Automatic Classification of Students on Twitter Using Simple Profile Information

Md Sadaf Noor
20366022

1.Introduction

Usefulness:

- Professional networking
- Spread of Misinformation

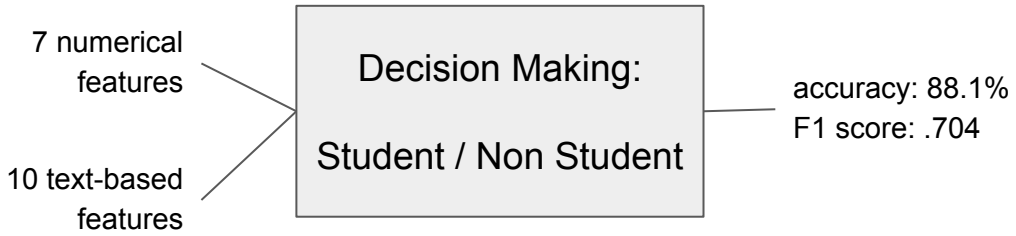
Successful classification:

- Gender
- Age
- Organization

Existing Techniques:

- surveying individuals,
- analyzing geographic proximity to educational institutions
- manually annotating users one by one
- Twitter and match them with professional mentors on LinkedIn

Proposed Technique:



S.M. Mushfiqur Rahman
16101040

1.1 Project Motivation

- Project part of investigation into students interact with misinformation
- Twitter is still considered the most misinformation
- Misinformations put people at risks
- Find out if and how students spread false information
- Try the best education system for students to stop misinformation

1.2 Ethical Consideration

- Students are more vulnerable to scammers and online predators
- Students might get targeted with spam and harmful content

Jahid Hasan Mamun
21166047

2. DATA

Sample Type	S	N-S	Total
General Stream	11	53	64
Hunter College	127	743	870
“Who to Follow”	86	8	94
Total	224	804	1028

Dataset:

Students(“S”) =225

Non-Students(“N-S”) =812

Total =1037

S = 21.7%, N-S = 78.3%

Manually Labeled Based on:

Twitter Bios, Tweet Content, LinkedIn and Instagram Profiles.

Other indicators:

Job Status, Graduation Status, followers of “Hunter College”, “Who to follow” feature

Limitations

1. 50% samples could not be identified.
2. Time-intensive manual labeling process.
3. Uneven distribution.

Aryan Shrestha
16201074

❑ 3.1 Feature Extraction

- To Training the models a combination of metadata based features and custom text based features were used. (Table 2)
- Users without descriptions, zeros were recorded for all description-based features .
- Features were scaled to similar ranges using scikit-learn's StandardScalerto improve model's performance.
- Only profile information was incorporated into this mode.
- User features were extracted using the Twitter API in batches of 100 user IDs which limits requests to 900 per 15 minutes, that means 90,000 users to be extracted per 15 minutes.

Feature Name	Value Type	Feature Description
Student?	Binary	Has "student", "estudiante", or "studying" in description. Not "students".
Friends	Continuous Numerical	Number of users followed by the account.
Occupation?	Binary	Has an occupation in description in occupation dataset.[6] Not "aspiring" or "future".
Emojis	Continuous Numerical	Number of emojis in user description squared.
Liked Posts	Continuous Numerical	Number of posts the account has 'liked'.
Parent?	Binary	Has "mom", "mama", "mother", "dad", "papa", or "father" in description.
Consecutive Upper	Continuous Numerical	Number of cons. uppercase letters in screen name.
Name Emojis	Continuous Numerical	Number of emojis in screen name.
Name Title	Binary	Has "mr.", "ms.", "mrs.", "ph.d", "ph. d", "phd", "m.d", "m. d", "doctor", or "dr." in screen name.
Link?	Binary	Has an associated link.
Tweet Rate	Continuous Numerical	Tweets posted per year by this account.
Tweet Count	Continuous Numerical	Total number of tweets (including retweets) posted by this account.
Year?	Binary	Has " '2_ ", " 2_' ", "class of 202x", "freshman", or "sophomore" in description.
Followers	Continuous Numerical	Number of users following the account.
Alum?	Binary	Has "alum" in description.
Views My Own?	Binary	Has "(views)/(opinions) (mine)/(my own)" or "rts not endorsements" in description.
Verified?	Binary	Has been verified.
Created At*	Continuous Numerical	Timestamp for creation time of account.
Account Age*	Continuous Numerical	Time (in years) since creation time of account.
Last Tweet Time*	Continuous Numerical	Time (in years) since the account's last tweet.

