Automatic Classification of Students on Twitter Using Simple Profile Information

Christopher Wun[*]
Hunter College High School
New York, NY, 10128
ckywun123@gmail.com

Lili-Michal Wilson[*]
Hunter College High School
New York, NY, 10128
lilimmwilson@gmail.com

Abstract

Obtaining social media demographic information using machine learning is important for efficient computational social science research. Automatic age classification has been accomplished with relative success and allows for the study of youth populations, but student classification—determining which users are currently attending an academic institution—has not been thoroughly studied. Previous work (He et al., 2016) proposes a model which utilizes 3 tweet-content features to classify users as students or non-students. This model achieves an accuracy of 84%, but is restrictive and time intensive because it requires accessing and processing many user tweets. In this study, we propose classification models which use 7 numerical features and 10 text-based features drawn from simple profile information. These profile-based features allow for faster, more accessible data collection and enable the classification of users without needing access to their tweets. Compared to previous models, our models identify students with greater accuracy; our best model obtains an accuracy of 88.1% and an F1 score of .704. This improved student identification tool has the potential to facilitate research on topics ranging from professional networking to the impact of education on Twitter behaviors.

1 Introduction

In recent years, social media has become the focus of more scientific research studies (Khang et al., 2012), and is being used as a valuable means of understanding human tendencies and social circles. Due to rapid developments in technology, social media has become a hub for scientific communication and research (Rowlands et al., 2011), a source of entertainment and news, and a place for networking and socialization (Whiting and Williams, 2013). Twitter is one of the most popular social media platforms, especially among young people; 44% of young adults aged 18-24 and 31% of people aged 25-29 use Twitter, with 25% using it daily and 14% using it several times per day.[1]

The ability to group Twitter users by demographic information is crucial in understanding the behaviors of certain subsets of the population. Twitter provides limited information on user demographics, necessitating the development of machine learning models to predict non-explicit user features. Several studies have demonstrated that it is possible to classify users by gender (Knowles et al., 2016), age (Morgan-Lopez et al., 2017; Simaki and Iosif Mporas, 2018; Smith and Gaur, 2018), whether or not they are an organization (Wood-Doughty et al., 2018), and more.

In this paper, we aim to classify Twitter users as students or non-students. While many age classification models have been proposed, understanding this subset of the age classification problem may provide new insights into behavior on Twitter. Grouping users into students and non-students has applications ranging from professional networking (He et al., 2016) to understanding these users' role in the spread of misinformation, and by extension, the role that education plays in their online behaviors (Chen et al., 2015). Previous studies of students' social media interactions have obtained student samples using relatively low-yield, low-accuracy, or low-coverage techniques such as surveying individuals, analyzing geographic proximity to educational institutions, or manually annotating users one by one (Chen et al., 2015; Miller and Melton, 2014; Moreno et al., 2016; Veletsianos and Kimmons, 2016; Hanson et al., 2013). Research into automatic demographic prediction of student users could facilitate such student-based studies.

[*] Authors contributed equally

[1] https://www.pewresearch.org/fact-tank/2019/04/10.

The most prominent application of an automatic classification model to student demographic research is described in He et al. (2016). This study develops a machine learning model that uses user tweet content to identify students on Twitter and match them with professional mentors on LinkedIn. Despite the model's relatively high accuracy (84%), its use of user tweet content is restrictive because it is not always possible to obtain complete tweet data. Protected accounts cannot be classified using this model, and researchers without advanced Twitter API access may not be able to obtain sufficient tweet data.

