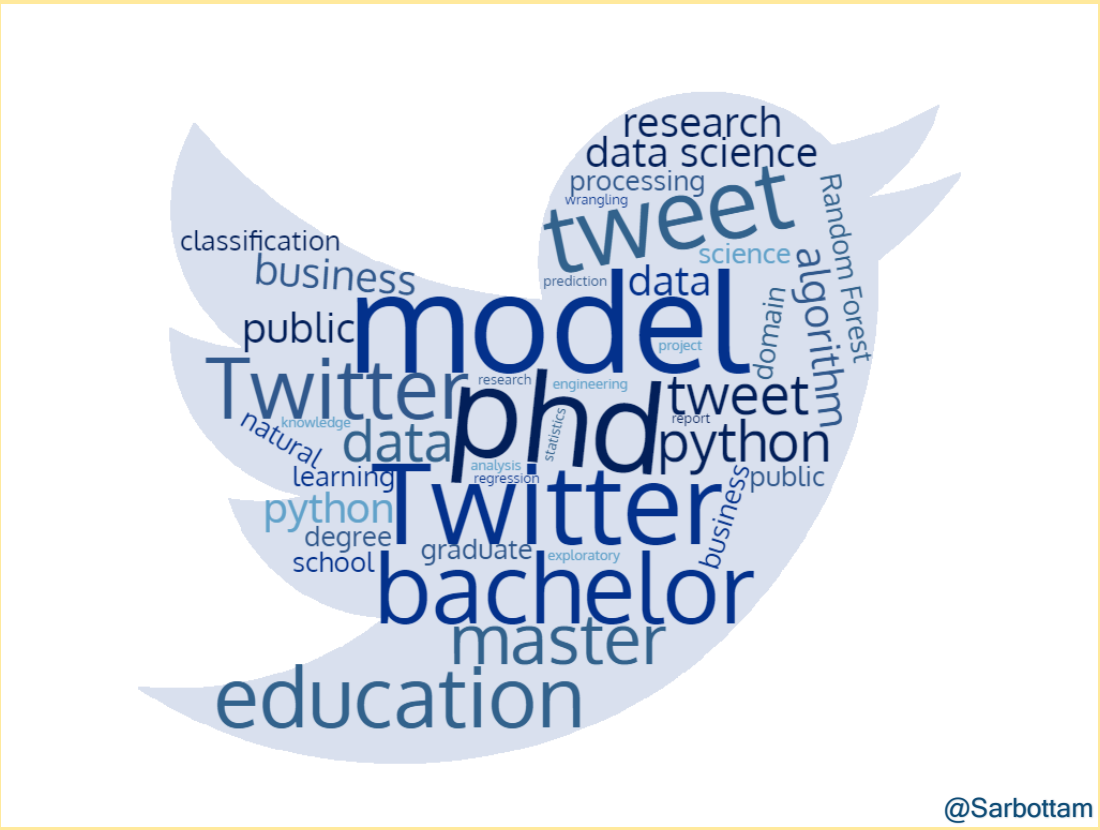


Use machine learning to predict twitter user's education level based on their tweets

Sarbottam Piya, PhD

Sarbottam.piya@gmail.com
Github.com/spiya

TABLE OF CONTENTS

INTRODUCTION

Education has been identified as one of the most important indicators of life outcomes such as employment, income and social status (ESRI 2021). In addition, an individual's perception about social and political changes are also highly dependent on their education level. People with a higher level of education show higher interest in politics, health and wellbeing. Also, people with a higher level of education show higher social trust and lower levels of political cynicism (ESRI 2021). Therefore, it might be possible to predict an individual's education level based on the opinions they express.

Social media platforms such as Facebook and Twitter are a source of large amounts of publicly available user-generated data. Applications of social media are rapidly increasing in the field of marketing, politics and health to understand people's opinions (O'Connor et al. 2010, Dredze et al. 2012, Gopinath et al. 2014). Understanding the demographics of social media users has tremendous advantages as it ensures that public messaging campaigns are reaching the targeted demographic (Culotta et al. 2016).

Understanding the demographics of social media users are useful for targeted advertising and polling across different demographics.

GOAL AND SIGNIFICANCE OF THE STUDY

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

In addition, other information such as number of times the tweets were retweeted, number of times tweets were liked, number of followers, number of people the users are following, and the total number of tweets posted by the users could also be used in the model to improve the performance of the model.

