

Use machine learning to predict
Twitter user's education level based on
their tweets



Sarbottam Piya, PhD

Sarbottam.piya@gmail.com

[Github.com/spiya](https://github.com/spiya)

Introduction



Education is the most important indicators of life outcomes



Education impacts an individual's perception about social and political changes



Can the opinion of an individual be used as a predictor to identify their education level?




Introduction



 Social media platforms such as Facebook and Twitter are a source of large amounts of publicly available user-generated data



 Understanding the demographics of social media users is beneficial for targeted public messaging campaigns

Develop a machine learning model that predicts the education level of Twitter users based on the opinions they expressed in Twitter

Goal of the project



Targeted audience

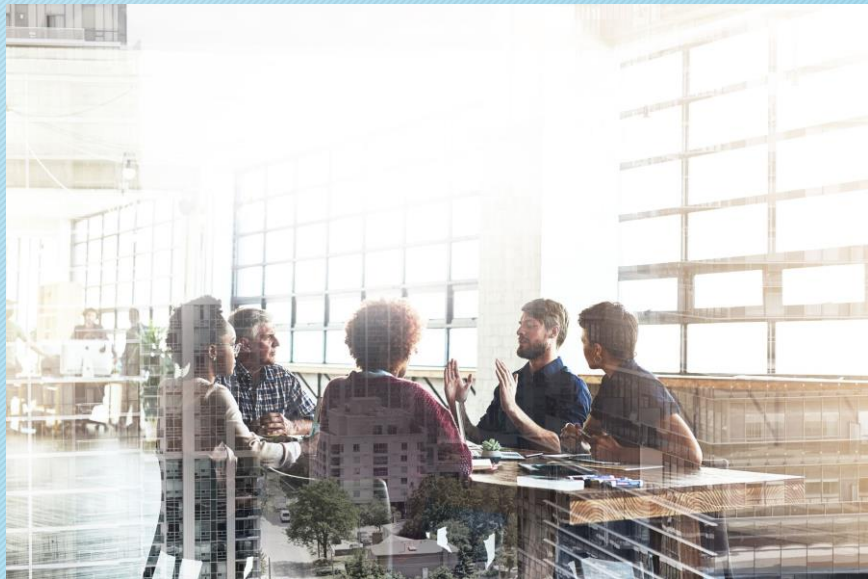


Business/companies

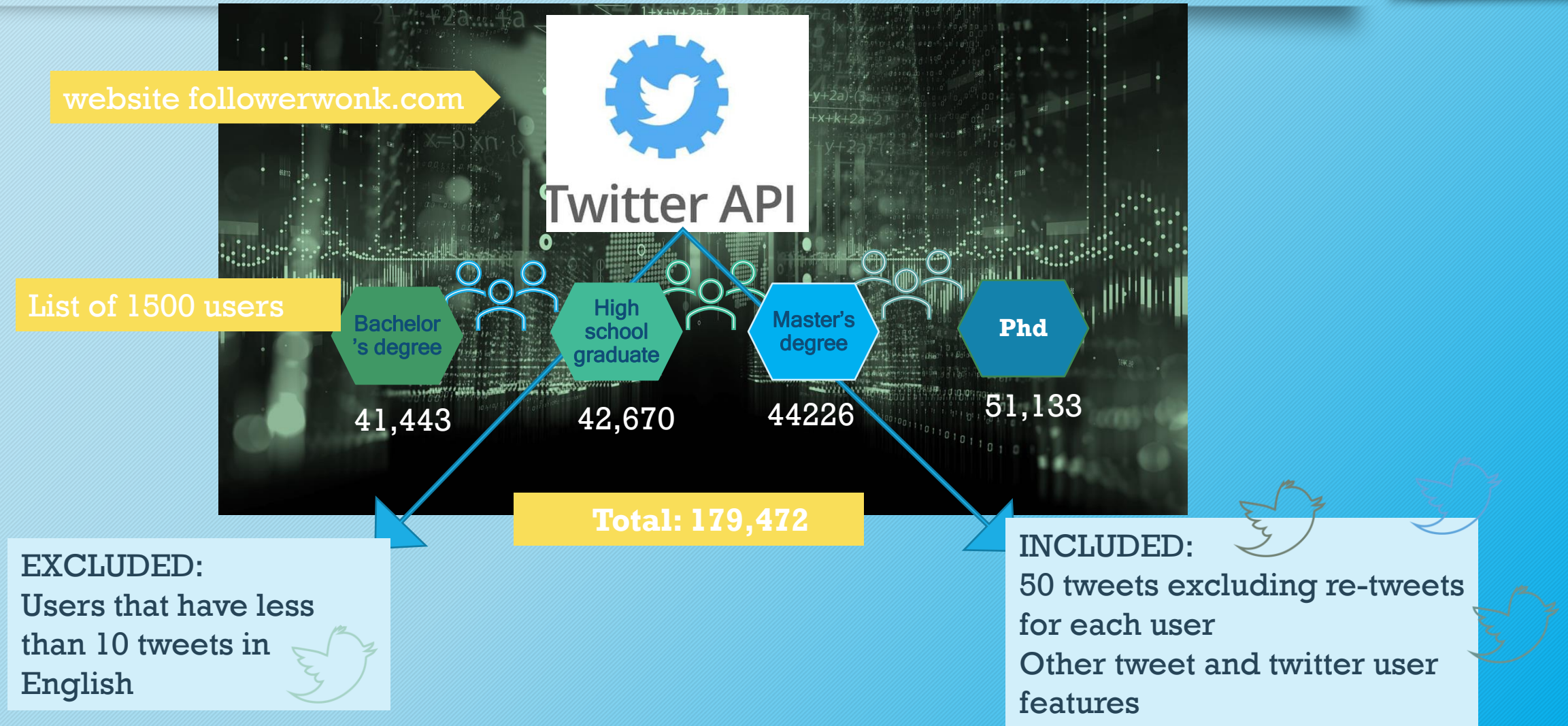
- Targeted advertisement

Policymakers

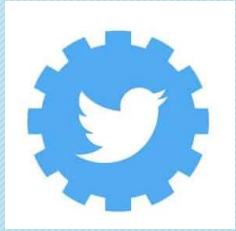
- Understand opinion of people



Dataset

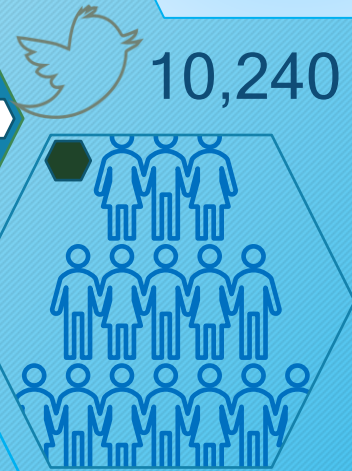
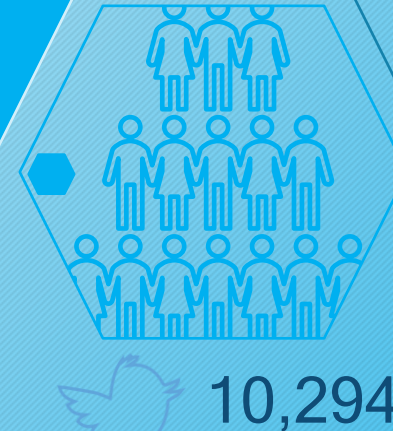
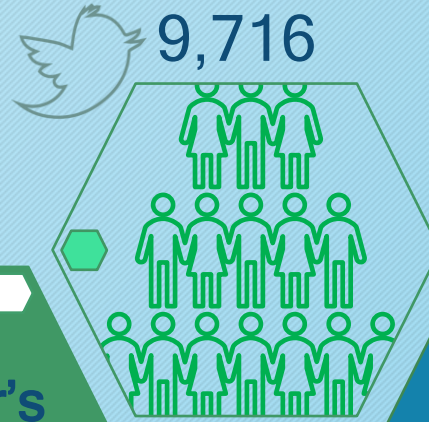
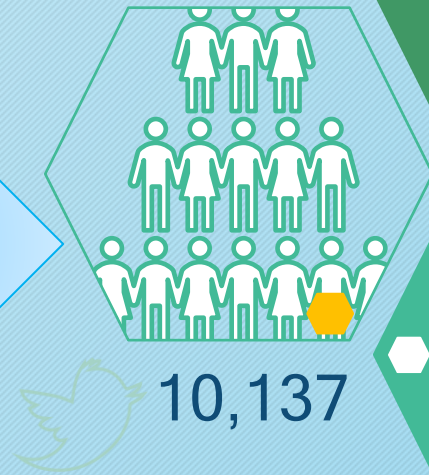