Here, we propose new models for student classification on Twitter which achieve high accuracy using only simple profile information. Our use of profile-based features obviates the need for tweet content, which allows for the classification of protected accounts and makes our model more accessible to researchers without access to advanced Twitter APIs. In addition, user profile information can be requested from the Twitter API at a faster rate than a sample of user tweets. These novel student identification models thus show potential for more accurate, accessible, and efficient classifications in demographic studies of social media users.[2]

1.1 Project Motivation

The development of this student identification classifier is part of a larger investigation into the way that students interact with misinformation on Twitter. Though the platform has implemented fact-checking guidelines, it is still considered one of the worst social media outlets in terms of spreading false information.[3] This spread of misinformation has necessitated research into ways of detecting rumors in tweets and predicting which kinds of users are most susceptible to resharing it (Vosoughi et al., 2018; Khan and Idris, 2018; Bodaghi et al., 2019; Margolin et al., 2017). This work is especially important in the current pandemic, where false claims on Twitter have pushed people to put themselves or others at risk (Barua et al., 2020; Rosenberg et al., 2020).

We hypothesize that users who are currently receiving education may behave differently on the platform due to a recent emphasis placed on fact checking at academic institutions.[4] Understanding if students play a different role in the spread of false information is relevant when determining what kinds of education can be implemented to prevent the spread of misinformation and how certain users could be targeted with that education.

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

Table 1: Dataset Distribution, Sample Types are defined in §2.1-2.3. S denotes student, while N-S denotes non-student.

1.2 Ethical Considerations

With the creation of a student classification model comes potential for misuse. Being able to determine whether or not a user is a student, even if their tweets are not public, could enable people to target student users with spam or other harmful content. As previously noted, students are generally a younger population who may be more vulnerable to online predators or scammers who are using this tool.

2 Data

Creating the dataset used to train the student identification model required obtaining a set of Twitter users who could be labeled as student or non-student and then manually labeling each user. These users were obtained from a combination of a general sample of Twitter users, a sample of the followers of a university Twitter account, and Twitter's “Who to Follow” feature. Ultimately, 1,037 users were labeled, with 225 students and 812 non-students. The distribution of the sample is shown in Table 1, with “S” representing students and “N-S” representing non-students. Users were manually labeled based on their Twitter bios, tweet content, or other online profiles (e.g. LinkedIn, Instagram).

A user was labeled as a student if their accounts indicated that they were in high school, undergraduate school, or graduate school in the spring of 2020. General examples of labeled students included users that provided a future graduation year

[2] Annotated datasets and student classifiers are publicly available at https://github.com/christopherwun/twitter-student-classifier

[3] https://www.washingtonpost.com/technology/2020/04/07/twitter-almost-60-percent-false-claims-about-coronavirus-remain-online-without-warning-label/

[4] https://www.nytimes.com/2020/02/20/education/learning/news-literacy-2016-election.html

in their bio; listed a current school in an associated LinkedIn account; or discussed homework, test scores, and/or their school's policies in their tweets.

A user was labeled as a non-student if their online profile showed that they had graduated from a higher institution, listed a full-time job, or were organizations or spam bots. General examples included noting a past graduation date or degree in their Twitter bio or listing jobs or graduation dates on an associated LinkedIn profile.

Labeling was conducted by the listed authors, and inter-annotator agreement was measured on a representative subset of 435 users. The subset demonstrated substantial agreement as shown by a Cohen's Kappa of 0.73 (McHugh, 2012). In the case of a disagreement, the first ascribed label was used in the final dataset. To avoid annotator bias in assigning these labels, each annotator labeled half of the dataset then reviewed the other annotator's labels blind.

2.1 General Twitter Stream

In order to get a sample of the overall Twitter population, we collected tweets written in English from the Twitter stream for ten minutes. This general stream was used to obtain a variety of users across Twitter for labeling. However, the proportion of identifiable students within this set was low, necessitating other sampling to increase the number of student examples.

2.2 College Followers

In order to increase the number of students in the dataset, followers of @Hunter_College (the verified account of a prominent NYC college) were sampled directly and manually labeled. This sample was taken under the assumption that the pool of followers of this school account would contain an increased proportion of students.

