Capstone 2: Classifying Twitter Text

Predicting whether a student or teacher tweeted using NLP Machine Learning Techniques

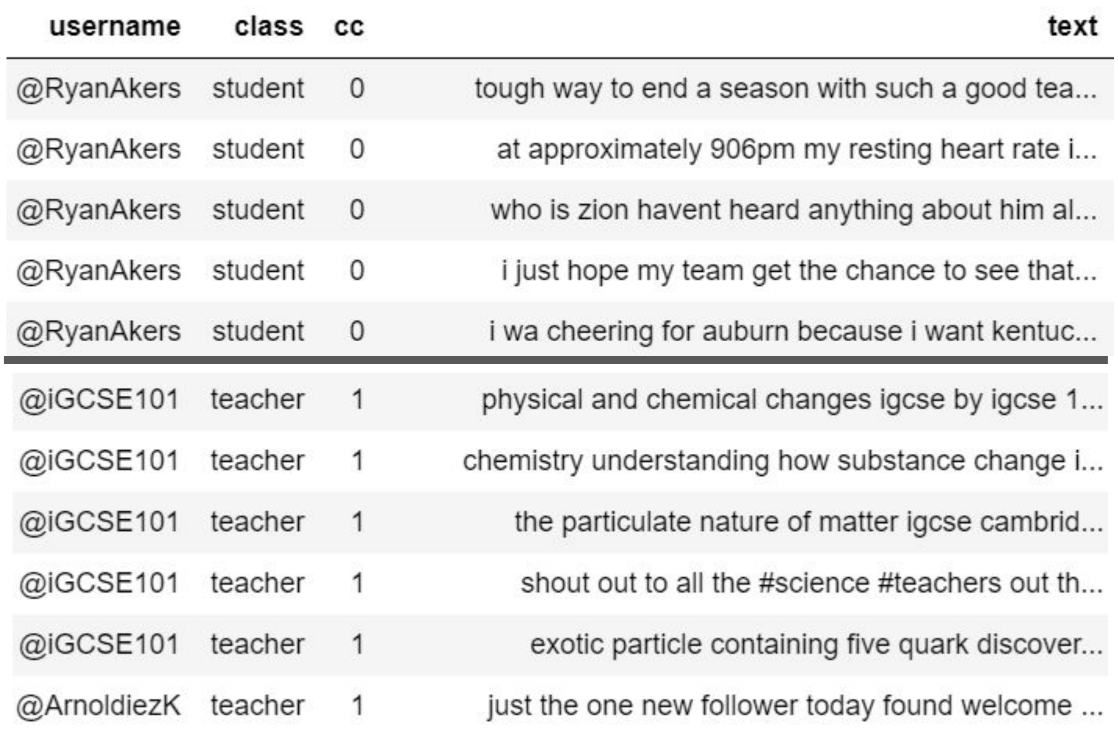
Christos Magganas - Springboard

NLP Classification Prediction

Objective: Twitter Classification – gather twitter data from two similar groups of users with a clear identifiable keyword in their profile, specifically “Students” and “Teachers.” Then create an Natural Language Processing prediction model that attempts to identify them based on tweet text.

DATA SCRAPING / WRANGLING / CLEANING

Here we see some examples of the tweets after they have been cleaned (tweet tokenization and lemmatization)



username	class	cc	text
@RyanAkers	student	0	tough way to end a season with such a good tea...
@RyanAkers	student	0	at approximately 906pm my resting heart rate i...
@RyanAkers	student	0	who is zion havent heard anything about him al...
@RyanAkers	student	0	i just hope my team get the chance to see that...
@RyanAkers	student	0	i wa cheering for auburn because i want kentuc...
@iGCSE101	teacher	1	physical and chemical changes igcse by igcse 1...
@iGCSE101	teacher	1	chemistry understanding how substance change i...
@iGCSE101	teacher	1	the particulate nature of matter igcse cambrid...
@iGCSE101	teacher	1	shout out to all the #science #teachers out th...
@iGCSE101	teacher	1	exotic particle containing five quark discover...
@ArnoldiezK	teacher	1	just the one new follower today found welcome ...

