Subreddit Classification with Natural Language Processing

By: Micah Luedtke

Problem:

- **Quantitative:** Can word frequency and usage be used to predict where a reddit post came from the r/Linguistics or the r/LanguageLearning subreddits?
- **Qualitative:** What insights can we gain from the difference in keywords between two different groups studying language in different ways?

Two types of linguists:

- **Someone who speaks another language:** r/LanguageLearning is a subreddit where people ask questions and share advice about effectively learning other languages
- **Someone who studies linguistics:** r/Linguistics is a subreddit where people discuss various topics related to the structure of different languages

Web Scraping and Data Cleaning:

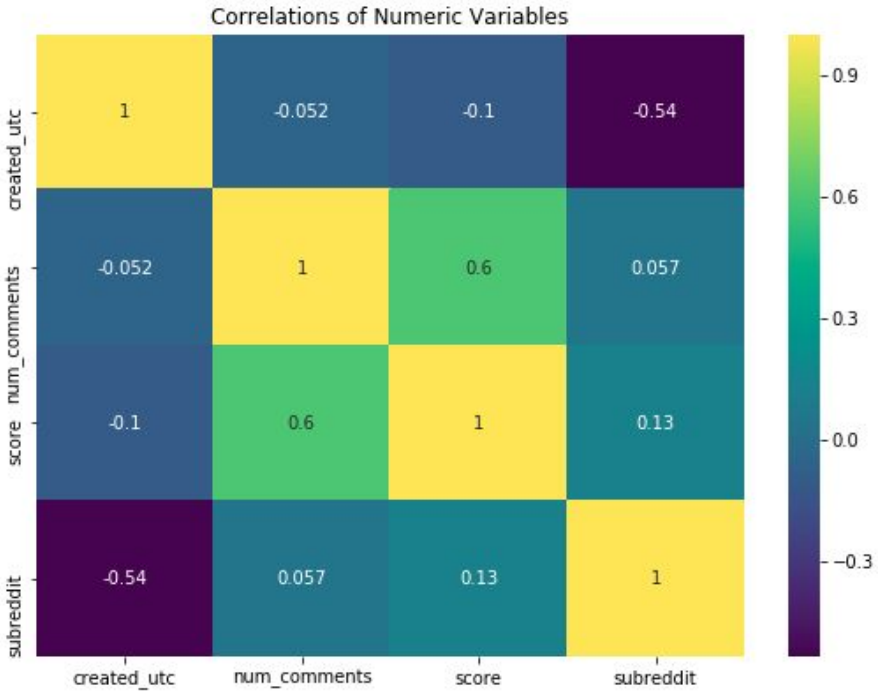
In [19]:
```
#Function for pulling
def pull_merge(subreddit_1, subreddit_2, posts=100):
    '''Default for posts is 100, maximum request is 1000 posts'''
    headers = {'user-micah':'my-user-micah'}
    url_1 = "https://api.pushshift.io/reddit/search/submission/?subreddit="+subreddit_1+"&size="+str(posts)
    url_2 = "https://api.pushshift.io/reddit/search/submission/?subreddit="+subreddit_2+"&size="+str(posts)
    res_1 = requests.get(url_1,headers=headers)
    res_2 = requests.get(url_2,headers=headers)
    df_subreddit_1 = pd.DataFrame(res_1.json()['data'])
    df_subreddit_2 = pd.DataFrame(res_2.json()['data'])
    df = df_subreddit_2.append(df_subreddit_1, sort=False)
    df.to_csv('{} and {}_'.format(subreddit_1,subreddit_2)+str(datetime.datetime.now()))
```

In [4]:
```
key_columns= ['author','created_utc','num_comments','permalink',
              'score','selftext','subreddit','title'
             ]
```

