Influential Text-Based Features in Predicting Admission Status of Online Degree Applicants

Farahnaz Soleimani
Georgia Institute of Technology
Atlanta, Georgia, USA
soleimani@gatech.edu

Jeonghyun Lee
Georgia Institute of Technology
Atlanta, Georgia, USA
jonnalee@gatech.edu

Meryem Yilmaz Soylu
Georgia Institute of Technology
Atlanta, Georgia, USA
msoylu6@gatech.edu

Saurabh Chatterjee
Georgia Institute of Technology
Atlanta, Georgia, USA
schatterjee47@gatech.edu

ABSTRACT

This paper presents the progress made towards developing an equitable predictive model for admission success to an online Master's program with a large pool of applicants. The overarching goal of this project is to help the future development of a systematic evaluation tool for programs with large applications. In the first phase of the project, we collected and processed data on 9,044 applications and have trained a predictive model using applicants' profile information such as demographic data, academic background, and test scores. In an ongoing phase, we seek to expand the applicants' database by incorporating the information in the letters of recommendation (LORs) and statements of purpose (SOPs) that are essential components of the application package for graduate programs and are extensively used to make decisions on granting admission. In this study, we assess various aspects of the LORs and SOPs using natural language processing to extract a comprehensive list of text features that are used to develop a classifier. We implement machine-learning algorithms such as Gradient Boosting to predict admission status and to identify the text features with the highest weight on the applicants' success. This work provides an understanding of the level of significance of a variety of text features that eventually helps the development of a comprehensive predictive model.

CCS CONCEPTS

• **Applied computing** → **Distance learning**.

KEYWORDS

Applicant success, machine-learning, admission criteria, predictive analytics

ACM Reference Format:
Farahnaz Soleimani, Jeonghyun Lee, Meryem Yilmaz Soylu, and Saurabh Chatterjee. 2022. Influential Text-Based Features in Predicting Admission Status of Online Degree Applicants. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1–3, 2022, New York City, NY, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3491140.3528318

L@S '22, June 1–3, 2022, New York City, NY, USA.

ACM ISBN 978-1-4503-9158-0/22/06.
https://doi.org/10.1145/3491140.3528318

1 INTRODUCTION

Georgia Tech's online Master's in Analytics (OMSA) which was initiated in Fall 2017 received more than 10,000 applications to date. Reviewing these application packages imposes an intense workload on the responsible faculty and staff. As a result, the institution leadership has requested predictive models that can facilitate their decision process. To this end, the current work presents the progress made as a part of an ongoing project to predict applicants' success in getting admission to the OMSA program. In [11], an initial predictive modeling test was performed by primarily using traditional application features including demographic information, academic background, and standardized test scores variables. The results suggested that Gradient Boosting performed the best. Additionally, the applicant's grade point average (GPA) played the most influential role in predicting the admission status of the applicants while the duration that the applicant spent in college ranked the second.

The current study aims to improve this modeling approach by extending the scope of relevant predictors. Specifically, we incorporate a set of text features into our model and evaluate the impact of those features on predicting the OMSA applicants' admission status. The text features are extracted specifically from the required application package including letters of recommendations (LORs) and statements of purpose (SOPs).

2 RELATED WORK

2.1 Text-based Application Features in Admissions Process

Researchers have actively examined the predictive validity of various application features that are required for admissions to academic or professional degree programs [11]. Some of the most widely studied features include GPA and scores from standardized tests such as the Scholastic Assessment Test (SAT) and the Graduate Record Examinations (GRE). A related line of research has focused on examining the extent to which a set of text-based application features would predict successful admission to academic programs. For example, some researchers conducted a comprehensive review

of the literature to compare the predictive validity of various screening measurement tools that are used in medical school admissions [9]. They found mixed evidence regarding the validity of subjective tools such as LORs or SOPs due to the issue of low inter-rater reliability, despite the value that this type of tool can bring into the admissions process.