Data wrangling



Deleted tweets









- <5 words
- Non-English
- Duplicates



4075 Twitter users (≤ 10 tweets per user)


Text normalization




-  Delete the words that were used for tagging other users
-  Extend the shortened words
-  Characters were converted to lowercase
-  Numerical characters, punctuation marks, single characters, whitespace, accented characters were deleted
-  Emojis were replaced with words
-  Weblinks were replaced with their domain name only.
-  Delete the stop words
-  Text lemmatization

Word embedding

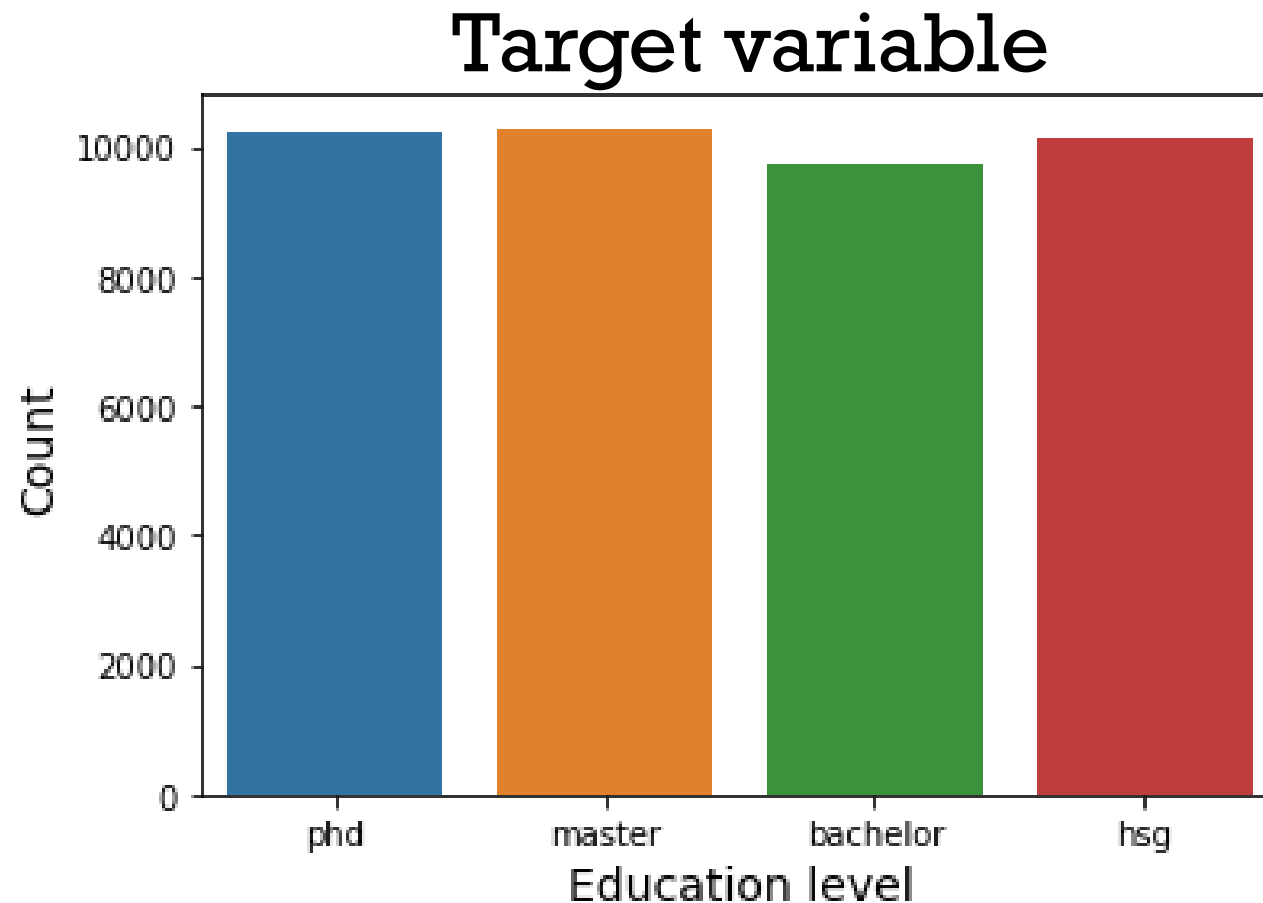


 Delete the tweets that contain words 'phd', 'master', 'bachelor', 'high school', or 'degree'

 Delete the entire tweets if the tweet contains word that has higher than 0.5 similarity with the above mentioned five words.