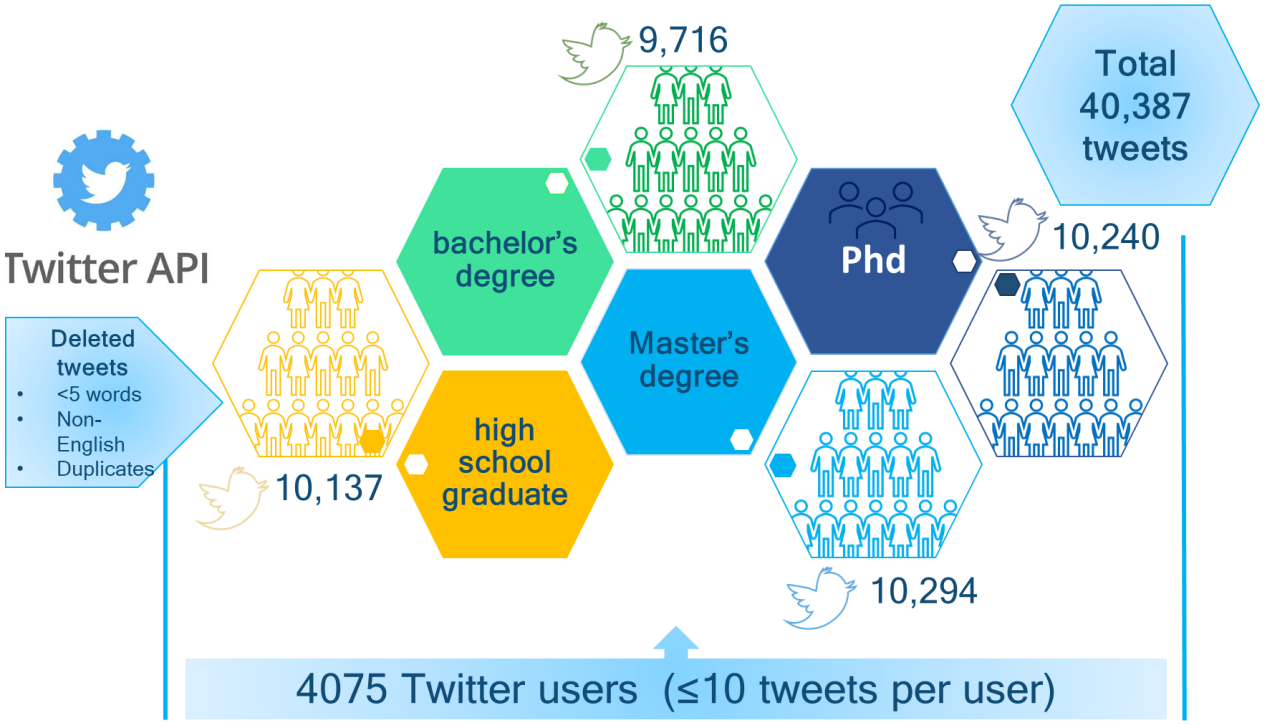
The output of this project will be useful for various purposes such as customer segmentation based on education level and understanding the opinion of people with different level of education about products and policies. Therefore, the output of this project will be useful to business/companies or policymakers who want to target Twitter users with a specific education level.

DATASET

Several Twitter users mention their education level in their Twitter bio. I used the website followerwonk.com to extract lists of 1500 users with the words “high school graduate”, “bachelor’s degree”, “master’s degree” or “PhD” degree in their Twitter bios. For each user, I extracted 50 tweets excluding re-tweets from the Twitter API. I included only those users who had at least 10 tweets in English. In addition, I also extracted: number of people following the user, the number of other users that the user follows on Twitter, the total number of tweets they posted, number of times each of their tweets were liked by other Twitter users, and number of times each of their tweets were retweeted. This resulted in a total of 179,472 tweets with 51,133 tweets for phd, 44226 tweets for master, 42,670 tweets for bachelor and 41,443 tweets for high school graduates.

DATA WRANGLING

Some of the users tweeted the same post multiple times. Therefore, I deleted duplicate tweets from the same user. In the next step, I deleted all the tweets that were in languages other than 'English'. Also tweets that were shorter than five words were deleted. After that, if any user still had more than 10 tweets, then additional tweets were deleted. This resulted in a total of 40,387 tweets with 10,294 tweets for phd, 10,240 tweets for master, 10,137 tweets for bachelor and 9,716 tweets for high school graduate.