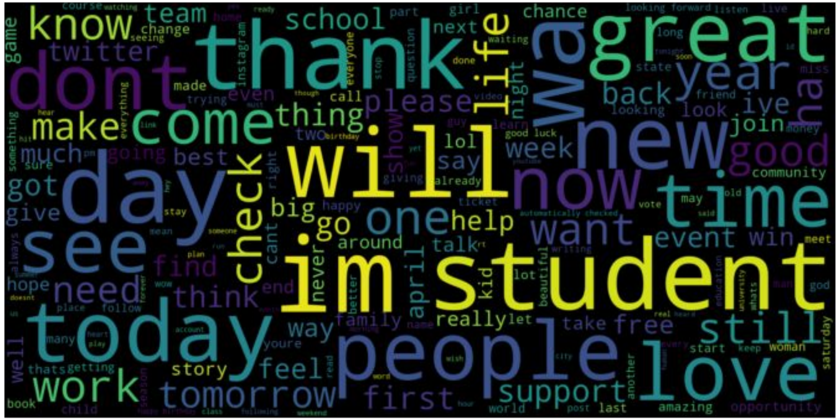
Feature Importance

Each word becomes a feature that helps the model predict class

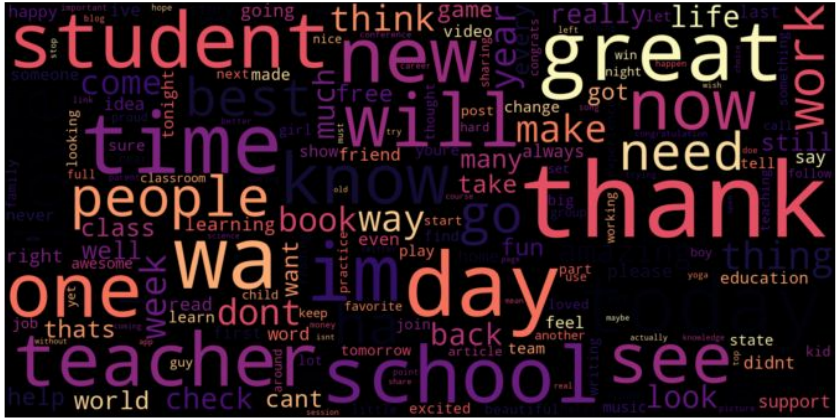
Predictive Term Frequency Table

	students	*teachers*	*total*
"school"	26	72	98
"best"	20	50	70
"teacher"	6	47	53
"book"	14	35	49
"class"	10	29	39