2.3 "Who to Follow": Similar Users

Finally, to further increase the number of students in the set, Twitter's "Who to Follow" feature was applied to students that had been identified through the general stream and Hunter College followers. The "Who to Follow" feature identifies users similar to a given user. When applied to previously labeled students, it often identified potential students with similar roles in the social network. These potential students were then labeled based on the same criteria as the previous two samples.

2.4 Limitations

Approximately 50% of sampled users could not be identified with certainty as a student or a non-student, and such users were not labeled and not included in the dataset.

Another potential issue with the dataset is its relatively small size, due to the time-intensive manual labeling process. In addition, the methods used to upsample students led to an uneven distribution of student users across data sources (Table 1). It is important to keep these limitations in mind when considering the results of our model.

3 Methods

3.1 Feature Extraction

A combination of metadata-based features and custom text-based features was used to train the models (Table 2). For users without descriptions (approximately 7.4% of the dataset), zeros were recorded for all description-based features. This includes both explicitly student-leaning features like "Student?" and "Year?" as well as explicitly non-student-leaning features like "Alum?" and "Occupation?". Features were scaled to similar ranges using scikit-learn's StandardScaler (Pedregosa et al., 2011) in order to improve model performance.

Unlike other student-identifying models, only profile information was incorporated into this model (i.e., no tweet data or other user data was considered). User features were extracted through the Twitter API in batches of 100 user IDs. The Twitter API limits requests to 900 per 15 minutes, meaning that our method allows 90,000 users to be extracted per 15 minutes.[5] A tweet-content-based approach such as the one used in He et al. (2016) must acquire user features by accessing their tweets, and requests for tweets from different users' timelines cannot be batched. Therefore, a tweet-based method allows the features of only 900 individual users to be extracted per 15 minutes.

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 (See Table 2). 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, which

[5] https://developer.twitter.com/en/docs/twitter-api/rate-limits

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

showed that all remaining features had a positive importance averaged across all six models (Fig. 1).

3.3 Model Selection

We looked into six different machine learning models implemented in scikit-learn (Pedregosa et al., 2011) to create our student identification classifier: 1) Logistic Regression, 2) Random Forest, 3) SVM, 4) K-Nearest Neighbors, 5) AdaBoost (a Random Forest model retrained multiple times, each time increasing the weight of previously misclassified users), and 6) a Stacked Classifier (a Logistic Regression model using the outputs of models 1-4 as inputs). We performed a grid search of model hyperparameters with 10-fold cross-validation to optimize each model for the highest F1 score.

After noticing that models placed a heavy emphasis on the “Student?” feature (Fig. 1), a simple

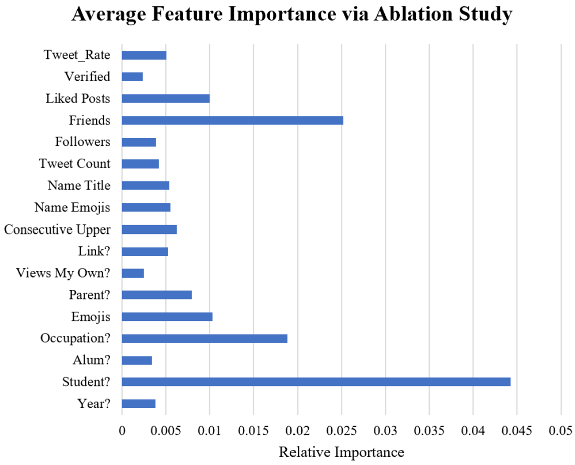

Figure 1: Relative Importance of User Features (excluding removed features)

[6] https://gist.github.com/wsc/1083459

"if-statement classifier" was created to determine whether or not the machine learning models were adding anything to the classification. It labeled a user as a student if the "Student?" feature was equal to 1. We also created a tweet-content-based SVM model to use as a baseline (described in §4.1).