Similarly, other researchers conducted a meta-analysis study on the predictive power of LORs in general college and graduate admissions [7]. According to their results, LORs were found to have positive but weak correlations with students' academic performances in post-secondary education. Nevertheless, the results suggested that LORs seem to serve as an incrementally valid predictor of degree attainment, corroborating the current usage of this measurement tool. Yet, extant research has primarily measured the estimated effect of the inclusion of such application features (i.e., whether an applicant submitted LOR or not) on admissions, rather than the effect of specific text features embedded in those materials. Our study aims to address this research gap by extracting a set of important text features from both LORs and SOPs through natural language processing (NLP) techniques and then building a predictive model for the successful admissions outcome.

2.2 Predicting Applicant Success

While existing educational research has primarily focused on predicting the success of currently enrolled students [8, 10], a growing body of research has recently been conducted to predict the success metrics of applicants [11]. Previous studies in this line of research have used a range of predictive analysis techniques [3], mixed-method approaches [2], and balancing weights approaches [1]. Machine-learning (ML) techniques have also been applied to measure the predictive power of application features. For example, some researchers used a deep neural network model to predict an applicant's chance of getting admitted to a graduate program [4] using their college GPA, GRE and TOEFL scores, university rating, LOR and SOP. However, only a single index score was used to represent the feature for LOR and SOP. Other researchers tested whether a set of text features extracted from SOPs could predict applicants' success in getting admissions to a computer science graduate program [6]. Nevertheless, further research is required to test if the ML-based approach would be applicable to estimate the predictive power of the combined application features consisting of both standardized and non-standardized application features such as comprehensive text features in LORs and SOPs. We aim to address this research gap by testing a predictive model that incorporates a variety of features extracted from text-based application materials.

2.3 Research Questions

In this study, we seek to address the following research questions: I) Can we predict the applicants' admission status based on the information provided in their LOR and SOP documents? II) Which aspects or information of these documents have the highest impact on being admitted to the program?

3 METHOD

The traditional narrative LORs do not typically provide uniform knowledge about the applicants due to the absence of a standardized format. In addition, there is a lack of standardized relative value given for internal grading and evaluation. People also use a variety of terminologies in their narratives which may cause different perceptions. To tackle these challenges, we process the text in the PDF documents to provide a list of standardized metrics corresponding to different characteristics of LORs and SOPs. We use these metrics as the input features to develop a predictive admission model.

3.1 Data Collection and Processing

For the current study, we collected application packages of around 1,700 applicants (for the years 2019-2021) from the Georgia Tech CollegeNET API. Later, we intend to apply the analysis framework that is presented in this work on an extended application pool (around 10,000 applicants) covering the years 2017-2021. In the initial step, we processed the PDF documents to clean and convert their narrative contents to a text format using Python's Natural Language Toolkit (NLTK) that is typically used to work with human language data.

Next, Part-of-Speech (POS) Tagging [12] was performed to grammatically tag the words in the text and also to differentiate the informative contents from the personal details. This categorization was done using Python's NLTK POS-tagger based on nouns, verbs, adjectives, adverbs, etc. Then, the text features were engineered by analyzing Text Statistics and Linguistic Attributes [5]. Text statistics quantify a variety of factors such as word count, sentence count, average word length, average sentence length, the average frequency of words, and the ratio of unique words. Besides, the quality of language used in the LORs and SOPs was measured by various Linguistic Attributes. For example, Lexical Diversity measures the richness of the text. The aforementioned text feature extraction was performed on the LORs and SOPs to produce two sets of 48 features (a total of 96 features) for each applicant (Table 1).

3.2 Gradient-Boosted Classifier

We initially conducted a pilot study to test the suitability of a set of candidate ML models using (a) support vector classifier (SVC); (b) gradient boosting; (c) k-nearest neighbors; (d) Random Forest; and (e) Multi-Layer Perceptron (MLP). According to the comparison of the statistical performance of the generated models, the gradient boosting classifier outperformed the other ML algorithms.

Tree-based approaches provide simple, interpretable ML-based models for classification problems. Gradient boosting [11] is such an approach that develops a classifier in the form of an ensemble of multiple learners in which it sequentially builds a series of relatively simple models to compensate the errors made by each preceding tree. The collection of the created models contributes to the output of the model and the final prediction score. Since we plan to classify the applicants based on their admission status (admitted vs not-admitted) a binary classifier model was developed. Using the constructed boosted trees, the importance score for each attribute was retrieved.

Table 1: Description of the extracted text features.

