said “Just go up to an adult and whisper in their ears: ‘I know everything about that dirty little secret of yours, now give me ten bucks, or else’ ; I’ve tried it on my parents last night and it totally worked!”\n\nExcited, Jerry went home after sc hool and immediately tried it on his mother. He walks up to her and whispered in her ears: “I know everything about t hat dirty little secret of yours, now give me ten bucks, or else.” His mother was immediately stunned, she gathered h erself together and replied: “Alright, here’s ten bucks, just whatever you do, do not tell your father.”\n\nJoyed, Je rry ran upstairs, found his father in the reading room and tried the trick on his father. He walks up to him and whis pered in his ears: “I know everything about that dirty little secret of yours, now give me ten bucks, or else.” Shock ed, his father scrambled for words: “What? But how did you... Never mind, fine, here. Just whatever you do, don’t tel l your mother.” And Jerry’s father handed him ten bucks.\n\nPocket full of cash and heart filled with excitement, Jer ry ran outside to the front yard and found their neighbor, Mr. Smith mowing his lawn. Jerry decides to try the trick on Mr. Smith as well. \n\nJerry ran up to Mr. Smith and said: “Mr. Smith, sir, there’s something I have to tell you.” Mr. Smith turned off his lawn mower and asked: “Well what is it, little buddy?” Jerry closed in and whispered in his ears: “I know everything about that dirty little secret of yours, now give me ten bucks, or else.”\n\n“I’m afraid I o we you more than ten bucks.” Said Mr. Smith: “So your mother told you huh? Well what’re you waiting for? Come here an d give me a hug, son.” True

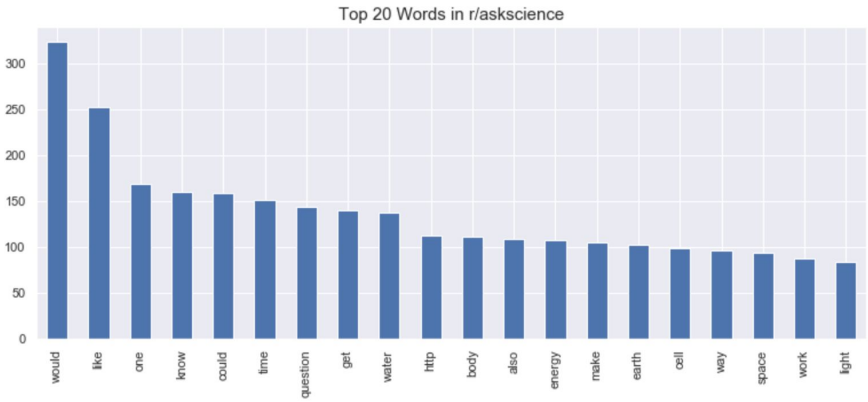
Exploratory Data Analysis

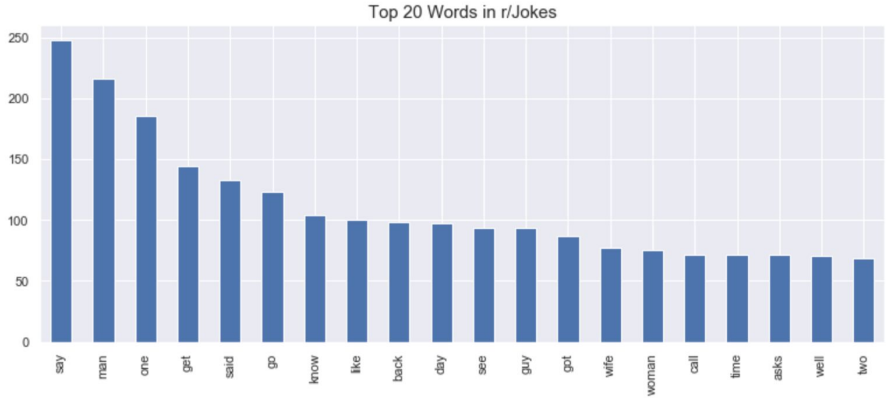
r/Jokes?

r/askscience?

Exploratory Data Analysis

Top 20 words in both subreddits

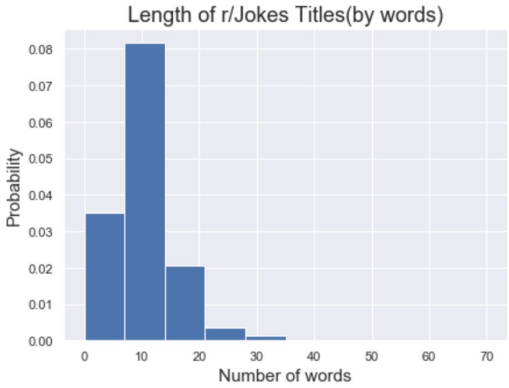


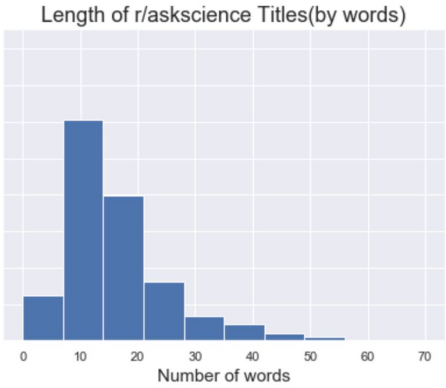


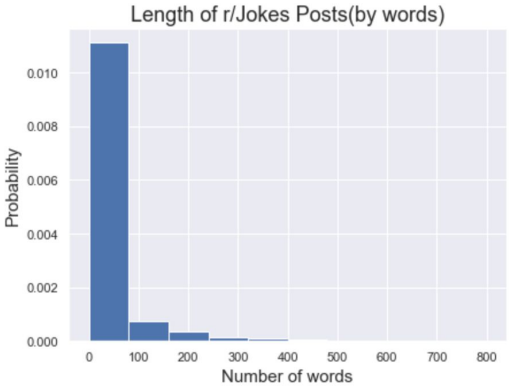
Common words found in both subreddits!

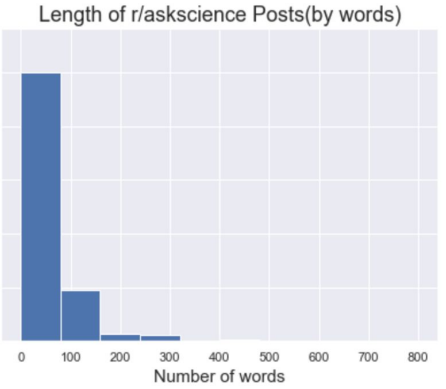
Exploratory Data Analysis

- What other relationships can we find that might aid us in classifying a particular post's subreddit?



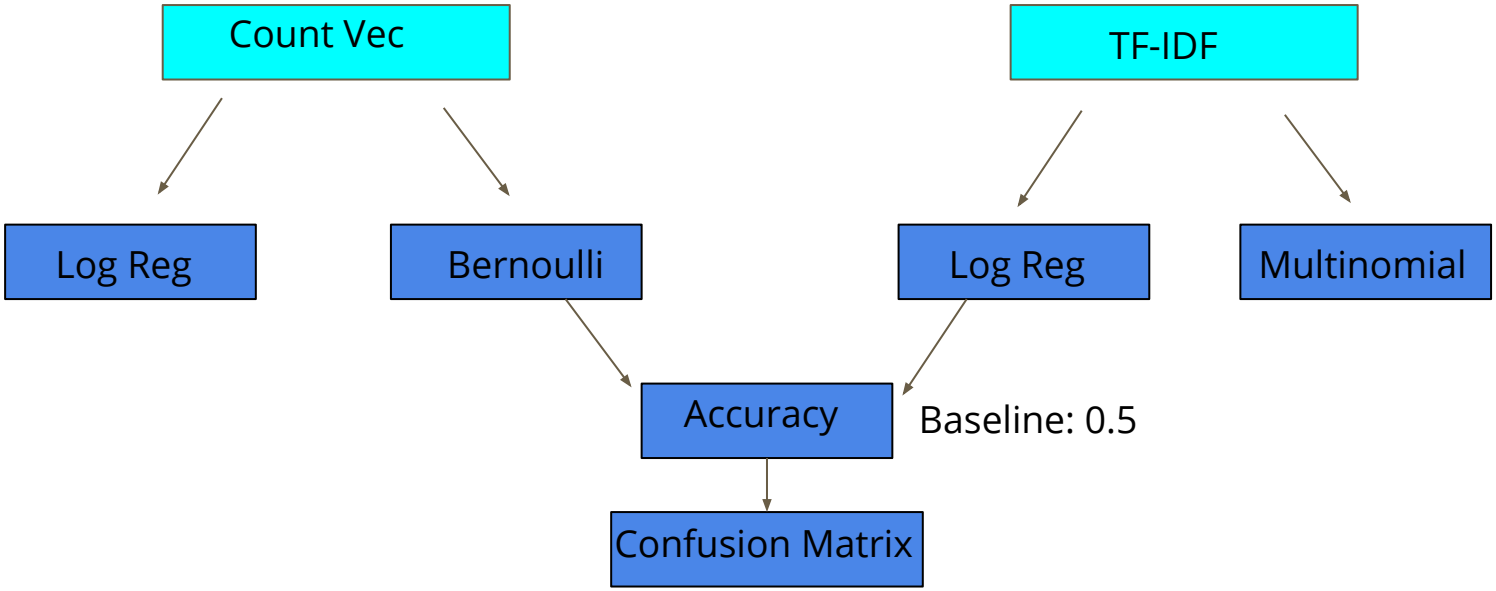






Techniques Used

- CountVectorizer, Tfidf (on **both** titles and selftext), then...
- logreg/MultinomialNB/BernoulliNB



Hyperparameters

Logistic Regression:

-'penalty': ['l1'],['l2']

-'C': np.arange(1, 5, 0.1),

-'warm_start': [True, False],

-'solver': ['lbfgs', 'liblinear']

Multinomial/Bernoulli:

-'fit_prior': [True,False],

-'alpha': np.arange(0, 1, 0.1)

Count Vectorizer with BernoullilNB

- Train accuracy score: 0.807
- Test accuracy score: 0.776

139 TN	18 FP
52 FN	104 TP

Count Vectorizer with Logistic Regression

- Train accuracy score: 0.835
- Test accuracy score: 0.783

139 TN	18 FP
49 FN	107 TP

TF-IDF with Logistic Regression

- Train accuracy score: 0.850
- Test accuracy score: 0.799

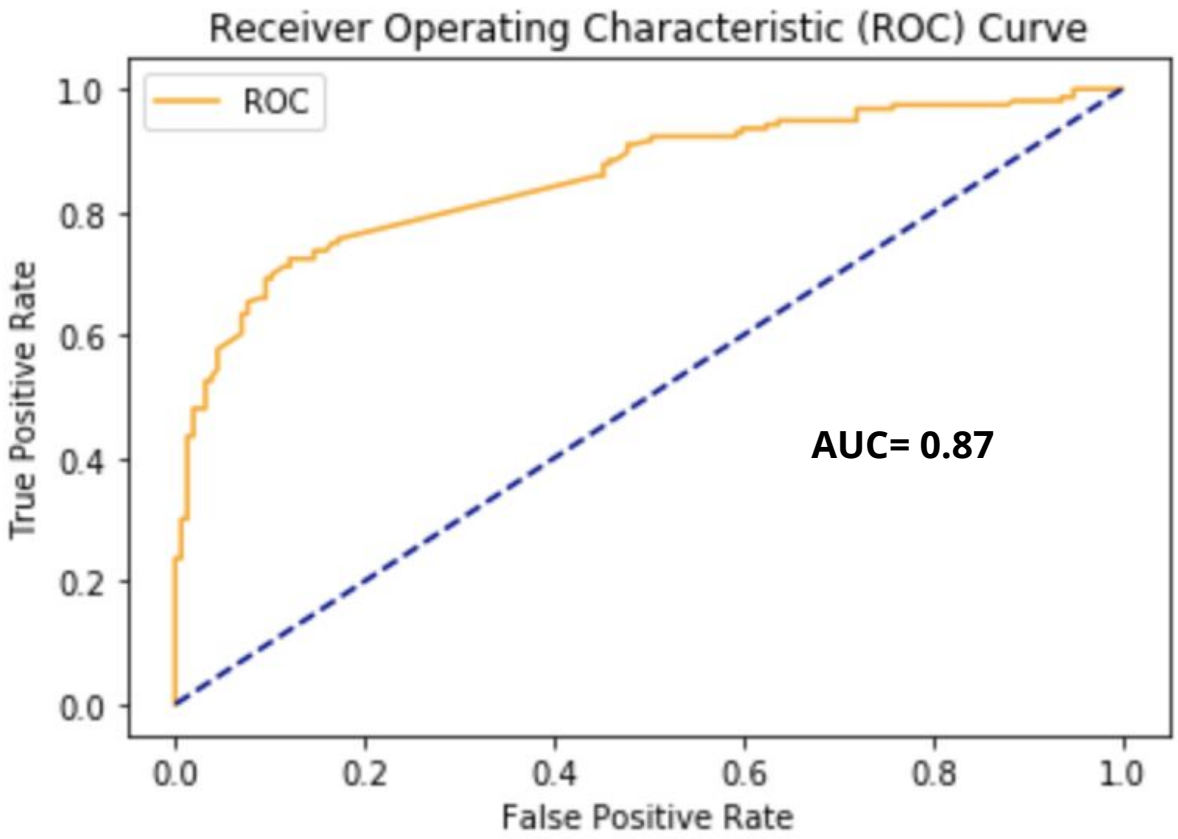
133 TN	24 FP
FN 39	TP 117

TF-IDF with MultinomialNB (Best Model)

- Train accuracy score: 0.821
- Test accuracy score: 0.793

- We are predicting words from Jokes subreddit 42% of the time!
- Misclassifying words only 20% of the time

130 TN	27 FP
38 FN	118 TP


Receiver Operating Characteristic (ROC) Curve
ROC
AUC= 0.87
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

From best model, hone further!

- NLP on **only titles**, NOT selftext
- **SelectKBest(chi2)** to reduce more features
- **Let algo itself pick the best** of: TF-IDF params, and best of logreg/mn/bn (pipe + gridsearchCV)

Why?

1. Titles hold the **essence** of the selftext
2. Chi2 **drops irrelevant features** relative to target
3. Best model only "good for the day": **won't always be mn, nor default TF-IDF params**

Titles ~~selftext~~

TF-IDF **(+gridsearchCV)** — 3632 features

↓

+SelectKBest(chi2) — 2500 features :) !

↓

+Best of logreg/mn/bn

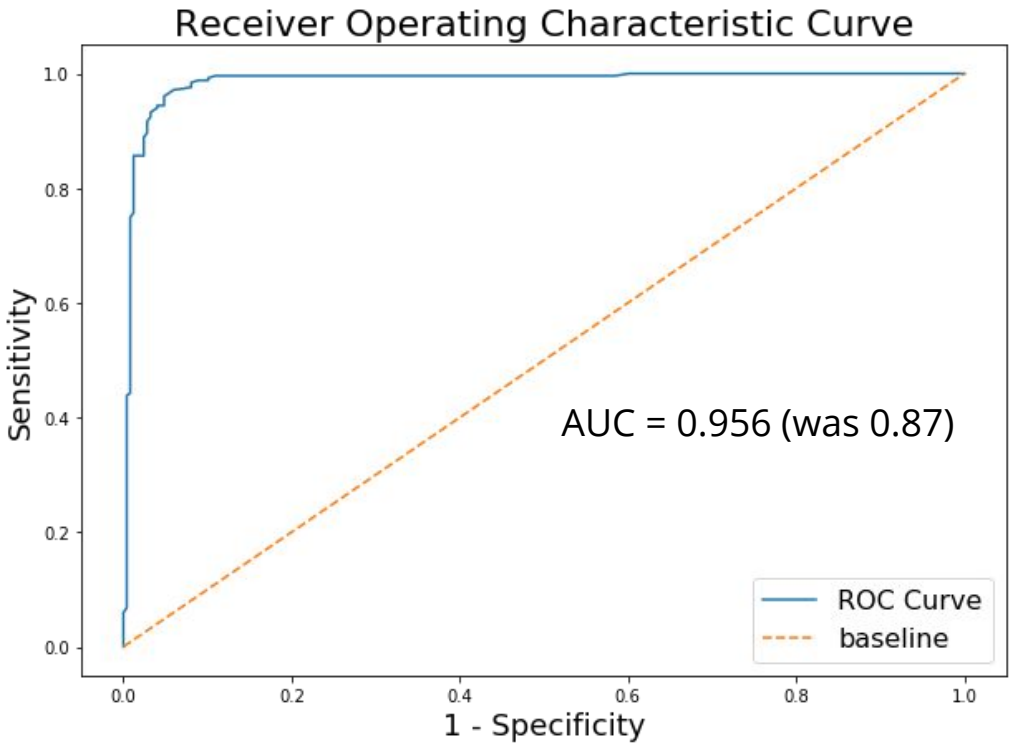
TF-IDF, *+SelectKBest (chi2)*, MultinomialNB *+only on titles* (Improved model)

- Train accuracy score: 0.977 (was 0.821)
- Test accuracy score: 0.956 (was 0.793)
- Test ROC-AUC score: 0.956 (was 0.87)

Significant improvement!

234 TN	14 FP
8 FN	243 TP

Results of improved model (ROC)



Results of improved model

- Classifies well even if with similar words!

‘Ever have amnesia and deja vu at the same time?’

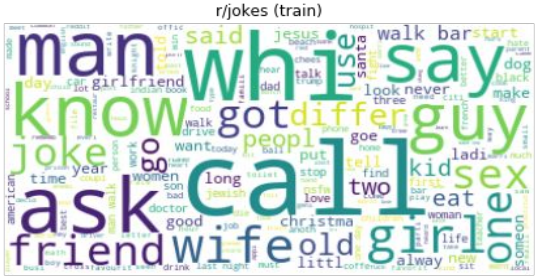
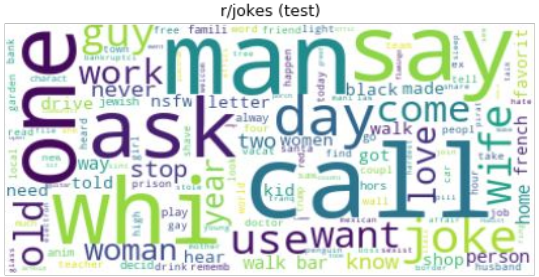
VS

‘Does a person suffering from amnesia retain the personality traits formed from/during the experiences they can no longer remember?’

Actual

Predict

Results of improved model (wordcloud)

	train	test
r/jokes		
r/askscience		

Conclusion

- Best-est model: **TF-IDF >> SelectKBest(chi2) >> multinomialNB (for now)** (accuracy, ROC-AUC >90+%)

- Should only analyse **titles**

- Need periodic re-training to stay “in-trend”. Proposed algo promises to be a **current and dynamic** model - it **self decides best** hyperparams, and best of logreg/mn/bn, for each re-training

Limitations & Future work

- Does not recognise puns, double-entendres etc. N-grams, stronger techniques, eg. Bert may be useful

'NASA sent a probe to all of the planets in our solar system, but quit after Uranus'

VS

'Why are the rings of Uranus turned sideways?'

Actual

Predict

- Need to periodically re-train model ("keep up with in-trend words"), for accurate classification

Q & A

Thank you