In the next step, text normalization was carried out. To normalize the text, first I deleted the words that were used for tagging other users (words starting with @). Then, the words that were shortened by dropping letters and replacing them with an apostrophe were expanded. After that, all the letters in the text were converted to lowercase and any numerical characters, punctuation marks, single characters, whitespace, accented characters were deleted. Also, emojis were replaced with words and weblinks were replaced with their domain name only. After that stop words were deleted from the text. Finally, text lemmatization was done to group together the different inflected forms of a word.

If the tweets contain words like 'degree', 'phd', 'master', 'bachelor', and 'hsg', most likely the tweet is mentioning about the education level of the user. Therefore, I deleted any tweets that

contain any of these five words. Furthermore, I used word2vec and spaCy's prebuilt word embeddings algorithms to find words that are similar to these five words ('degree', 'phd', 'master', 'bachelor', & 'hsg') and deleted the entire tweets if the similarity was higher than 0.5.

EXPLORATORY DATA ANALYSIS

Target variable

The target variable was education level that comprised four classes - phd, master, bachelor and hsg. Number of tweets in the four classes ranged from 9,716 to 10,294 (Figure 1). Since the number of samples in each class are similar, there is no issue of class imbalance in the dataset.

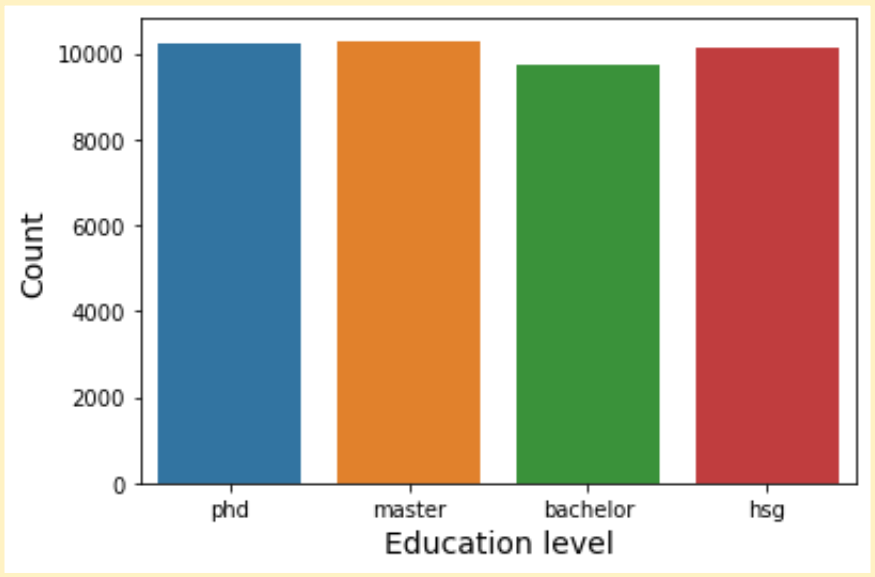


Figure 1: Total number of tweets for each class in the dataset

Twitter users' and tweet information

In general, users with PhD degree are more active in Twitter as they post the largest number of tweets than any other classes, their tweets are retweeted more, and are liked more than users of other classes. In addition, users with PhD are followed by more people and they also follow more Twitter users than users from other classes (Figure 2). After that, users with master's and bachelor's degrees showed a similar degree of activeness in Twitter. Users with only a high school degree were the least active among the four classes. The users from this class posted the least number of tweets and their tweets were also retweeted at low frequency. Also, the users from this class had the least number of followers and they also followed the least number of other users among the four classes (Figure 2).

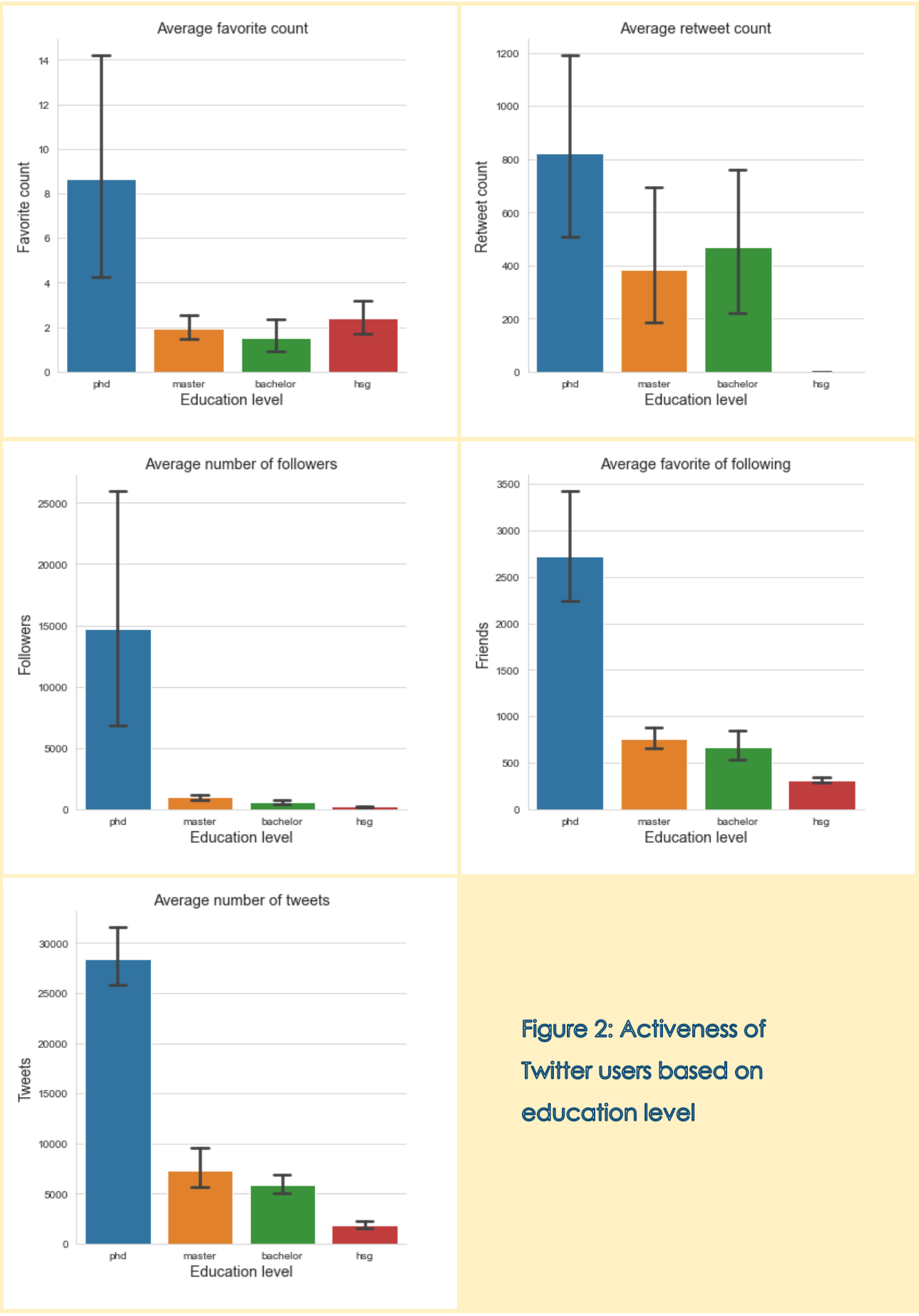


Figure 2: Activeness of Twitter users based on education level

Tweet features

Text length of tweets varied a lot; however, most of the tweets had a length of 140 characters. Furthermore, most tweets had a length of 5 to 25 words in a tweet with the majority of tweets having average word length between 4 to 7 characters (Figure 3).

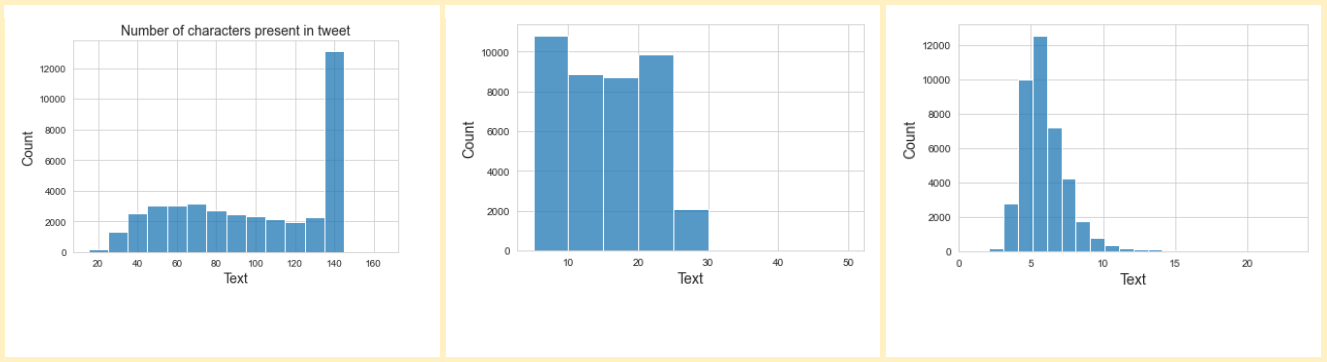


Figure 3: Tweet features of overall sample

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 (Figure 4). However, there was no specific pattern in average word length of tweets among different education levels.