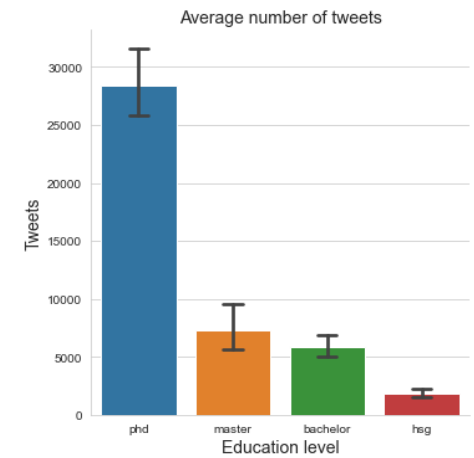
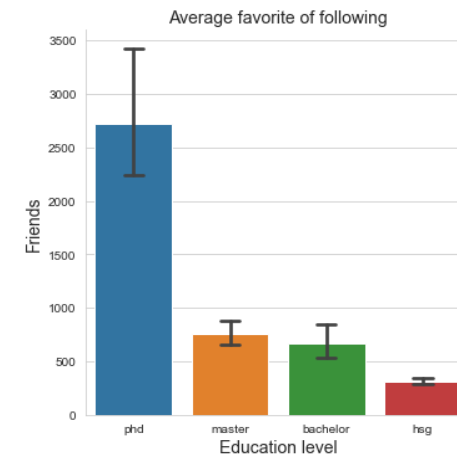
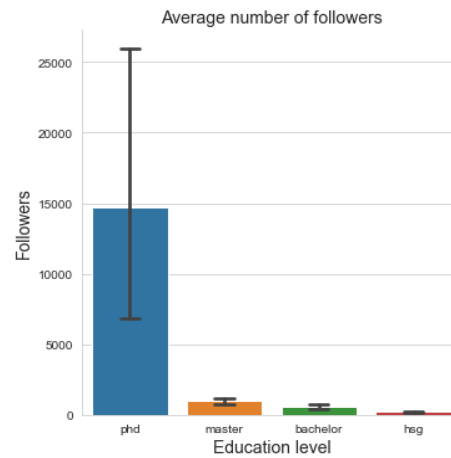
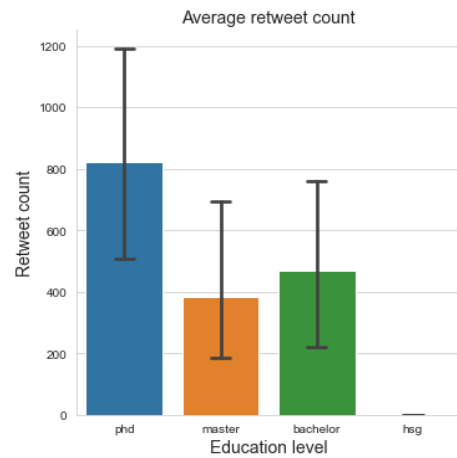
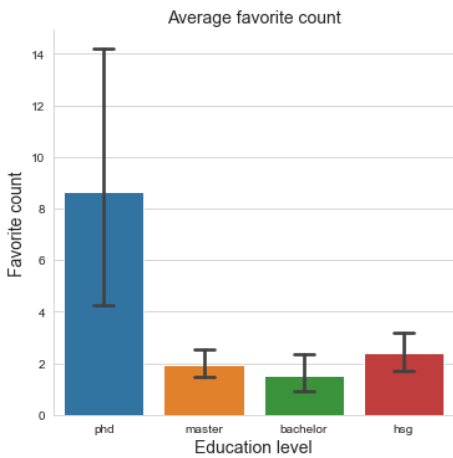