3.4 Model Tuning

After selecting the top three model types based on F1 score and AUROC (defined in §4), we further improved our results by adding regions within our prediction probabilities, which we call "gray areas," where our models would classify a user as "uncertain." These regions were identified by testing 39 candidates with 10-fold cross-validation. Two gray area candidates were selected based on accuracy and F1 score for each of the three model types.

4 Results

Table 3 reports the results from running the six different classifiers and two baseline models without implementing a gray area. The test set reflected the distribution of the overall dataset, with a ratio of students to non-students of around 1:4. We report the accuracy (percent of correctly predicted students or non-students over the total), F1 score (the harmonic mean of precision and recall), and area under the Receiver Operator Characteristic curve (AUROC; an ROC curve plots the true positive rate against false positive rate at different decision boundaries).

Before considering gray area, the Stacked Classifier outperformed all other classifiers in accuracy, F1, and AUROC. Every model outperformed the "if-statement classifier" in both accuracy and F1 score, indicating that the additional complexity and extra features considered in the machine learning models contributed to their success.

	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

Table 4 reports the results from running the top three classifiers (Stacked Classifier, Logistic Regression, and Random Forest) using two gray area boundaries that were identified through 10-fold cross validation as described in §3.3. AUROC is not reported here as it is the same in Table 3. Instead, we report coverage, which is the percentage of test examples that the gray area model labeled. The thresholds for the gray area are reported in the table as (lower bound, upper bound). Of these gray area models, the Stacked Classifier with thresholds at (0.25, 0.55) outperformed the rest, with an accuracy of 90.3% and an F1 of .761. The model also retained a relatively high coverage, labeling nearly 90% of the test examples.

It is important to note that, while the gray area models outperform the normal models in both accuracy and F1, much of that can be attributed to the fact that certain false negative examples are being ignored. Depending on the intended use, a higher accuracy version of the Stacked Classifier may be used sacrificing coverage, or vice versa.

	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

4.1 Comparison to Previous Models

In order to better understand how our model compares to previous student identification attempts, we recreated the tweet content classification model described in He et al. (2016). They used the relative frequency of three expressions—"HAHA"/"LOL", emojis, and hashtags—among the 200 most recent user tweets as features in their LIBSVM model. We trained and tested this model on our labeled dataset, then optimized the model's hyperparameters as described in §3.3. The tuned model achieved an accuracy of 77.1%, an F1 score of 0.222 and an AUC of 0.646 (Table 3). The accuracy of this model was lower than the reported 84% in He et al. (2016). The F1 and AUC scores cannot be compared since they were not reported in the original paper.

A likely explanation for the inconsistency in accuracy lies in dataset construction. He et al. (2016) partially automated their data annotation by using regular expressions from tweets to identify student candidates before manually labeling these users. This technique may have increased the proportion of high-tweeting users in their dataset and made their tweet-based approach to classification more effective. In contrast, our manual annotations did not rely on vast amounts of tweet content to assign a label. Thus, our model is more reliable when analyzing users with lower tweet counts.

Our profile-based student classifier outperformed the tweet-content classifier in accuracy, F1 score, and AUROC when applied to our dataset. Therefore, this model improves upon existing models and can be used as a tool in future student demographic research.

4.2 Error Analysis

Organizations that are tailored towards students and university accounts commonly appear among the misclassified users. In order to mitigate this misclassification, our student classifier could be used in conjunction with an individual versus organization classifier similar to the one presented in Wood-Doughty et al. (2018). Organizations could be filtered out before applying the student classifier, and the model's false positive rate would likely decrease.

Among the 24 users without descriptions (7.7% of the test set), only 2 of them were misclassified: one student, and one non-student. The model's accuracy is thus comparable between users without descriptions (accuracy = 91.7%) and the general test set (accuracy = 88.1%). As the sample size is small, we cannot conclude that users without descriptions are labeled more accurately than the general set; however, this result does indicate that users without descriptions can still be accurately classified by our model.