Student WordCloud



Teacher WordCloud

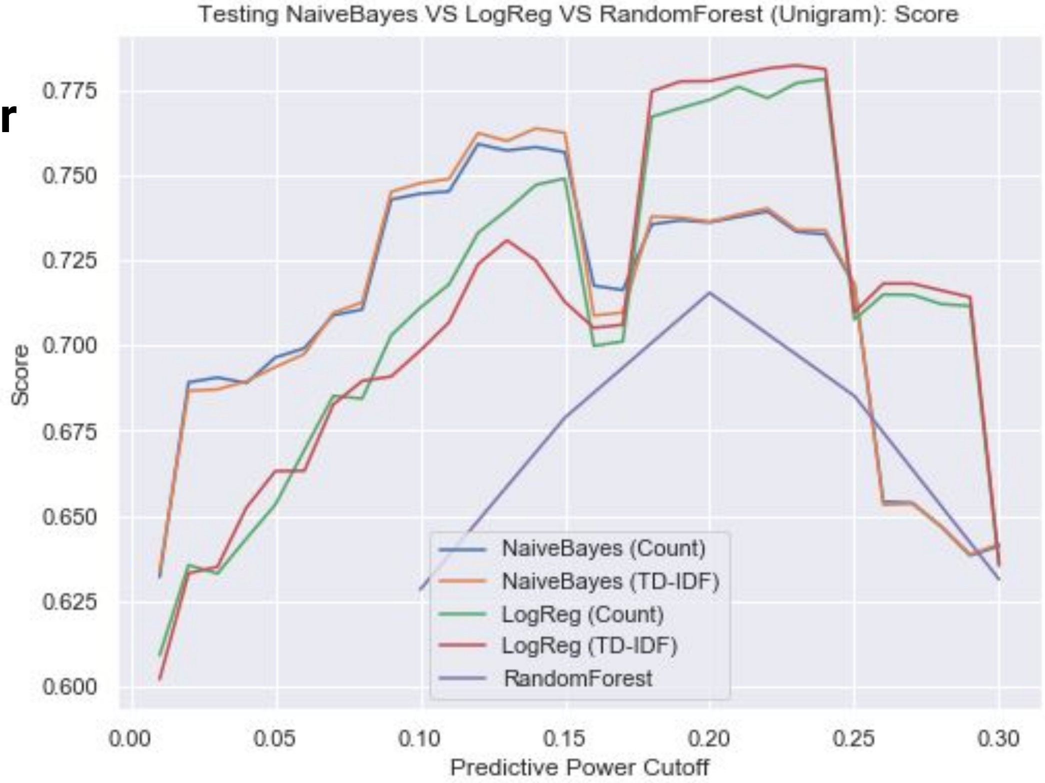


Vectorizers / Classifiers / Stopwords / Cross-Validation

Count**Vectorizer** & Tfidf**Vectorizer**

MultinomialNB &
LogisticRegression &
RandomForest**Classifier**

Stopwords: Non-predictive features are based on a changing range of feature predictiveness from lowest predictiveness to a given cutoff



Results

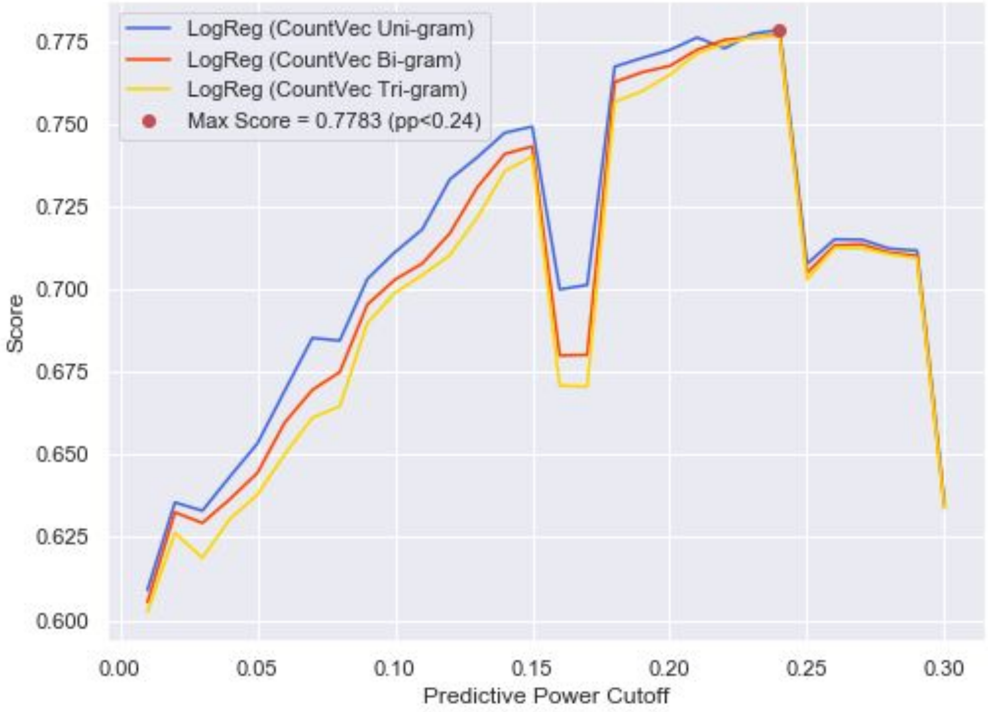
Score by count/tf-idf vectorizer and uni/bi/tri_gram

NaiveBayes: count uni 0.7384 bi 0.7333 tri 0.7262
tf-idf uni 0.739 bi 0.7356 tri 0.7274

LogReg: count uni 0.7613 bi **0.7692** tri 0.7674
tf-idf uni 0.7696 bi 0.7689 tri 0.7667
(**logreg-count**vec-**bi**gram has the highest score)

RandForest: count uni 0.7109 bi 0.7075 tri 0.7043
tf-idf uni 0.7151 bi 0.7128 tri 0.7079

(all scores above w/ custom_stop_words such that predictive_power < 0.2)



Reflection

Why does this problem matter? - The ability to identify individuals from the masses by a certain class, whether that is from Twitter or some other platform, is an extraordinarily tool to have for any data project.

How could this technology be used? - Being able to classify a group of people based on their tweets can be useful for advertising, but also for…

- Filtering out spam
- Identify abusive or obscene content
- Group similar frequently asked questions to streamline response
- Compare positive and negative user reviews for improvement
- Identifying individuals that might pose a security threat

Links:

https://twitter.com/,

https://followerwonk.com/,

https://github.com/Jefferson-Henrique/GetOldTweets-python/blob/master/LICENSE