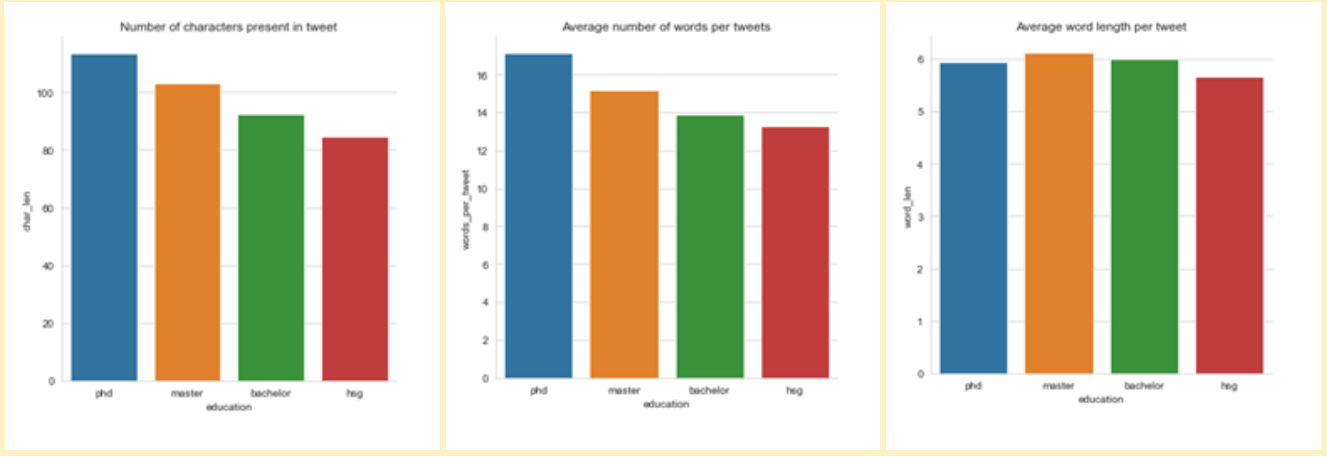


Figure 4: Tweet features based on education level

Most predictive words

To identify the most predictive words for each class, CountVectorizer was fitted in the training data and Naïve Bayes classifier was used to compute the probability of the word to be in each class for each word in the corpus. Words with the highest probability for each class were identified and word clouds were generated using the 100 most predictive words for each class (Figure 5). The top three most predictive words for PhD are paper, video and memorial. It is not unexpected that "paper" is the most predictive word for PhD as the majority of PhD discuss journal papers. Likewise, many users with PhD wished and showed respect to fallen heroes during Memorial Day compared to other classes making it the third most predictive word for PhD. The most predictive words for a master's degree were 'technology', 'program' and 'information'. Twitter users with master's degrees showed a lot of awareness and interest in technology. Here are some of the example tweets from users with master's degrees to support the claim.

- "Education Emerging Technology Steps - what is new in technology http://t.co/kpDxyfHU0v via @ThinkDevGrow"
- "People feel more empowered by internet and technology, which reflects on their buyer's journey, expectations, and eâ€¦ https://t.co/ftHJohGbJK"
- "The epitome of technology useâ€¦ 3-D printing 6th grade projects on one computer in the building I am in while I remâ€¦ https://t.co/thIxGPWpeB"
- "Globalisation has been made easier with technology, this has created a more diverse and international workforce thaâ€¦ https://t.co/R7x59H5oeG"

The top 100 most predictive words for bachelor were very diverse and included words such as 'pic', 'automatically', 'update', 'earn', 'dark_skin_tone', 'check', 'enter', 'person', 'mother' and 'jesus'. The most predictive words for bachelor included 'stat', 'ur', 'senior', 'b_tch', 'suck', 'ass', 'shit', 'f_ck', 'damn' and 'sleep' (Figure 5). The inclusion of several curse words among the most predictive words for the hsg is in accordance with the results from previous studies that have shown that users with higher social ranking and with greater number of followers are less likely to use curse words in their tweets (Steinmetz, 2014).

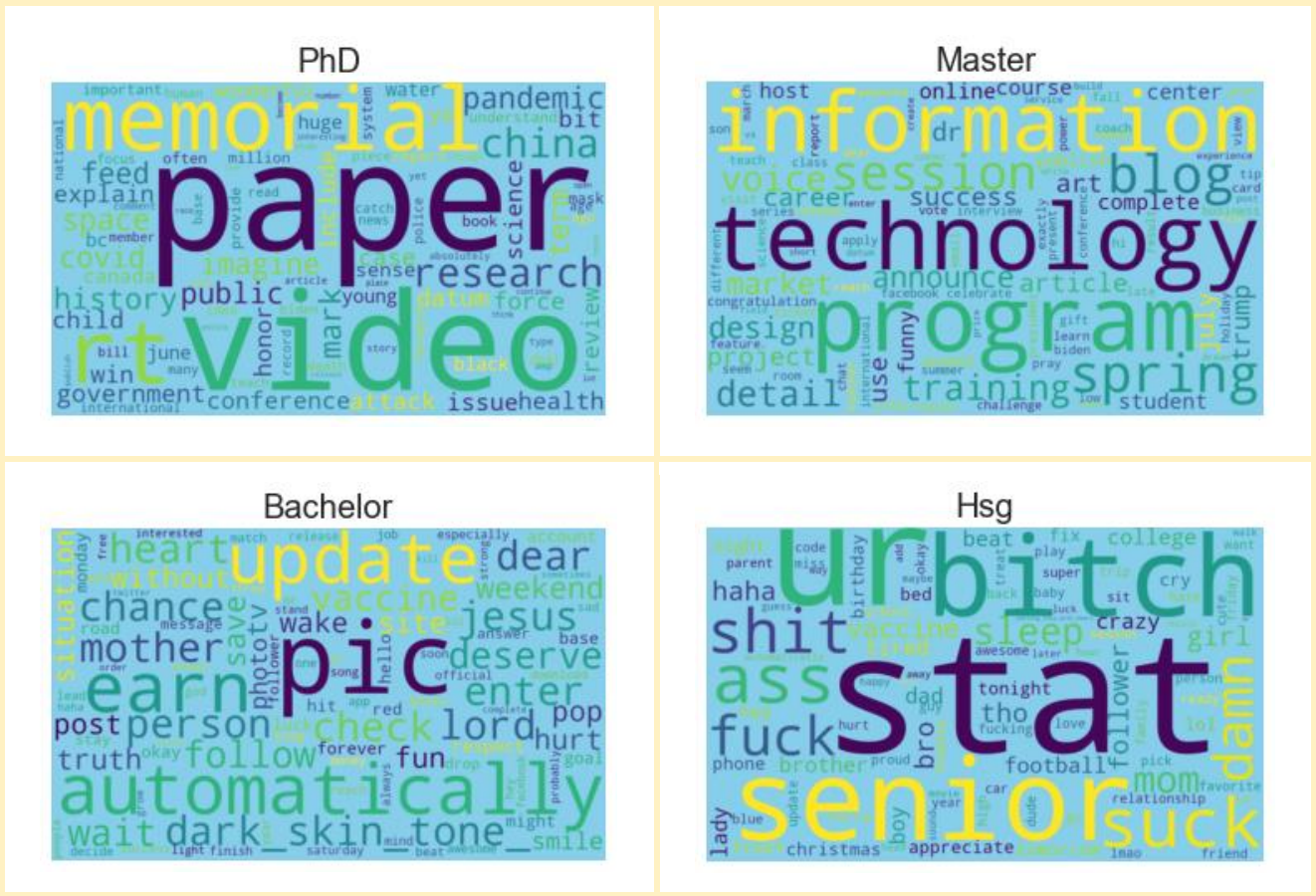


Figure 5: Word cloud of 100 the most predictive words for different education level

MACHINE LEARNING MODEL

Text Vectorization

In this project, I evaluated both CountVectorizer and TfidfVectorizer to convert text into numerical representation with unigram and bigram. CountVectorizer transforms a given text into a vector based on the frequency of each word that occurs in the entire text while TfidfVectorizer uses the term frequency (tf) and the inverse document frequency (idf) to create weighted term frequencies.

Classifiers

For each of the text vectorizations, I evaluated four different models:

1. Naive Bayes
2. Logistic regression

3. Random Forest model
4. Linear Support Vector Classification

To identify the best hyperparameters, I performed grid search with 5-fold cross-validation using accuracy as the indicator of performance metrics.

Comparison of the model

At first, I evaluated the above-mentioned machine learning model using only text data as predictors to predict the education level of Twitter users. Based on the evaluation of four different models with CountVectorizer and TfidfVectorizer, I found that the highest accuracy was 0.49 and I obtained this score using several models (Table 1).

Table 1: Accuracy of various model using only text data as predictors

Model	Grid search	CountVectorizer		TfidfVectorizer	
		Best hyperparamter	Accuracy	Best hyperparamter	Accuracy
Naïve Bayes	alpha: (1, 0.1, 0.01, 0.001, 0.0001, 0.00001)	alpha: 1	0.49	alpha: 1	0.48
Logistic regression	penalty: [l1, l2], C: [0.0001, 0.001,0.01,0.1,1,5,10,100]	C: 1, penalty: l2	0.49	C: 10, penalty: l2	0.49
Random Forest	criterion: [gini, entropy],n_estimators: [10, 50, 100, 200]	criterion: gini, n_estimators: 100	0.46	criterion: gini, n_estimators: 100	0.47
Linear SVM	C: [0.01, 0.1, 1.0, 5.0]	C: 0.1	0.49	C: 1.0	0.49

Then, I decided to evaluate each of these models with text data as well as other numerical features on Twitter user and tweet data such as number of times tweet was retweeted, number of followers etc as the predictor of the model. Again, I performed grid search analysis to identify the best hyperparameter for each of these models (Table 2). Based on the analysis, Random Forest model performed the best with the accuracy of 0.67 (Table 3).

Table 2: Accuracy of various model using text data and tweets and user information as predictors

Model	CountVectorizer		TfidfVectorizer	
	Best hyperparamter	Accuracy	Best hyperparamter	Accuracy
Naïve Bayes	alpha: 1	0.5	alpha: 1	0.5
Logistic regression	C: 5, penalty: l1	0.54	C: 5, penalty: l1	0.56
Random Forest	criterion: gini, n_estimators: 100	0.67	criterion: gini, n_estimators: 100	0.67
Linear SVM	C: 0.1	0.51	C: 1.0	0.54

Table 3: Confusion matrix of the best performing model (Random Forest model) with 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 different business cases

Sometimes it may be of interest for some businesses to predict the education level of a specific class, for example, for the military to target ads to high school graduates specifically. In such cases, the multiclass target variable may be converted to binary variable, for example, target variable with two classes- high school graduate vs others. To test if the performance of the model to predict the specific class will be further improved, I evaluated the performance of above mentioned machine learning models to predict the education level of bachelor and others. Table 4 shows the accuracy of various machine learning models.

Table 4: Accuracy of various model that predict HSG vs others

Model	CountVectorizer		TfidfVectorizer	
	Best hyperparamter	Accuracy	Best hyperparamter	Accuracy
Naïve Bayes	alpha: 0.0001	0.65	alpha: 0.00001	0.67
Logistic regression	C: 1, penalty: l2	0.78	C: 5, penalty: l2	0.78
Random Forest	criterion: gini, n_estimators: 100	0.80	criterion: gini, n_estimators: 100	0.79
Linear SVM	C: 0.1	0.78	C: 1.0	0.78

As seen in table 4, again Random Forest model with CountVectorizer performed the best with the accuracy of 80%. However, the recall value to predict the bachelor was just 0.19. It should be noted that the data is imbalanced as the ratio of class 'others' is 3 times higher than class 'bachelor'. Therefore, I did thresholding to find the best decision probability at which the model has the highest F1 score and found that at 0.231, the F1 score was the highest (Figure 6). At this threshold, the recall to predict the bachelor increased significantly to 0.59 (Table 5). Hence, converting the multiclass target variable to binary variable will be very useful if we are interested in a particular class of target variable. The performance of the model could improve further if we do under-sampling or over-sampling to balance the dataset.