Text-based Features	*Description*
NNP	Proper noun, singular
VBZ	Verb, 3rd person singular
VBN	Verb, past participle
DT	Determiner
NN	Noun, singular or mass
IN	Preposition or subordinating conjunction
CD	Cardinal Number
NNS	Noun, plural
CC	Coordinating conjunction
VB	Verb, base form
JJ	Adjective
VBP	Verb, non-3rd person singular present
TO	to
RB	Adverb
PRP$	Possessive pronoun
PRP	Personal Pronoun
WP	Wh-pronoun
MD	Modal
VBG	Verb, gerund or present participle
VBD	Verb, past tense
RP	Particle
WDT	Wh-determiner
JJS	Adjective, superlative
POS	Possessive ending
RBR	Adverb, comparative
WRB	Wh-adverb
RBS	Adverb, superlative
NNPS	Proper noun, plural
PDT	Predeterminer
JJR	Adjective, comparative
EX	Existential there
FW	Foreign word
SYM	Symbol
UH	Interjection
LS	List item maker
WP$	Possessive wh-pronoun
Word_count	Word count
Sentence_count	Sentence count
Avg_word_length	Average word length
Avg_sentence_length	Average sentence length
Avg_word_frequency	Average frequency of word
Unique_word_ratio	Ratio of unique words
RI_fre	Readability Index (Flesch Formula)
RI_gf	Readability Index (Gunning Fog Index)
RI_si	Readability Index (Smog Index)
RI_dcrs	Readability Index (Dale-Chall)
LD_yules_i	Lexical Diversity (Yule's I)
LD_yules_k	Lexical Diversity (Yule's K)

4 RESULTS

Based on the results from the pilot study, classification-based modeling was performed using Gradient Boosting provided by Python's scikit-learn library to predict applicants' admission status. To develop a predictive model using the extracted text-based features, the data was split into the training and testing subsets using scikit-learn's train_test_split library which partitions the data in a stratified fashion. 75% of the entire data was used to train the classification model, while the remaining 25% was used to evaluate the model's performance on unseen data. Within the training subset, grid search was performed using Cohen's Kappa as the measure to derive the best hyperparameters for Gradient Boosting. The tuned model with the best set of hyperparameters was used to run on the testing subset to report the final predictive scores. To evaluate the classifier performance, a 10-fold cross-validation approach was used in which the training subset was randomly divided into 10 equal-sized subsets. The classification model was trained using 9 subsets, while it was validated on the remaining subset. This process was repeated 10 times and the average results were considered as the overall model's performance to prevent over-fitting.

According to the evaluation results, the Gradient Boosting approach using 100 estimators with a learning rate of 0.1, maximum depth of 4, minimum samples leaf of 6, and a deviance loss function exhibited the best cross-validation score (0.4) and training score (0.95). Using these hyperparameters, a classification model was developed and tested on the test subset. Figure 1 displays the confusion matrix. Moreover, Gradient Boosting measured the relative importance of each considered feature in predicting the applicants' admission. The 15 top-ranked significant features are presented in Figure 2. According to the comparison of the measure of importance, LOR unique word ratio and Linguistic Attributes such as Readability Index and Lexical Diversity as well as some of the Text Statistic related features of SOPs were indicated as the highest-ranked features.

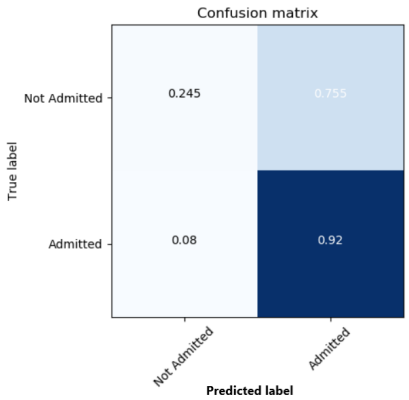

Figure 1: Gradient Boosting Matrix

5 CONCLUSION AND FUTURE WORK

As part of a more extensive study, in this paper, we reported on progress made towards a systemic evaluation of applicants to Georgia Tech's OMSA program. In addition to our former effort to identify variables to construct predictive modeling, through NLP and ML, we predicted the admission status of the applicants and identified the text-based parameters with the most significant impact on getting admission.

