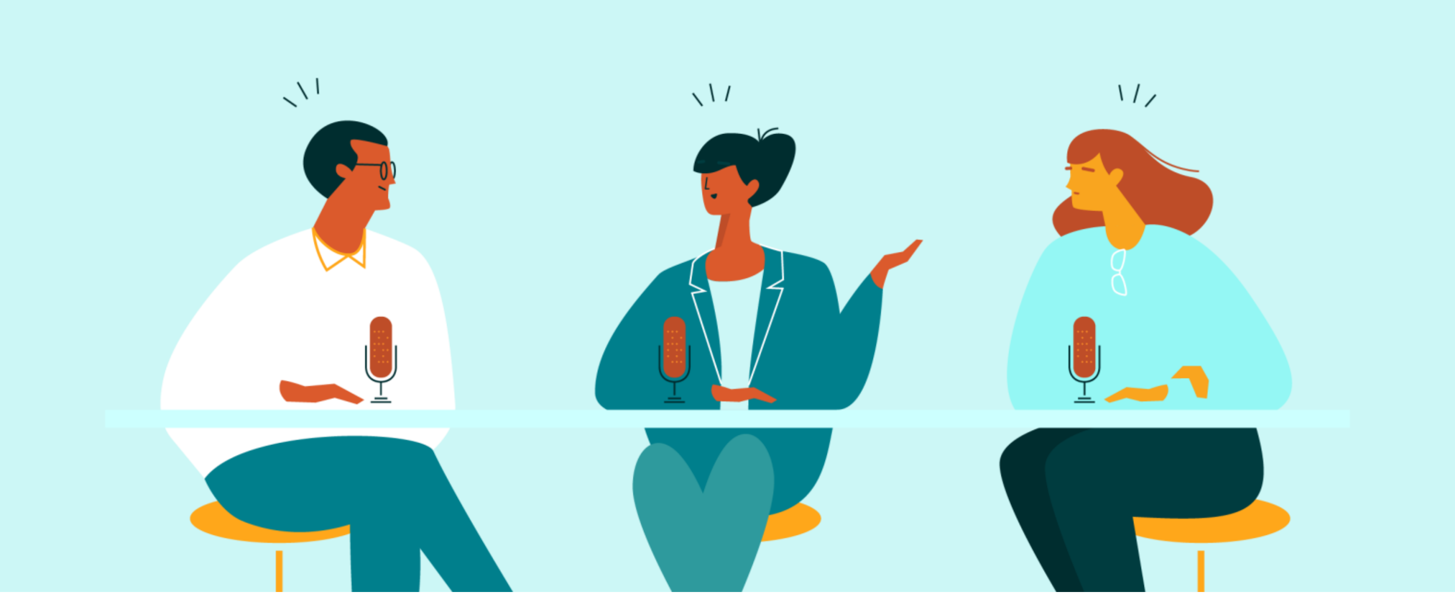
Subreddit Classification

Deepika Verma
Oct 7, 2022

Problem Statement

Panel Discussion-Science and Engineering

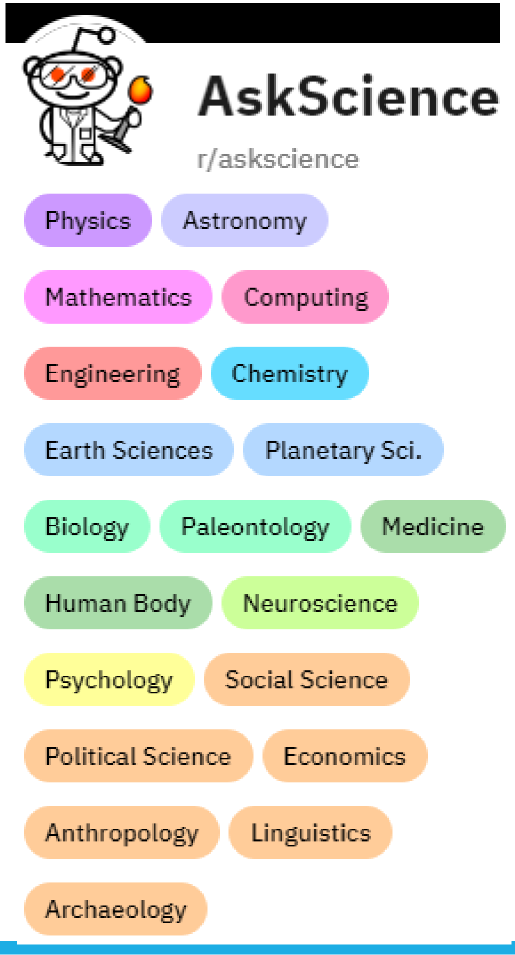


To train a classifier for Science and Engineering categories with accuracy >85% and f1-score>0.85

https://www.eventbrite.com.au/blog/organise-successful-panel-discussion-ds00/

Procedure/Methodology

- Data collected from two subreddits:
  - AskScience
  - AskEngineers
- 5074 submissions from each subreddit
- Data cleaning and EDA
- Preprocessing and modeling
- Model evaluation → production model



AskScience

r/askscience

Physics Astronomy

Mathematics Computing

Engineering Chemistry

Earth Sciences Planetary Sci.