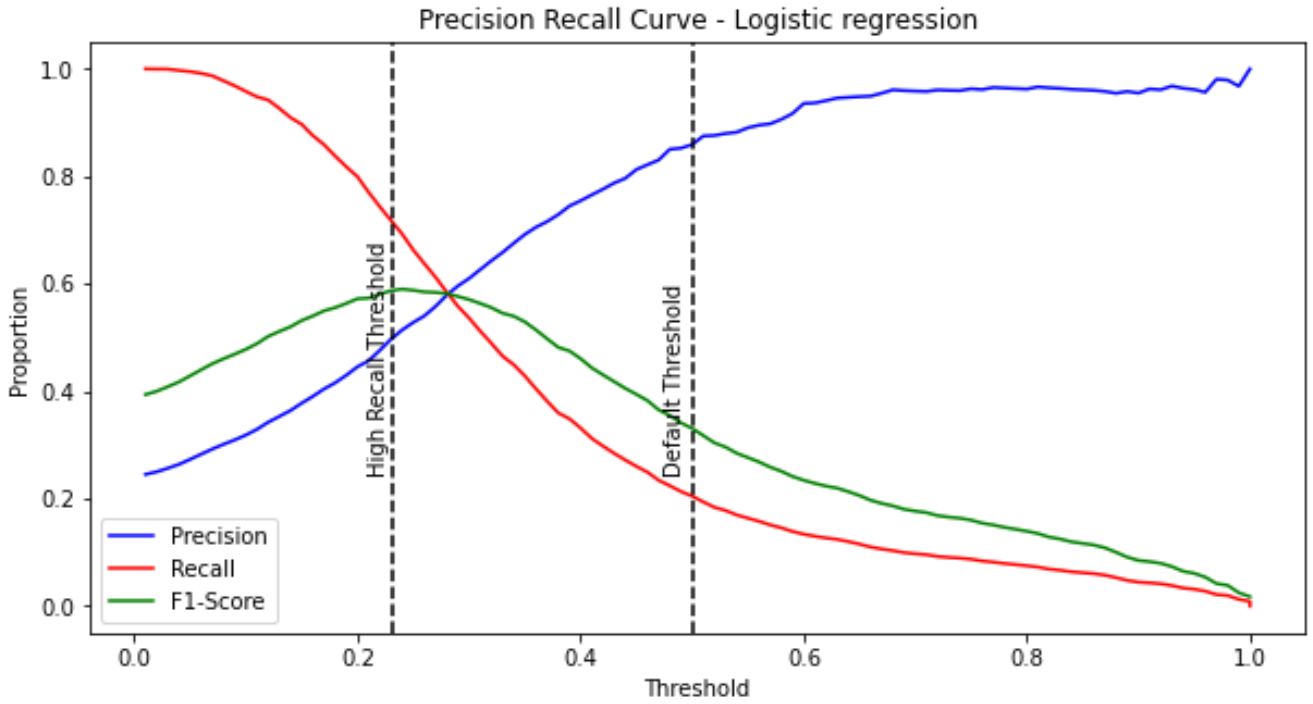


Figure 6: Plot demonstrating F1 score, precision and recall at different thresholds

Table 5: Confusion matrix of Random Forest model for HSG vs other with thresholding (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

CONCLUSIONS

Overall, the machine learning model performed very well to predict the education level of the Twitter user particularly PhD and high school graduate. With the Random Forest model, it was possible to predict the education level of Twitter users with 67% accuracy. Furthermore, if a business is interested to identify users with a specific education level for targeted advertisement, converting the multiclass target variable into a binary variable could significantly improve the performance of the model to predict the class of interest. In our case, by converting the multiclass target variable to binary variable and thresholding, the education level of 69% of the users with bachelor's degree was correctly classified by the model compared to 46% when multiclass dataset was used.

Excluding users that are posting several tweets advertising some products and additions of tweets from more users in the dataset may further improve the performance of the model.

REFERENCES

Culotta A, Ravi NK, Cutler J. 2016. Predicting Twitter user demographics using distant supervision from website traffic data. Journal of Artificial Intelligence Research, 55: 389-408.

Dredze M. 2012. How social media will change public health. IEEE Intelligent Systems, 27: 81-84.

Gopinath S, Thomas JS, Krishnamurthi L. 2014. Investigating the relationship between the content of online word of mouth, advertising, and brand performance. Marketing Science, 33: 241-258.

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA. 2010. From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11: 122-129.

Steinmetz, K. 2014. #Cursing Study: 10 Lessons About How We Use Swear Words on Twitter. https://time.com/8760/cursing-study-10-lessons-about-how-we-use-swear-words-on-twitter/

Volkova S, Wilson T, Yarowsky D. Exploring demographic language variations to improve multilingual sentiment analysis in social media. 2013. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 1815-1827.

Image credits:

www.freepik.com, www.firstpost.com, www.pxfuel.com, www.freepnglogos.com, Utsala Shrestha, Marten Bjork on Unsplash