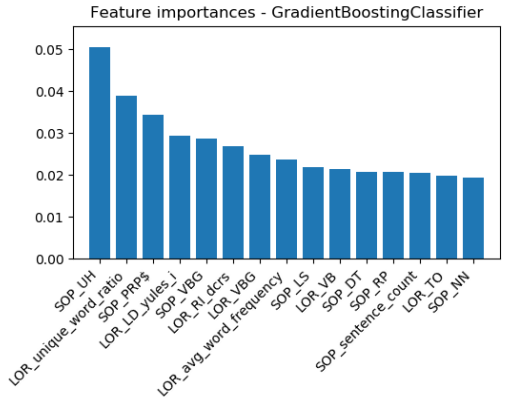

Figure 2: The level of importance of features calculated using Gradient Boosting algorithm

Our primary analyses suggested Gradient Boosting as a suitable ML algorithm to develop the predictive classifier. Moreover, the comparison results of the features' level of importance in developing the predictive model indicated that among the LOR-related features, the unique word ratio along with the Linguistic Attributes, including Lexical Diversity index and Readability Index, were encountered the most influential parameters in predicting admission success. However, the top-ranked SOP-related features belonged to Text Statistics.

In the next steps of the project, we plan to expand the scope of our dataset and feature processing. Furthermore, our predictive model will incorporate additional features, including but not limited to applicants' demographic information, educational background, English proficiency, GPA, and work experience. Additionally, we plan to use visual analytic tools to detect and alert users when their analysis behavior is biased. Ultimately, we hope that our research will offer helpful guidance for the OMSA program's admission process and help administrators make informed decisions contributing to program improvement.

REFERENCES

[1] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. 2020. Variation in impacts of letters of recommendation on college admissions decisions: Approximate balancing weights for treatment effect heterogeneity in observational studies. (2020).

[2] Chalee Engelhard, Rebecca Leugers, and Jenna Stephan. 2016. Effectiveness of pre-admission data and letters of recommendation to predict students who will need professional behavior intervention during clinical rotations in the United States. *Journal of educational evaluation for health professions* 13 (2016).

[3] Daniel V Girzadas, Robert C Harwood, Steve N Delis, Kathleen Stevison, George Keng, Nancy Cipparrone, Andrea Carlson, and George D Tsonis. 2001. Emergency medicine standardized letter of recommendation: predictors of guaranteed match. *Academic Emergency Medicine* 8, 6 (2001), 648–653.

[4] Md Omaer Faruq Goni, Abdul Matin, Tonmoy Hasan, Md Abu Ismail Siddique, Oishi Jyoti, and Fahim MD Sifnatul Hasnain. 2020. Graduate admission chance prediction using deep neural network. In *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 259–262.

[5] Michael Alexander Kirkwood Halliday. 2006. *Linguistic studies of text and discourse*. Vol. 2. A&C Black.

[6] Diptesh Kanojia, Nikhil Wani, and Pushpak Bhattacharyya. 2018. Is your statement purposeless? predicting computer science graduation admission acceptance based on statement of purpose. *arXiv preprint arXiv:1810.04502* (2018).

[7] Nathan R Kuncel, Rachael J Kochevar, and Deniz S Ones. 2014. A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment* 22, 1 (2014), 101–107.

[8] Jeonghyun Lee, Farahnaz Soleimani, John Hosmer IV, Meryem Yilmaz Soylu, Roy Finkelberg, and Saurabh Chatterjee. 2022. Predicting Cognitive Presence in At-Scale Online Learning: MOOC and For-Credit Online Course Environments. *Online Learning* 26, 1 (2022).

[9] Eric Siu and Harold I Reiter. 2009. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. *Advances in Health Sciences Education* 14, 5 (2009), 759.

[10] Farahnaz Soleimani and Jeonghyun Lee. 2021. Comparative Analysis of the Feature Extraction Approaches for Predicting Learners Progress in Online Courses: MicroMasters Credential versus Traditional MOOCs. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 151–159.

[11] Shawn Staudaher, Jeonghyun Lee, and Farahnaz Soleimani. 2020. Predicting Applicant Admission Status for Georgia Tech's Online Master's in Analytics Program. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 309–312.

[12] Atro Voutilainen. 2003. Part-of-speech tagging. *The Oxford handbook of computational linguistics* (2003), 219–232.