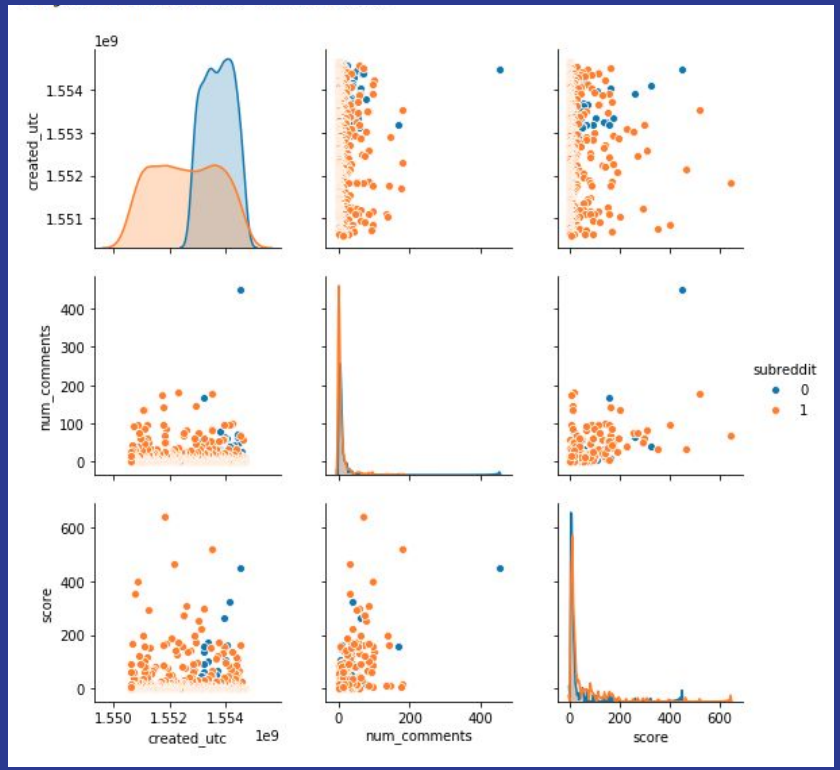
In [5]:
```
df.drop(df.columns.difference(key_columns), 1, inplace=True)
```

EDA

Heatmap of Numeric Variables



Pairplot of Numeric Variables



Vectorizers and Classification Models

Performance on Training and Testing Data

Models	Countvectorizer	TF-IDF
Logistic Regression	Training R^2 : 0.99	Training R^2 : 0.94
	Testing R^2 : 0.82	Testing R^2 : 0.86
Naive Bayes	Training R^2 : 0.87	Training R^2 : 0.97
	Testing R^2 : 0.85	Testing R^2 : 0.87

Keyword Comparison

CountVectorizer

	word	coef_log	coef_nb
405	improve	-0.754883	-9.213959
458	learned	-0.757520	-7.892203
360	happy	-0.757603	-8.202358
239	each	-0.774650	-7.166266
185	countries	-0.779477	-8.520812
2	10	-0.785569	-8.520812
36	already	-0.789205	-8.035304
241	easier	-0.789978	-8.520812
457	learn	-0.825139	-6.688230
597	order	-0.847582	-7.268049
866	tips	-0.872227	-9.501641
650	practice	-0.883569	-9.213959
52	another	-0.915332	-7.304416
470	level	-0.945847	-7.103746
576	now	-0.955422	-6.649009
448	la	-0.977578	-8.808494
105	best	-1.006394	-7.827664
236	duolingo	-1.317092	-9.907106
461	learning	-1.669505	-6.750106
695	removed	-1.996544	-8.654343

	word	coef_log	coef_nb
477	linguistics	2.058772	-5.709904
423	ipa	0.947227	-7.464759
189	create	0.839364	-7.767040
150	come	0.822442	-6.936692
130	change	0.814166	-7.166266
323	general	0.794598	-6.771612
274	explain	0.791427	-7.655814
706	say	0.788320	-5.936814
733	short	0.788292	-7.199056
476	linguistic	0.787974	-7.304416
662	pronoun	0.781667	-8.202358
266	example	0.765217	-6.218227
680	reading	0.760800	-6.522716
169	considered	0.754080	-7.381377
843	their	0.751190	-6.205804
419	interesting	0.747936	-7.073893
813	syntax	0.730545	-7.509211
536	morphology	0.719680	-7.827664
691	related	0.718507	-6.911374
700	rules	0.716953	-7.103746

Nazioarteko alfabeto fonetikoa (*The International Phonetic Alphabet*) © 2005 IPA

Kontsonanteak (Aire-etorriaz ekoiztutakoak).