Exploratory data analysis

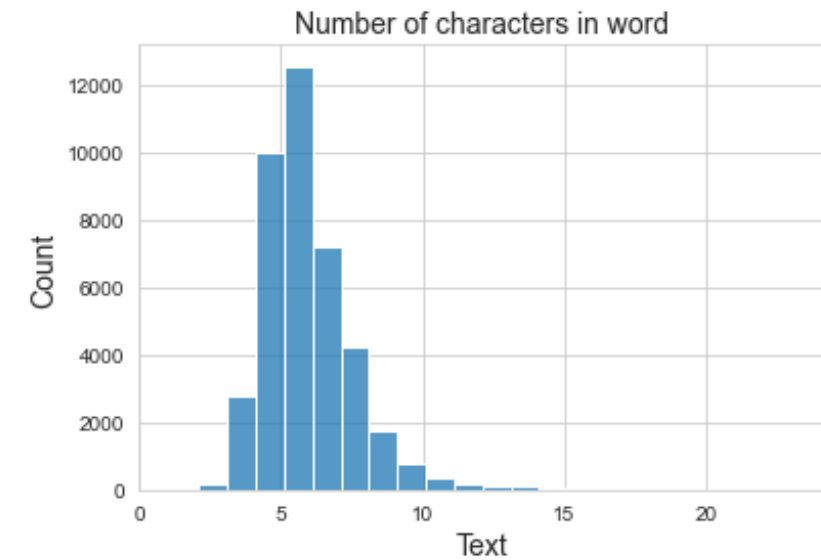
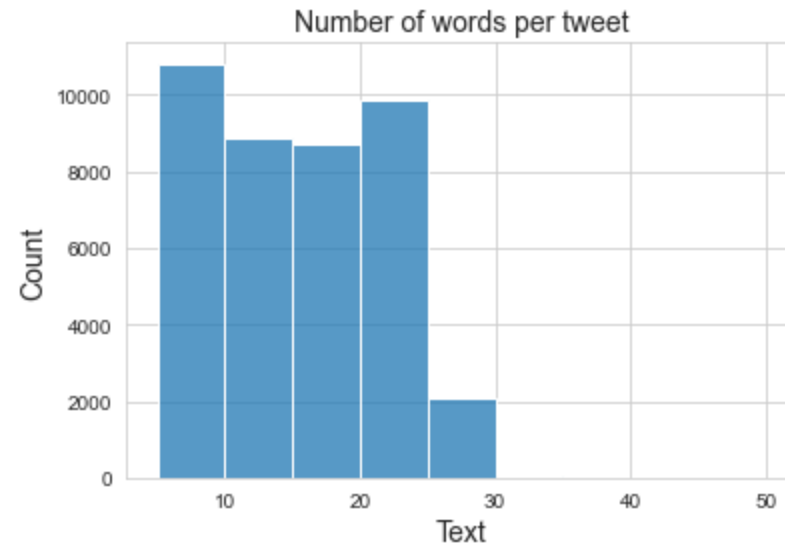
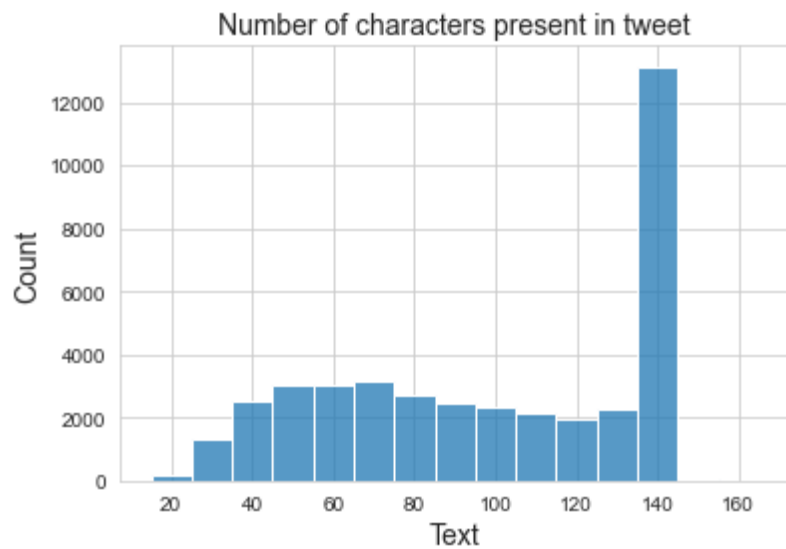


Total number of tweets for each class in the dataset

Twitter users with PhD are more active on Twitter

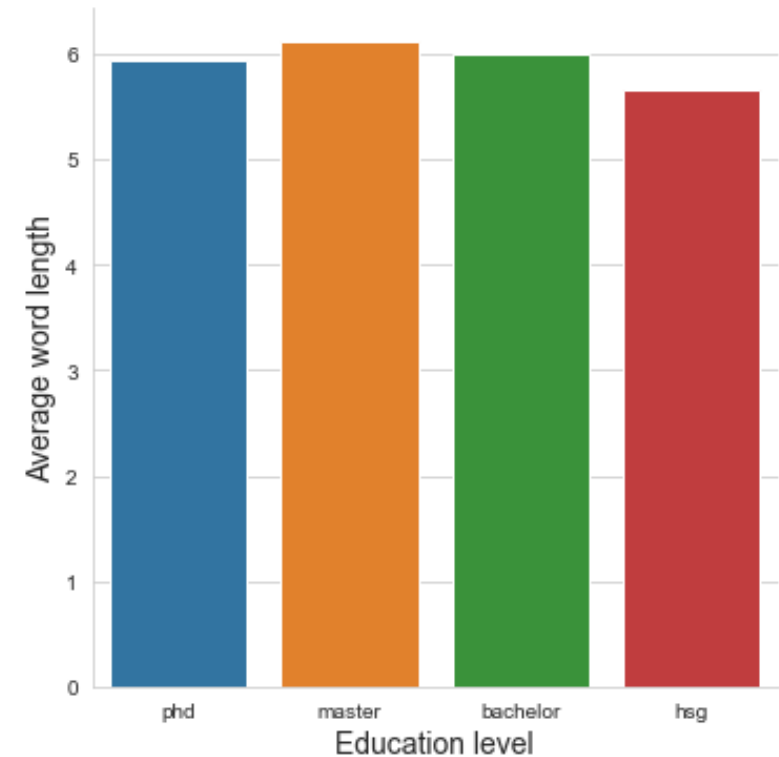
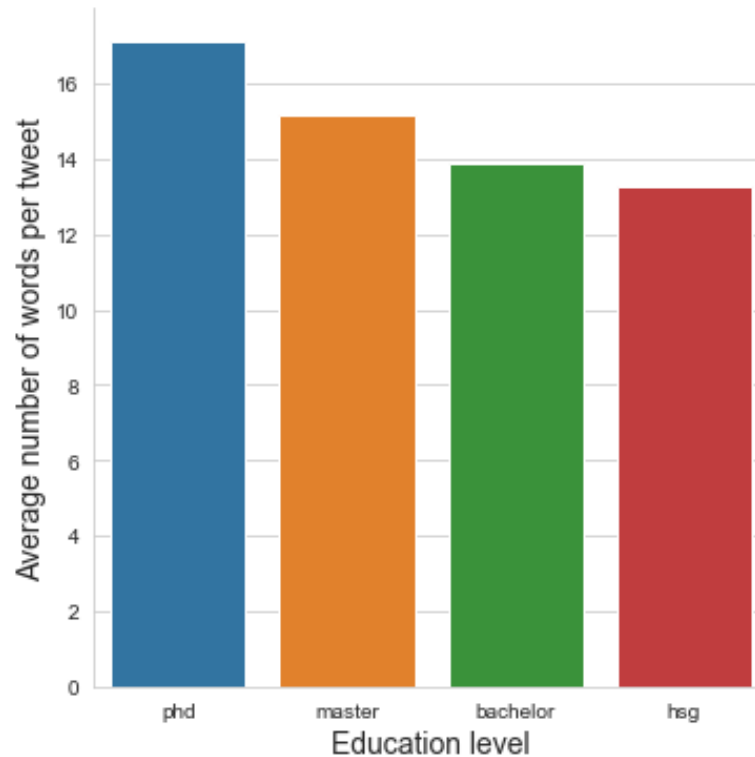
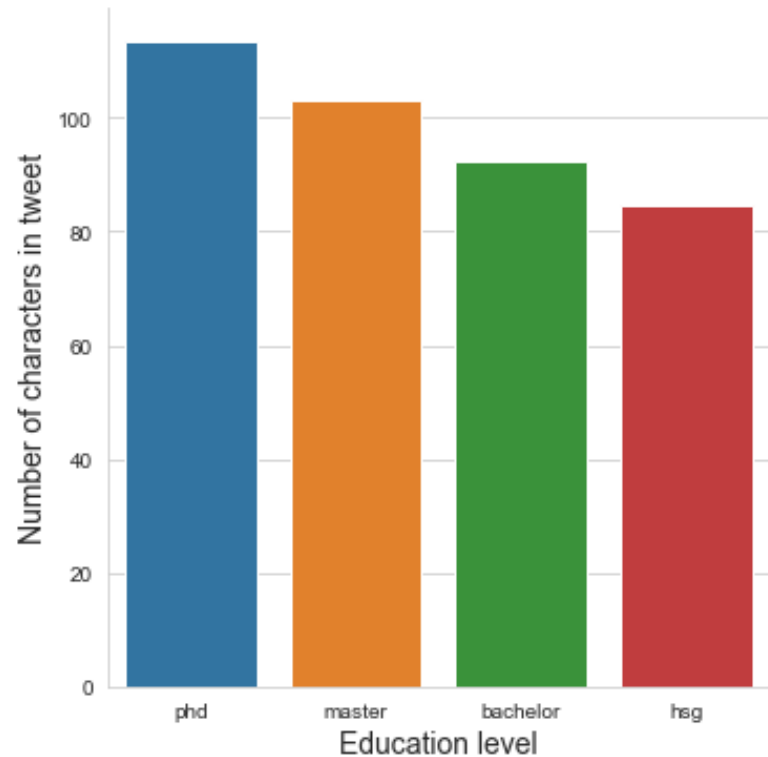


Tweet features



Most of the tweets were 140 characters in length with 5 to 25 words and average word length between 4 to 7 characters

Tweet features based on education level



Average number of characters in a tweet and average number of words per tweet significantly increased as the education level increased

Text Vectorization



 **CountVectorizer**

 **TfidfVectorizer**



Machine learning model



 Evaluated Naive Bayes, Logistic regression, Random Forest model and Linear Support Vector classification

 Hyperparameter tuning was done with grid search with 5-fold cross-validation

 Accuracy was used as the indicator of evaluation metrics.

Accuracy of models with text as predictors



Model	Accuracy	
	CountVectorizer	TfidfVectorizer
Naïve Bayes	0.49	0.48
Logistic regression	0.49	0.49
Random Forest	0.46	0.47
Linear SVM	0.49	0.49

Accuracy of models with text and tweets and user information as predictors



Model	Accuracy	
	CountVectorizer	TfidfVectorizer
Naïve Bayes	0.5	0.5
Logistic regression	0.54	0.56
Random Forest	0.67	0.67
Linear SVM	0.51	0.54



Classification report of the best performing model

Random Forest model (CountVectorizer)

	precision	recall	f1-score	support
HSG	0.58	0.86	0.69	2546
Bachelor	0.65	0.46	0.54	2383
Master	0.66	0.38	0.48	2405
PhD	0.8	0.97	0.88	2548
accuracy			0.67	9882
macro avg	0.67	0.67	0.65	9882
weighted avg	0.67	0.67	0.65	9882

Targeting specific class for business cases



 Sometimes it may be of interest for some businesses to predict the education level of a specific class

 For example, an university want to advertise about their graduate program to twitter users with bachelor degree

 Multiclass target variable could be converted to binary variable

 Bachelor vs others



Selection of the best model



Model	Accuracy	
	CountVectorizer	TfidfVectorizer
Naïve Bayes	0.65	0.67
Logistic regression	0.78	0.78
Random Forest	0.80	0.79
Linear SVM	0.78	0.78

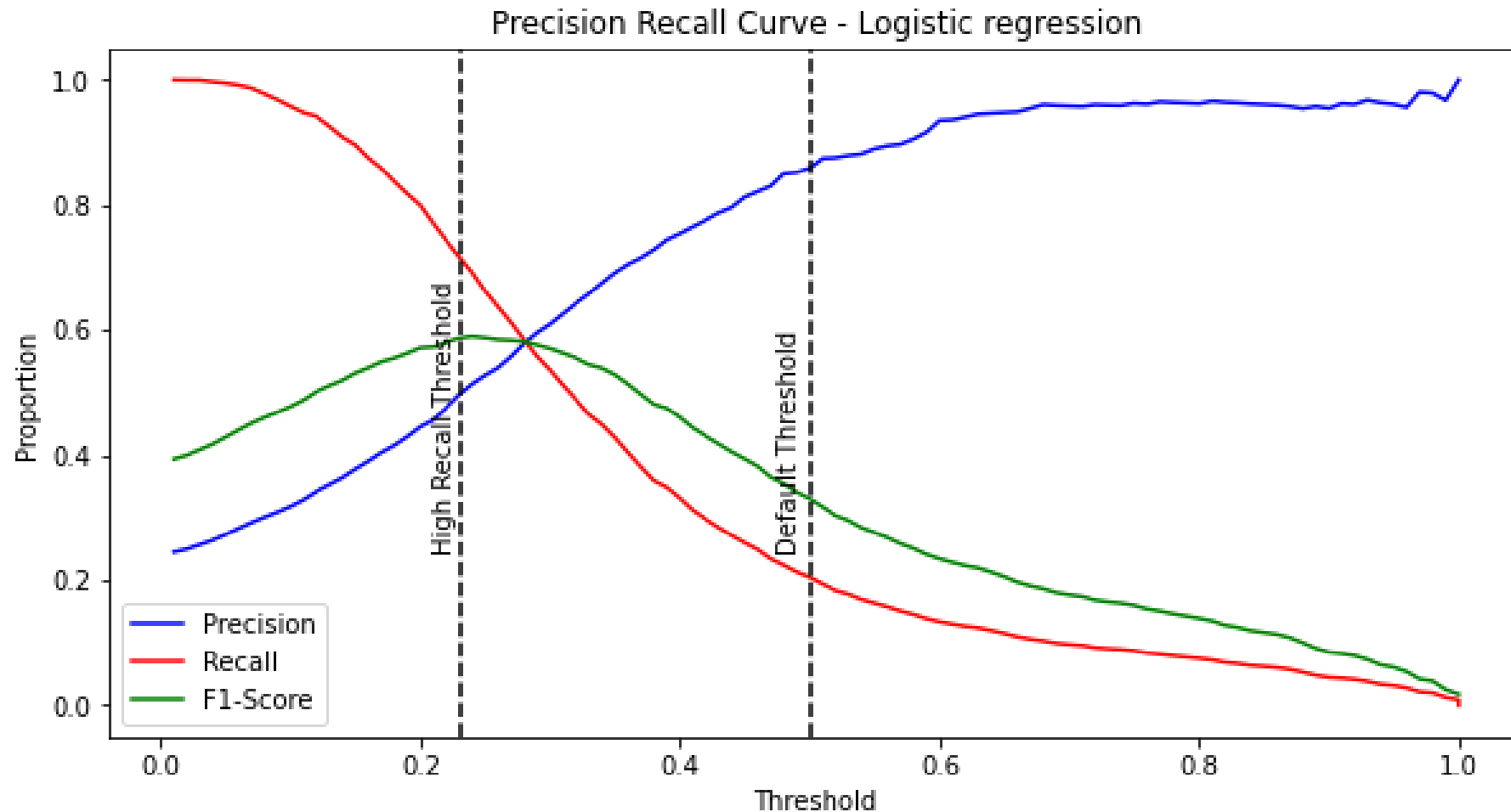


Classification report of the best performing model

Random Forest model (CountVectorizer), Threshold = 0.231

	precision	recall	f1-score	support
Others	0.89	0.79	0.84	7499
bachelor	0.51	0.69	0.59	2383
accuracy			0.77	9882
macro avg	0.70	0.74	0.71	9882
weighted avg	0.80	0.77	0.78	9882

Plot demonstrating F1 score, precision and recall at different thresholds



Conclusions



- ✈ Model performs better when information about the Twitter user and tweet were included as predictors along with text data
- ✈ Random Forest model performed the best with 67% accuracy
- ✈ Converting multiclass target variable to binary variable could improve the performance of the model to predict the specific class
- ✈ Excluding users that are posting several tweets advertising some products and additions of tweets from more users in the dataset could further improve the performance of the model



Thank You