Table 2: Profile-Based Features (ordered by importance)
*Indicates feature was removed from the final model

❑ 3.2 Feature Selection

- Twenty original features were extracted from each user, and three were removed due to low importance to the machine learning models .
- Importance was assessed via logistic regression importance rankings, decision tree rankings, random forest rankings, and LASSO rankings.
- This feature removal was verified via an ablation study, whichshowed that all remaining features had a positive importance averaged across all six models (Fig. 1)

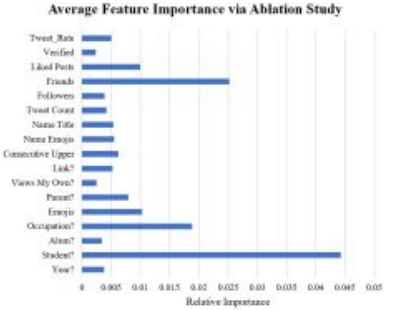


Figure 1: Relative Importance of User Features (excluding removed features)

Md. Mofaqkhayrul Islam Mim
20266013

3.3 Model Selection

- Six distinct machine learning models were implemented by the authors in scikit-learn (Pedregosa et al., 2011) to create student identification classifier: 1) Logistic Regression, 2) Random Forest, 3) SVM, 4) K-Nearest Neighbors, 5) AdaBoost and 6) a Stacked Classifier.

- To optimize each model for the highest F1 score, they used a grid search of model hyperparameters combined with 10-fold cross-validation

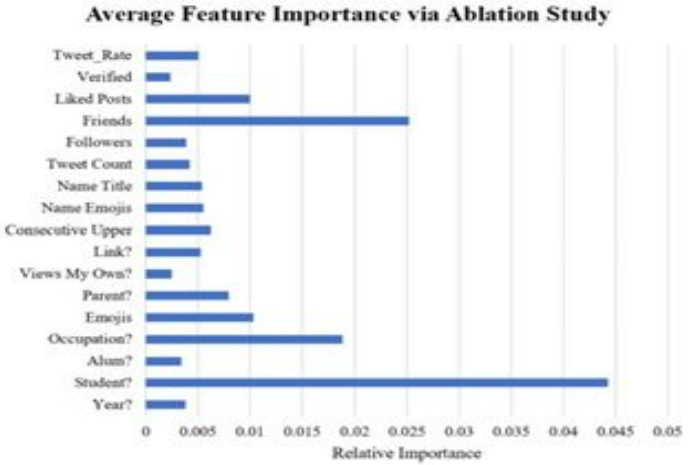


Figure 1: Relative Importance of User Features (excluding removed features)

- Fig-1, showed that, The "Student?" element was heavily emphasized by the models. A simple "if-statement classifier" was constructed to see if the machine learning models were adding anything to the categorization. It classified a user as a student if the “Student?” feature was set to 1.

Md. Mofaqkhayrul Islam Mim
20266013

3.4 Model Tuning

- They enhanced the results by adding regions within their prediction probabilities, which they term "gray zones," where their models would identify a user as "uncertain," after picking the best three model types based on F1 score and AUROC.

- By using 10-fold cross-validation to examine 39 candidates, these regions were discovered. For each of the three model types, two gray area candidates were chosen based on accuracy and F1 score.

Md Tarikul Islam
21166020

4.0 Results

	Accuracy	F1	AUROC
Logistic Regression	87.4	.678	.896
Random Forest	86.7	.677	.910
SVM	86.8	.643	.898
KNN	81.6	.601	.750
AdaBoost	87.1	.683	.785
Stacked Classifier	**88.1**	**.704**	**.917**
If-Statement	80.0	.340	-
Tweet-based SVM	77.1	.222	.646

Table 3: Model Comparison

	Accuracy	F1	Coverage
Logistic Regression (0.3, 0.4)	89.6	.760	89.7
Logistic Regression (0.35, 0.45)	89.8	.748	91.3
Random Forest (0.3, 0.5)	89.4	.749	88.4
Random Forest (0.35, 0.45)	87.4	.713	94.5
Stacked Classifier (0.25, 0.55)	**90.3**	**.761**	**89.7**
Stacked Classifier (0.35, 0.55)	89.3	.735	93.2

Table 4: Gray Area Model Comparison

S M Ibrahim
21166045

Conclusion

- In this paper, the author introduces a metadata-based machine learning model to predict student Twitter users accurately.
- They also introduce a gray-area model that achieves 90.3% accuracy without leaving many users unlabeled. Their models improve upon past research by providing more accurate, more efficient, and faster classifications due to their use of only simple profile information.
- Currently, they are working to apply this student classifier in a preliminary study of student interactions with COVID-19 related misinformation on Twitter.