Biology Paleontology Medicine

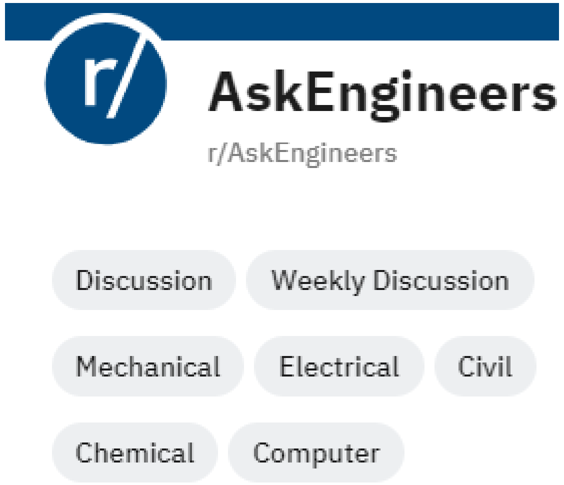
Human Body Neuroscience

Psychology Social Science

Political Science Economics

Anthropology Linguistics

Archaeology



AskEngineers

r/AskEngineers

Discussion Weekly Discussion

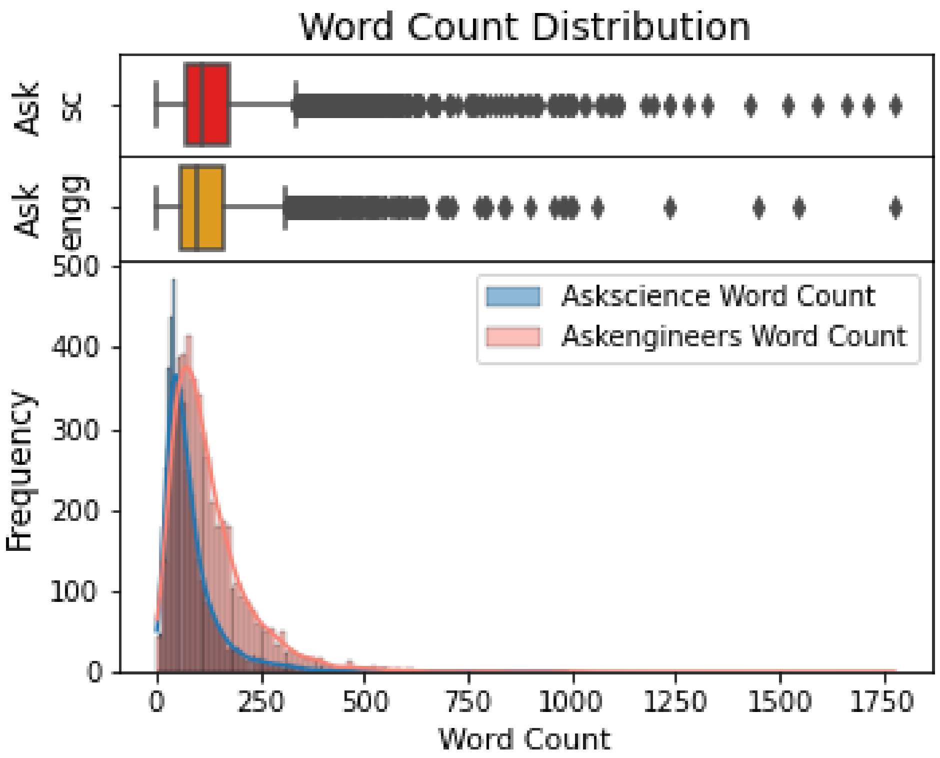
Mechanical Electrical Civil

Chemical Computer

Data Cleaning and EDA

Initial EDA –understanding the text and submission activity

- **Word Count Distributions**



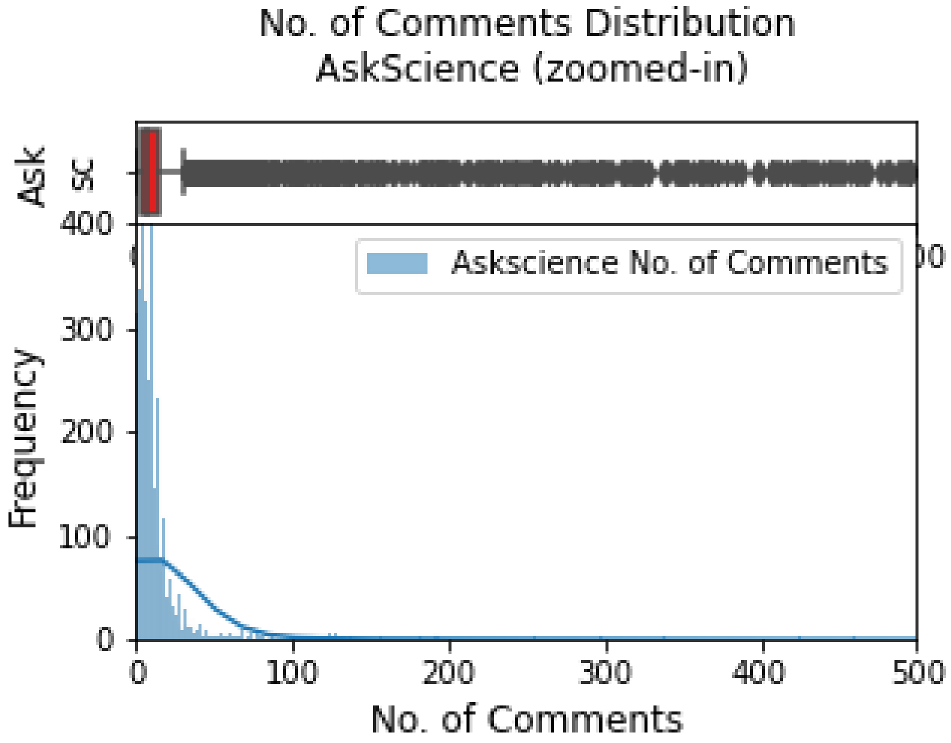
- **No. of authors in each category**

AskScience: 4433
AskEngineers: 4009

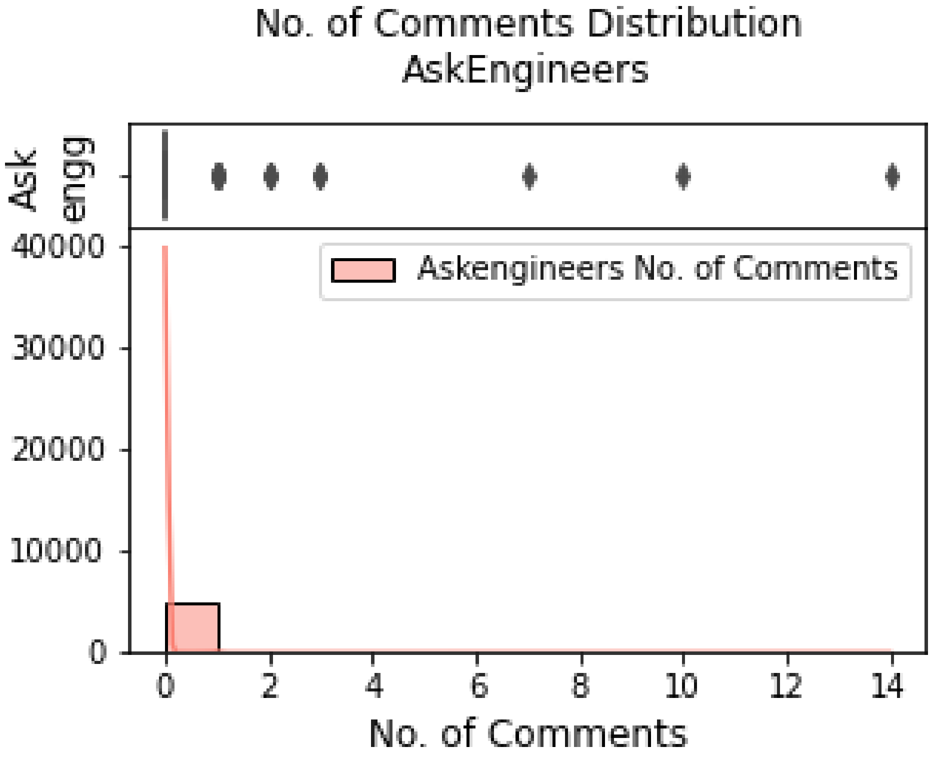
Data Cleaning and EDA

Initial EDA –understanding the text and submission activity

- **No. of Comments**



No. of Comments Distribution
AskScience (zoomed-in)



No. of Comments Distribution
AskEngineers

Data Cleaning and EDA

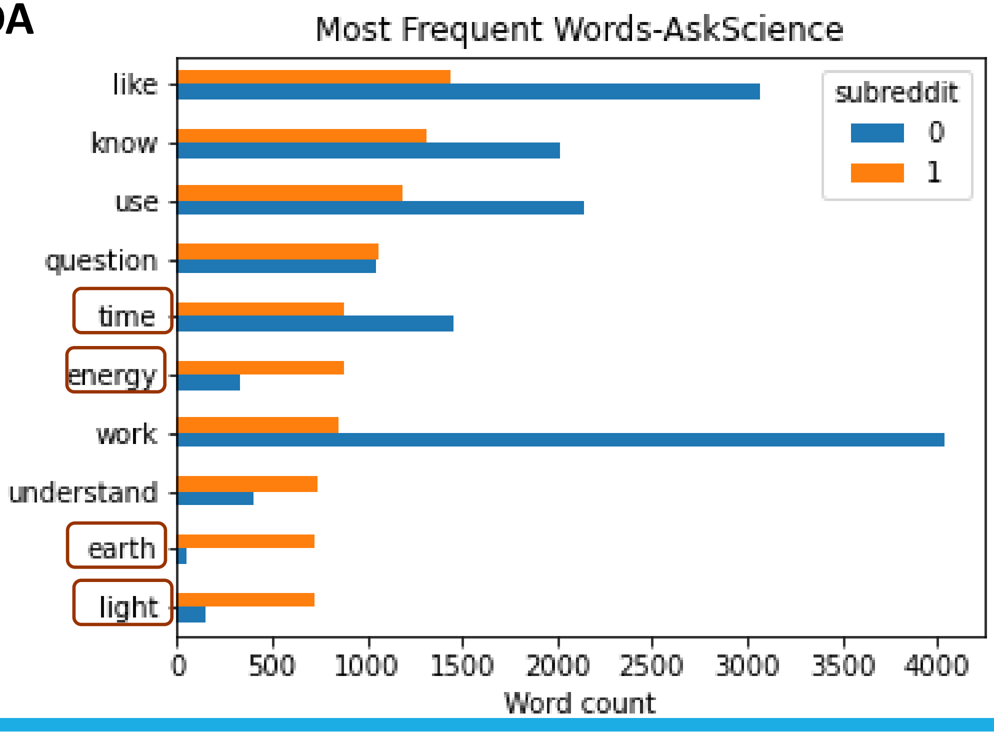
Cleaning

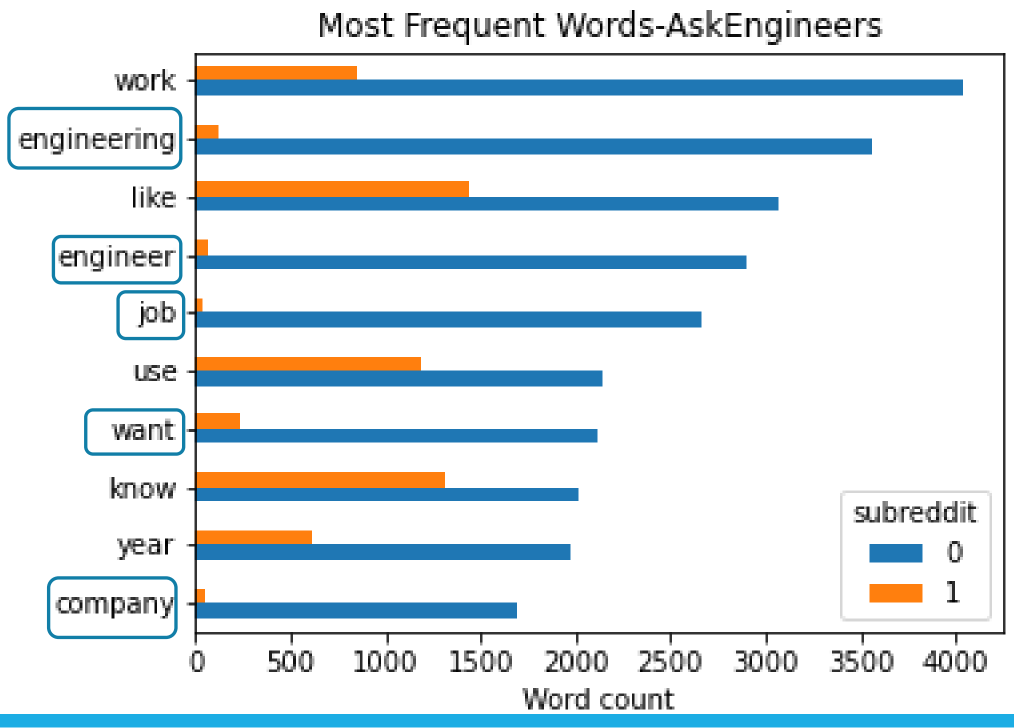
- Standard cleaning
- Patterns: Positive, negative no., float no., equation, dimension, percentage

Preprocessing

- Lemmatization
- Removal of stop words

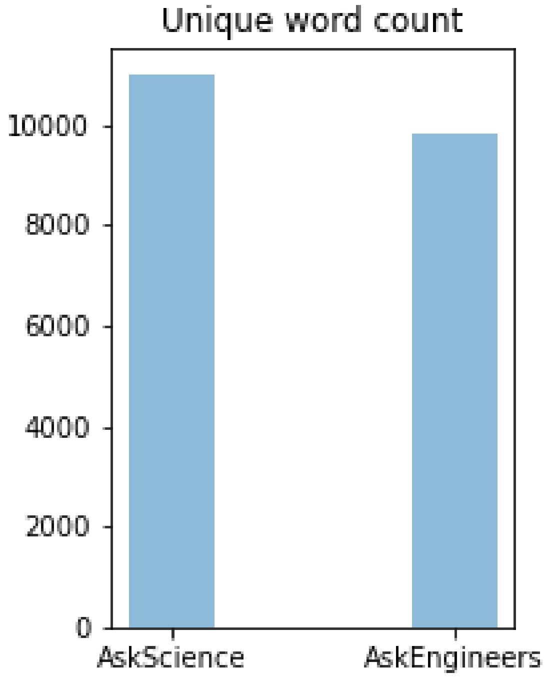
EDA



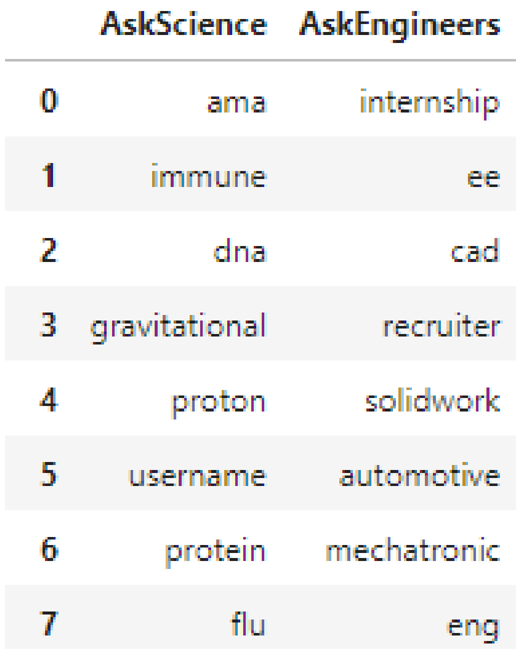


Data Cleaning and EDA

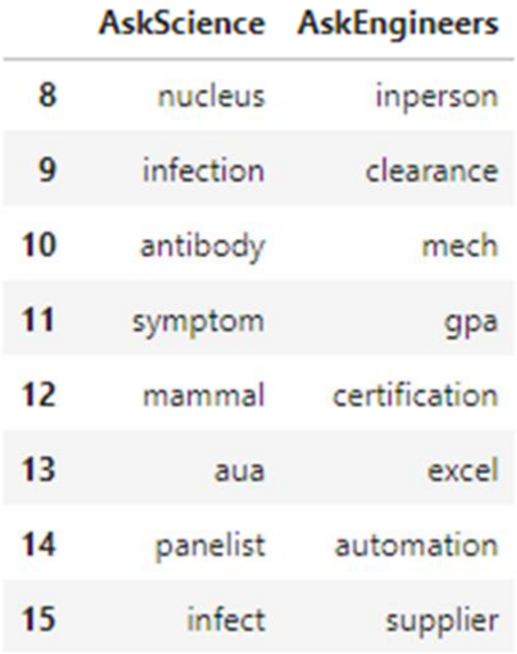
Unique Words



AskScience: **10,998**
AskEngineers: **9842**

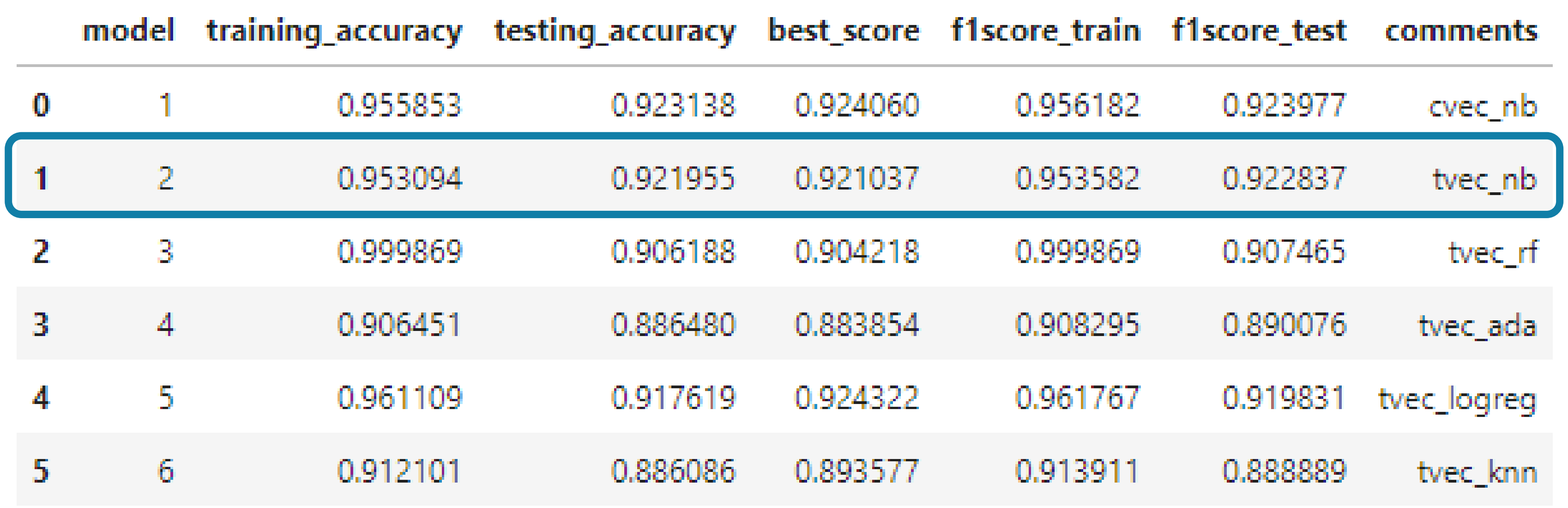


	AskScience	AskEngineers
0	ama	internship
1	immune	ee
2	dna	cad
3	gravitational	recruiter
4	proton	solidwork
5	username	automotive
6	protein	mechatronic
7	flu	eng



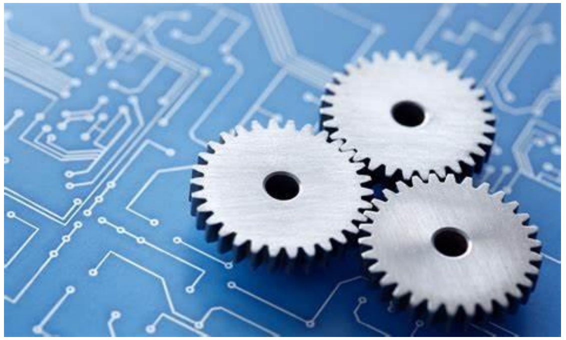
	AskScience	AskEngineers
8	nucleus	inperson
9	infection	clearance
10	antibody	mech
11	symptom	gpa
12	mammal	certification
13	aua	excel
14	panelist	automation
15	infect	supplier

Training different Classifiers



	model	training_accuracy	testing_accuracy	best_score	f1score_train	f1score_test	comments
0	1	0.955853	0.923138	0.924060	0.956182	0.923977	cvec_nb
1	2	0.953094	0.921955	0.921037	0.953582	0.922837	tvec_nb
2	3	0.999869	0.906188	0.904218	0.999869	0.907465	tvec_rf
3	4	0.906451	0.886480	0.883854	0.908295	0.890076	tvec_ada
4	5	0.961109	0.917619	0.924322	0.961767	0.919831	tvec_logreg
5	6	0.912101	0.886086	0.893577	0.913911	0.888889	tvec_knn

- **Naïve Bayes**
- **Random Forest**
- **Ada Boost**
- **Logistic Regression**
- **K-nearest Neighbors**



https://images.search.yahoo.com/search/images;_ylt=AwrjaRtK8z9jqNUKRqtXNyoA;_ylu=Y29sbwNn
cTEEcG9zAzEEdnRpZAMEc2VjA3BpdnM-?p=science&fr2=piv-
web&type=E211US105G0&fr=mcafee#id=20&iurl=http%3A%2F%2Fentertostudy.com%2Fwp-
content%2Fuploads%2F2020%2F03%2FGeneral-Science.png&action=click
https://getwallpapers.com/collection/science-backgrounds

Production Model

Naïve Bayes

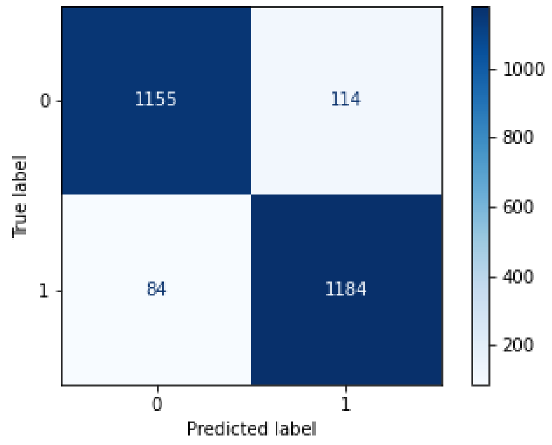
- Predicts 91% of negatives (AskEngineers) correctly
- Predicts 93% of positives (AskScience) correctly

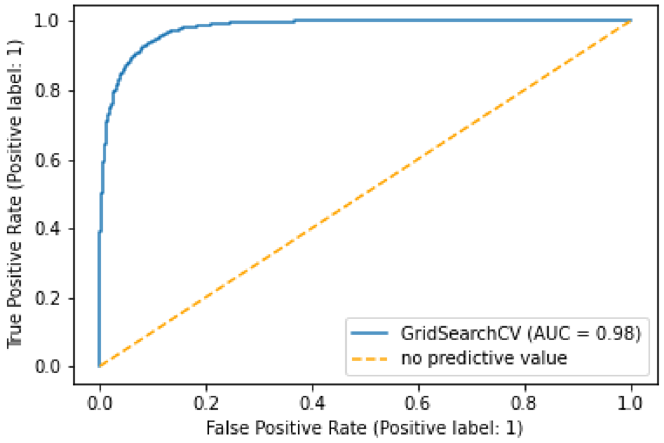
- Misclassification rate: 7.8%

(198/2535)

- Balanced accuracy: 0.92

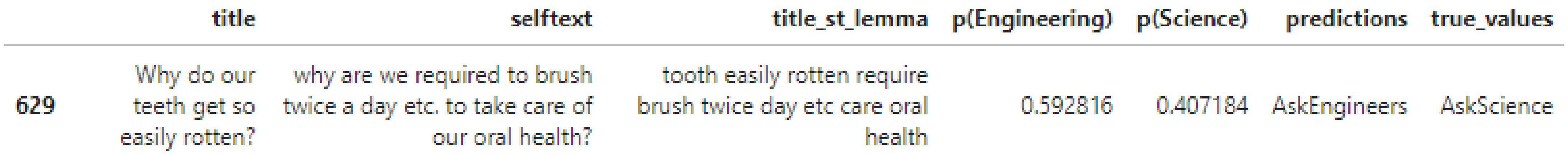
(ability to classify both the classes correctly)



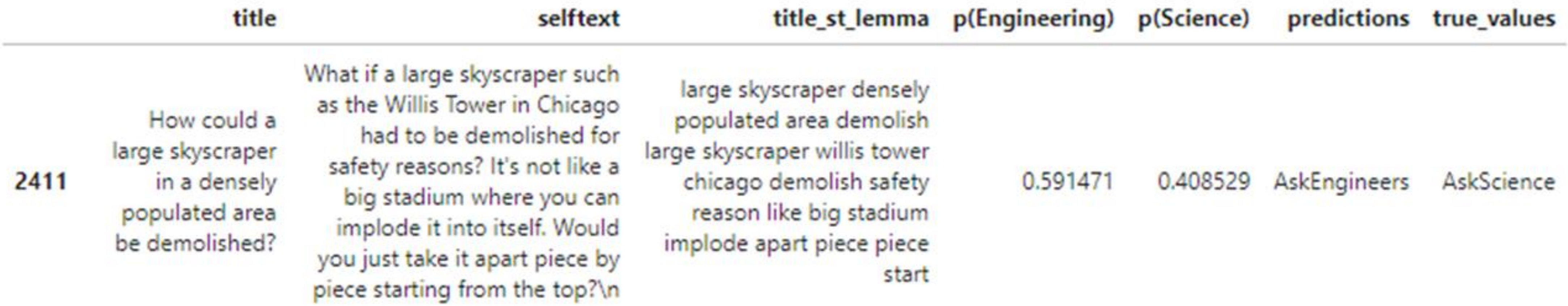


Misclassified posts EDA

On the fence



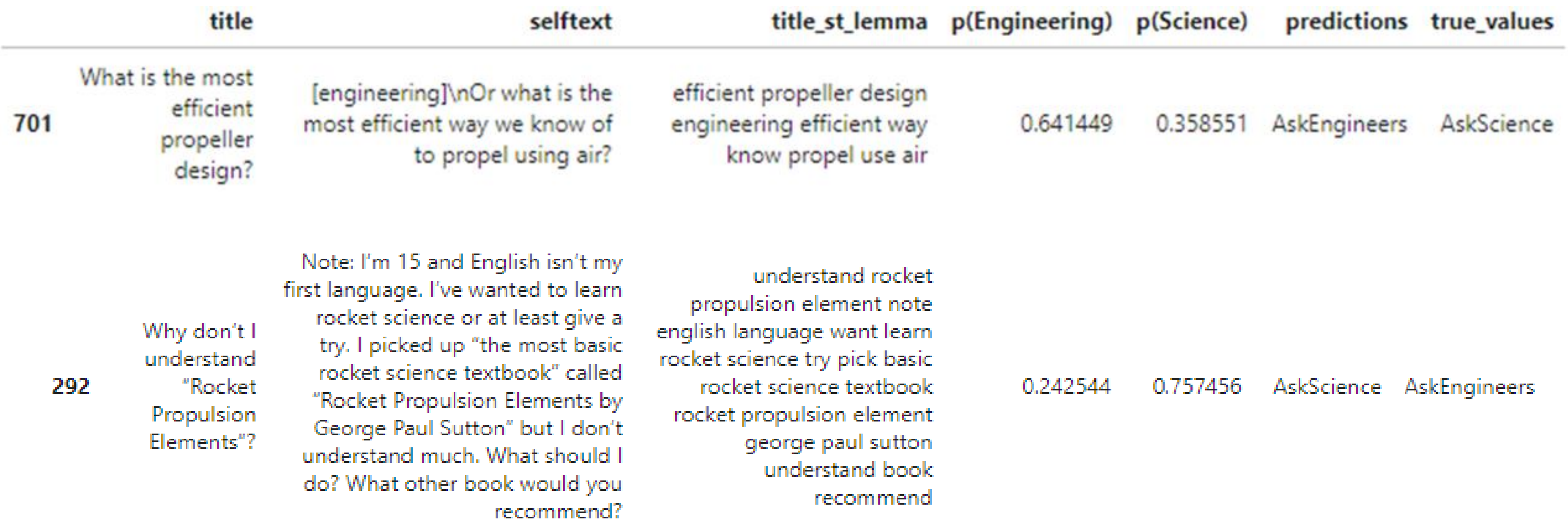
	title	selftext	title_st_lemma	p(Engineering)	p(Science)	predictions	true_values
629	Why do our teeth get so easily rotten?	why are we required to brush twice a day etc. to take care of our oral health?	tooth easily rotten require brush twice day etc care oral health	0.592816	0.407184	AskEngineers	AskScience



	title	selftext	title_st_lemma	p(Engineering)	p(Science)	predictions	true_values
2411	How could a large skyscraper in a densely populated area be demolished?	What if a large skyscraper such as the Willis Tower in Chicago had to be demolished for safety reasons? It's not like a big stadium where you can implode it into itself. Would you just take it apart piece by piece starting from the top?\n	large skyscraper densely populated area demolish large skyscraper willis tower chicago demolish safety reason like big stadium implode apart piece piece start	0.591471	0.408529	AskEngineers	AskScience

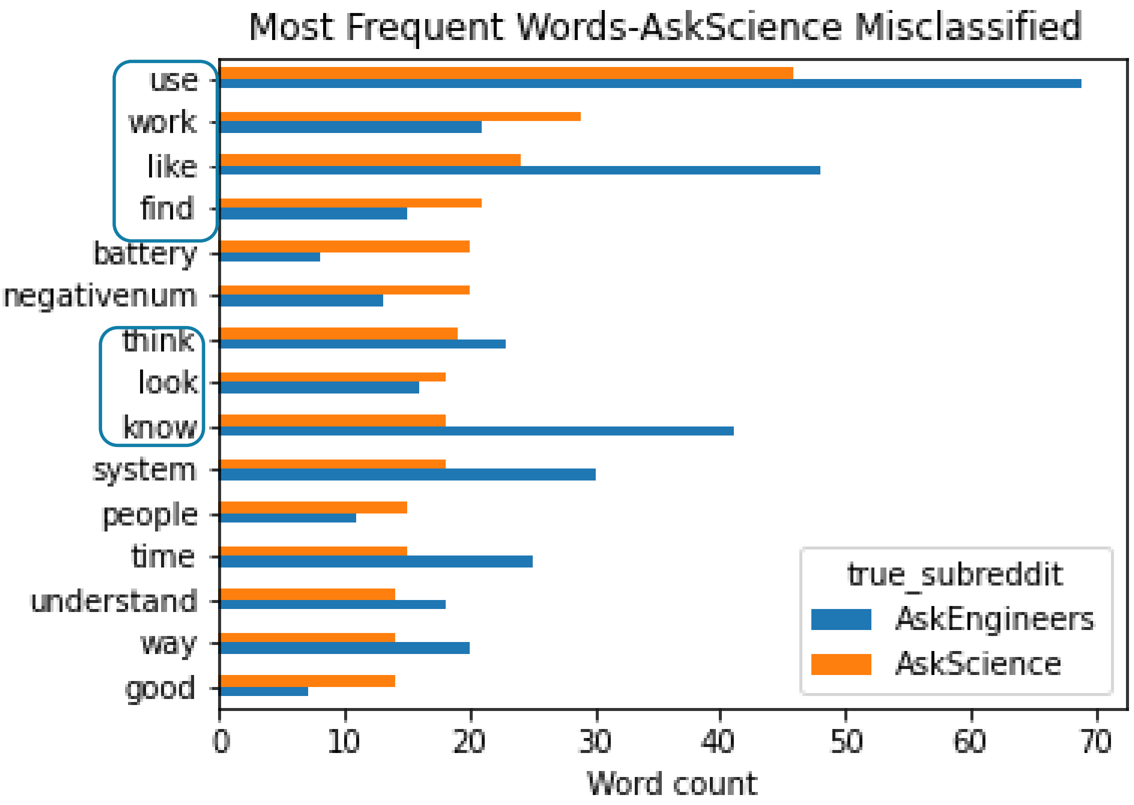
Misclassified posts EDA

Correctly misclassified??

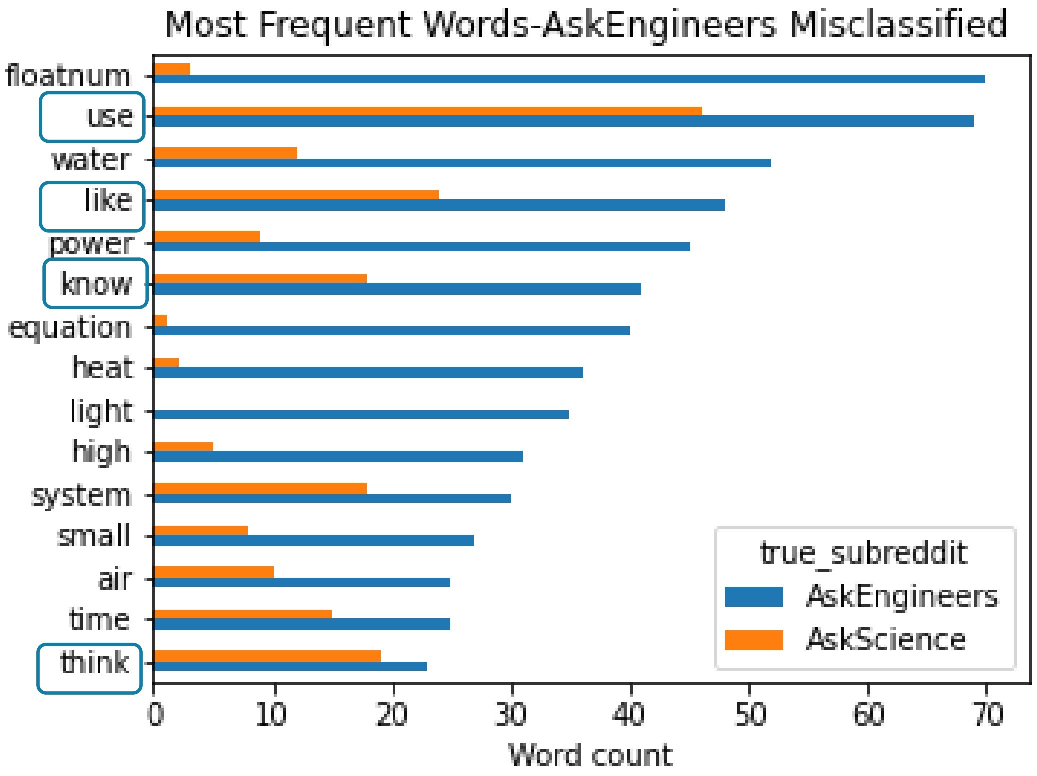


	title	selftext	title_st_lemma	p(Engineering)	p(Science)	predictions	true_values
701	What is the most efficient propeller design?	[engineering]\nOr what is the most efficient way we know of to propel using air?	efficient propeller design engineering efficient way know propel use air	0.641449	0.358551	AskEngineers	AskScience
292	Why don't I understand "Rocket Propulsion Elements"?	Note: I'm 15 and English isn't my first language. I've wanted to learn rocket science or at least give a try. I picked up "the most basic rocket science textbook" called "Rocket Propulsion Elements by George Paul Sutton" but I don't understand much. What should I do? What other book would you recommend?	understand rocket propulsion element note english language want learn rocket science try pick basic rocket science textbook rocket propulsion element george paul sutton understand book recommend	0.242544	0.757456	AskScience	AskEngineers

Misclassified posts EDA



Most Frequent Words-AskScience Misclassified



Most Frequent Words-AskEngineers Misclassified

Summary and Recommendations

- Naïve-Bayes predictive model (model 2) satisfies the success criteria of:
  - accuracy >85% and f1-score>85% in classifying the Science and Engineering submissions, and is the
  - best performing model

- This model is recommended to the University to use during their panel discussion to classify the questions in two categories-Science and Engineering.

Next Steps

- Tuning of model to improve f1-score, and reduce false positives and false negatives

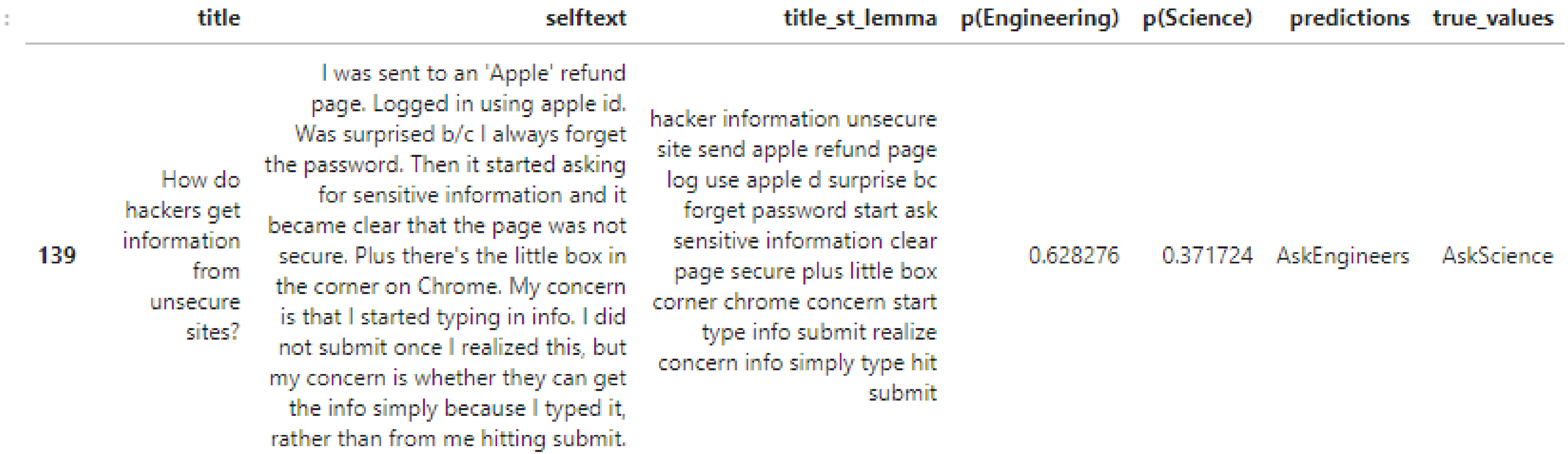
- Explore text of misclassified posts and see if there are any patterns, do some more data cleaning and tune the model

Thank You!!

Acknowledgment:

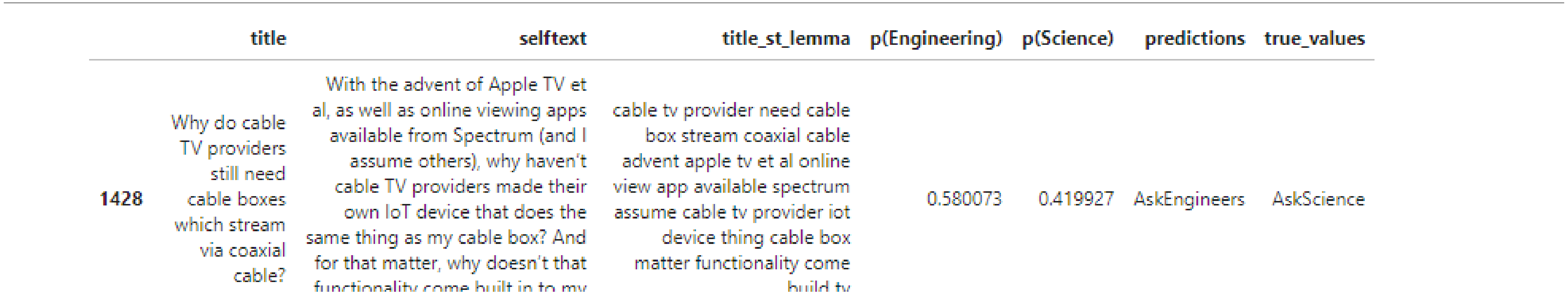
- Hank Butler
- Alanna
- Devin
- All the classmates!

Misclassified posts EDA

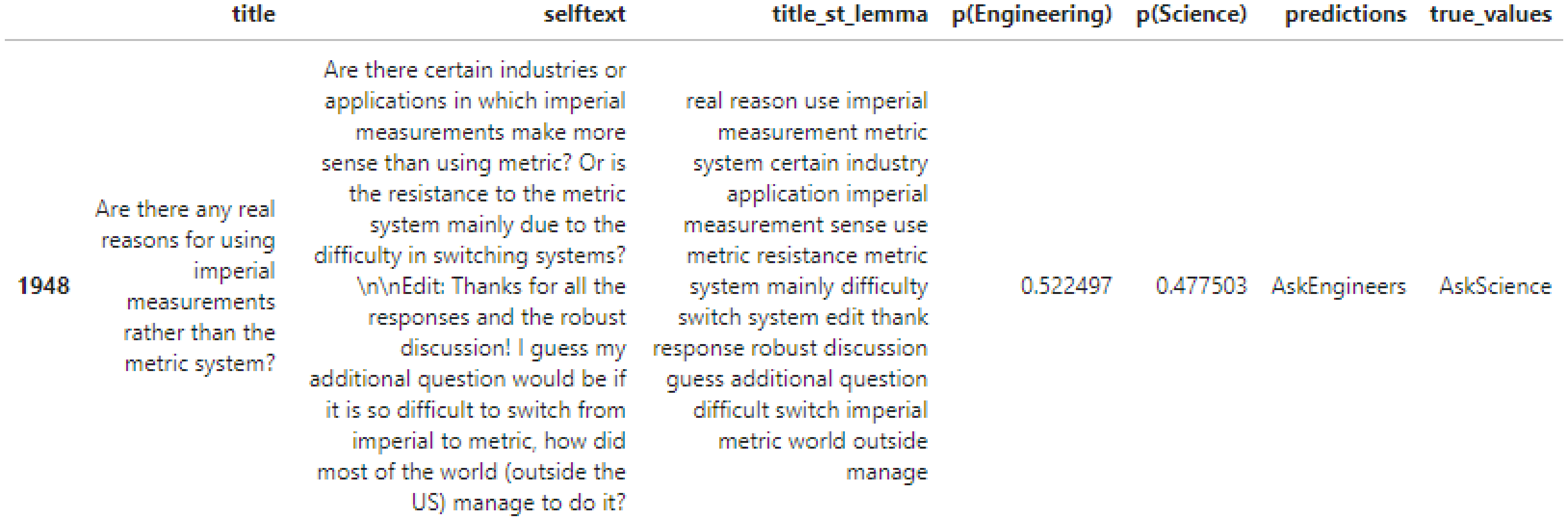


	title	selftext	title_st_lemma	p(Engineering)	p(Science)	predictions	true_values
139	How do hackers get information from unsecure sites?	I was sent to an 'Apple' refund page. Logged in using apple id. Was surprised b/c I always forget the password. Then it started asking for sensitive information and it became clear that the page was not secure. Plus there's the little box in the corner on Chrome. My concern is that I started typing in info. I did not submit once I realized this, but my concern is whether they can get the info simply because I typed it, rather than from me hitting submit.	hacker information unsecure site send apple refund page log use apple d surprise bc forget password start ask sensitive information clear page secure plus little box corner chrome concern start type info submit realize concern info simply type hit submit	0.628276	0.371724	AskEngineers	AskScience

Misclassified posts EDA



	title	selftext	title_st_lemma	p(Engineering)	p(Science)	predictions	true_values
1428	Why do cable TV providers still need cable boxes which stream via coaxial cable?	With the advent of Apple TV et al, as well as online viewing apps available from Spectrum (and I assume others), why haven't cable TV providers made their own IoT device that does the same thing as my cable box? And for that matter, why doesn't that functionality come built in to my	cable tv provider need cable box stream coaxial cable advent apple tv et al online view app available spectrum assume cable tv provider iot device thing cable box matter functionality come build tv	0.580073	0.419927	AskEngineers	AskScience



	title	selftext	title_st_lemma	p(Engineering)	p(Science)	predictions	true_values
1948	Are there any real reasons for using imperial measurements rather than the metric system?	Are there certain industries or applications in which imperial measurements make more sense than using metric? Or is the resistance to the metric system mainly due to the difficulty in switching systems? \n\nEdit: Thanks for all the responses and the robust discussion! I guess my additional question would be if it is so difficult to switch from imperial to metric, how did most of the world (outside the US) manage to do it?	real reason use imperial measurement metric system certain industry application imperial measurement sense use metric resistance metric system mainly difficulty switch system edit thank response robust discussion guess additional question difficult switch imperial metric world outside manage	0.522497	0.477503	AskEngineers	AskScience