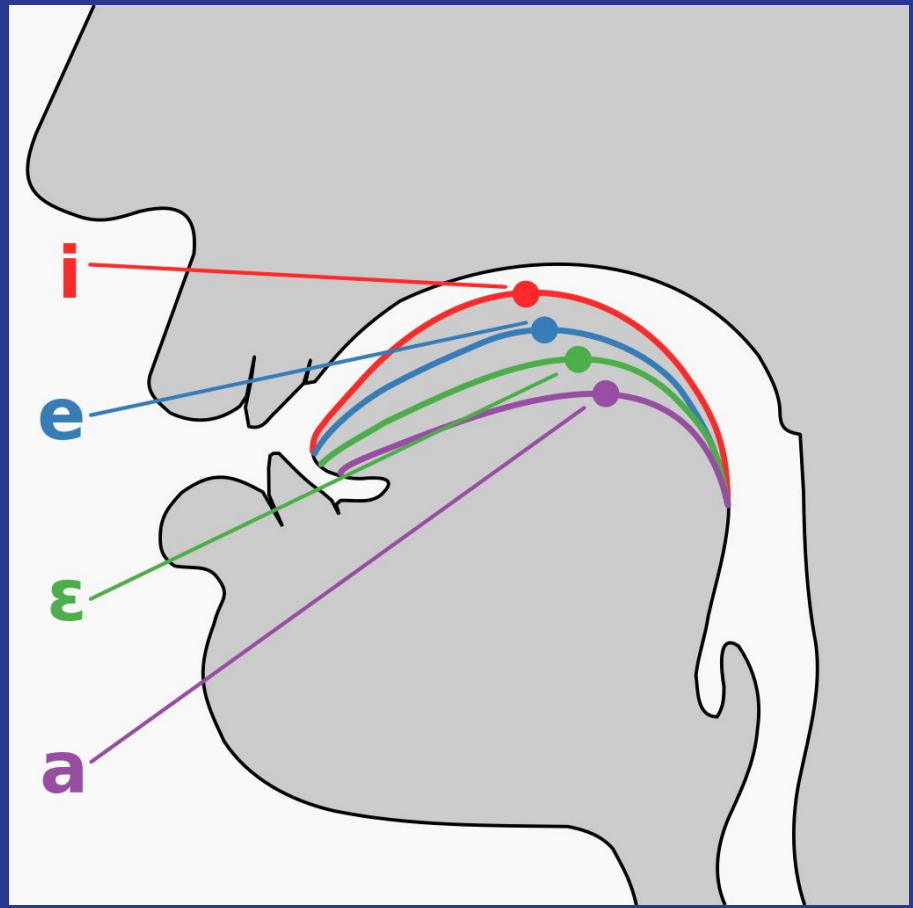
Ahoskunea → / Ahosmoldea ↓	Ezpainkariak	Ezpain-horzkariak	Horzkariak	Hobikariak	Sabaiaurrekoak	Irauliak (apikariak)	Sabaikariak	Belareak	Ubulareak	Faringealak	Glotalak
Leherkariak	p b		t̪ d̪	t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Sudurkariak	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Dardarkariak	ʙ			r					ʀ		
Ttak *(tap/flap)*		ⱱ		ɾ		ɽ					
Igurzkariak	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Albokari igurzkariak				ɬ ɮ							
Hurbilkariak		ʋ		ɹ		ɻ	j	ɰ			
Albokari hurbilkariak				l		ɭ	ʎ	ʟ			

Sinboloak, kutxa batean bi agertzen direnean: ezkerrekoa ahoskabea, eskuinekoa ahostuna. Kutxa grisetakoak ahoskagaitz.

VOWELS

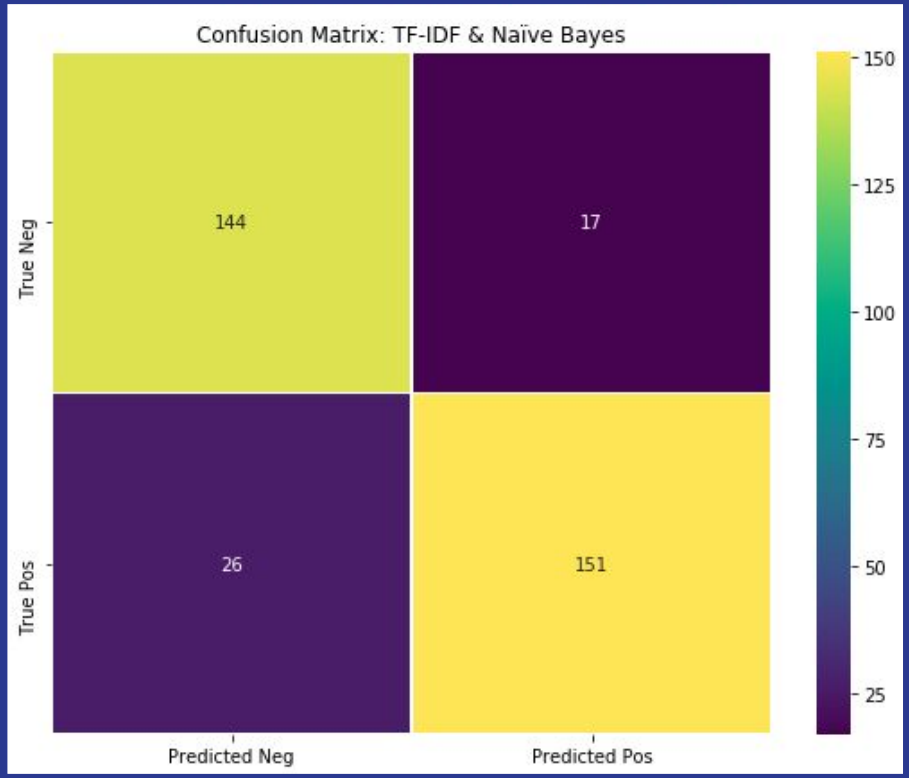
	Front	Central	Back
Close	i • y	ɨ • ʉ	ɯ • u
	ɪ ʏ		ʊ
Close-mid	e • ø	ɘ • ɵ	ɤ • o
		ə	
Open-mid	ɛ • œ	ɜ • ɞ	ʌ • ɔ
	æ	ɐ	
Open	a • ɶ		ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.



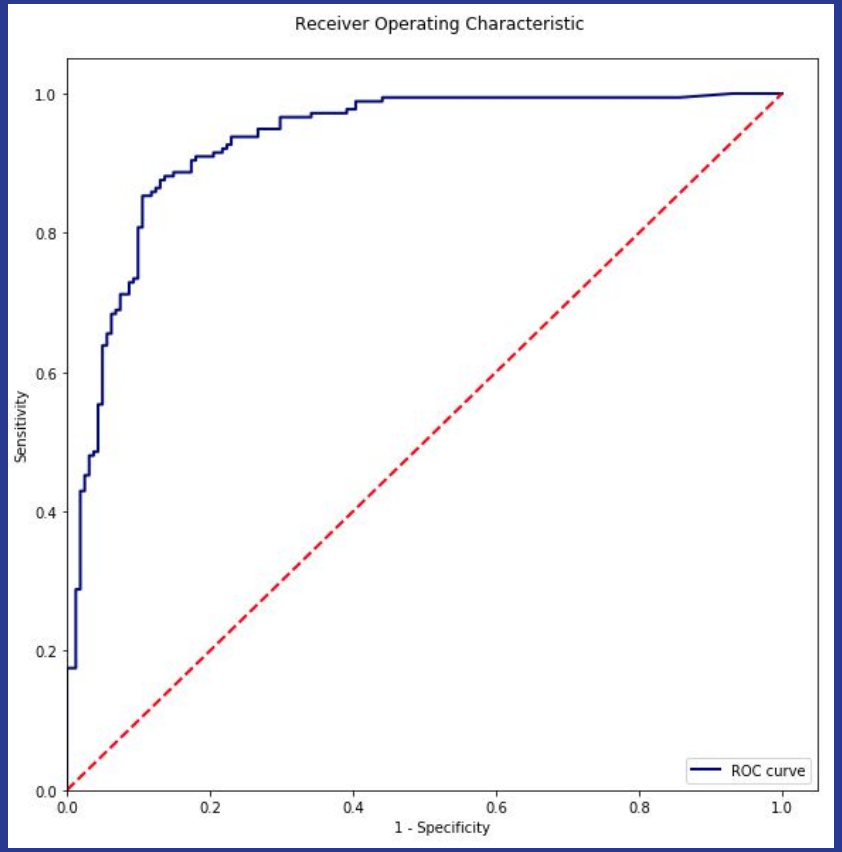
Confusion Matrix

TF-IDF and Naïve Bayes



ROC

TF-IDF and Naïve Bayes



BONUS: Constructed Languages

- Constructed languages are languages that were constructed by a single person or small group relatively quickly, rather than over time by natural languages development processes.
- r/Conlags is dedicated to people interested in creating constructed languages and discussing methods.
- Ran a sample of r/Conlags through each of the models and the posted were classified as r/Linguistics 65%-75% of the time.

Conclusions and Future Work:

- **People interested in linguistics and language learning approach similar subject matter from different perspectives**
- **Direct comparisons between r/LanguageLearning & r/Conlags and r/Linguistics & r/Conlags**
- **User-level comparison**