5 Conclusions

In this paper, we introduce a metadata-based machine learning model to accurately predict student Twitter users. We also introduce a gray-area model that achieves 90.3% accuracy without leaving many users unlabeled. Our models improve upon past research by providing more accurate, more efficient, and faster classifications due to their use of only simple profile information.

Currently, we are working to apply this student classifier in a preliminary study of student interactions with COVID-19 related misinformation on Twitter.

Acknowledgments

We thank Zach Wood-Doughty for his guidance on feature selection and model optimization, and the AACL-IJCNLP SRW reviewers and mentors for their comments and suggestions. We would also like to thank Gilana Reiss and Philip Frankel for their support during the project conception.

References

Zapan Barua, Sajib Barua, Salma Aktar, Najma Kabir, and Mingze Li. 2020. Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*.

A. Bodaghi, S. Goliaei, and M. Salehi. 2019. The number of followings as an influential factor in rumor spreading. *Applied Mathematics and Computation*, 357:167–184.

Xinran Chen, Sei-Ching Joanna Si, Yin-Leng Theng, and Chei Sian Lee. 2015. Why students share misinformation on social media: Motivation, gender, and study-level differences. *The Journal of Academic Librarianship*, 41.

Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. 2013. Tweaking and tweeting: Exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of Medical Internet Research*, 15(4):e62.

Ling He, Lee Murphy, and Jiebo Luo. 2016. Using social media to promote STEM education: Matching college students with role models. In *Machine Learning and Knowledge Discovery in Databases*, pages 79–95. Springer International Publishing.

M. Laeeq Khan and Ika Karlina Idris. 2018. Recognise misinformation and verify before sharing: a reasoned action and information literacy perspective. *Behaviour Information Technology*, 38:1194–1212.

Hyoungkoo Khang, Eyun-Jung Ki, and Lan Ye. 2012. Social media research in advertising, communication, marketing, and public relations, 1997–2010. *Journalism Mass Communication Quarterly*, 89:279–298.

Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics.

Drew B. Margolin, Aniko Hannak, and Ingmar Weber. 2017. Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, 35:196–219.

Marry L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, pages 276–282.

Robert Miller and James Melton. 2014. College students and risk-taking behaviour on twitter versus facebook. *Behaviour Information Technology*, 34:678–684.

Megan A. Moreno, Alina Arseniev-Koehler, Dana Litt, and Dimitri Christakis. 2016. Evaluating college students' displayed alcohol references on facebook and twitter. *Journal of Adolescent Health*, 58(5):527–532.

Antonio A. Morgan-Lopez, Annice E. Kim, Robert F. Chew, and Paul Ruddle. 2017. Predicting age groups of twitter users based on language and metadata features. *PLoS One*, 12(8).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. 2020. The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic. *Canadian Journal of Emergency Medicine*, 22:418–421.

Ian Rowlands, David Nicholas, Bill Russell, Nicholas Canty, and Anthony Watkinson. 2011. Social media use in the research workflow. *Learned Publishing*, 24:183–195.

Vasiliki Simaki and Vasileios Megalooikonomou Iosif Mporas. 2018. Age identification of twitter users: Classification methods and sociolinguistic analysis. *Computational Linguistics and Intelligent Text Processing*, pages 385–395.

Alan Smith and Manas Gaur. 2018. What's my age?: Predicting twitter user's age using influential friend network and dbpedia. arXiv:1804.03362v1.

George Veletsianos and Royce Kimmons. 2016. Scholars in an increasingly open and digital world: How do education professors and students use twitter? *The Internet and Higher Education*, 30:1–10.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146–1151.

Anita Whiting and David Williams. 2013. Why people use social media: a uses and gratifications approach. *Qualitative Market Research*, 16:362–369.

Zach Wood-Doughty, Praateek Mahajan, and Mark Dredze. 2018. Johns hopkins or johnny-hopkins: Classifying individuals versus organizations on twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics.