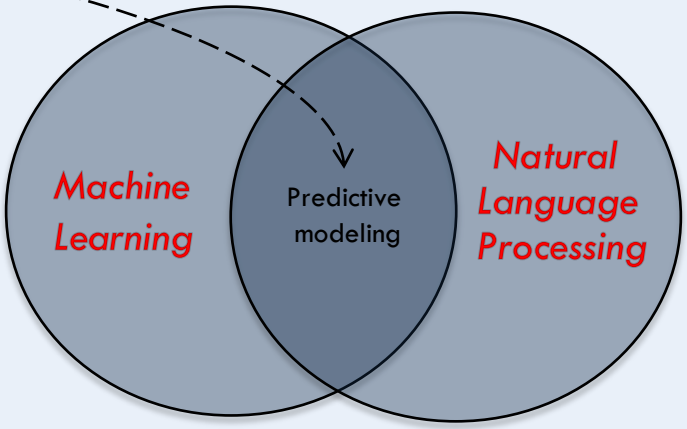
Research Statement

Is your Statement Purposeless?

Predicting Computer Science Graduation Admission Acceptance based on Statement Of Purpose



Authors: Diptesh Kanojia, ***Nikhil Wani***, and Dr. Pushpak Bhattacharyya

Center for Indian Language Technology, IIT Bombay, India

About Me

Currently: **Research Fellow at Indian Institute of Technology, Bombay(IITB)**

Machine learning and NLP track, CFILT, CSE Dept.

Past: **Nvidia Corporation**

Research intern, AI and Deep Learning track.

Education: Major in Computer Science, **Fresh B.Tech Graduate!** ☺

Motivation - Why?

1. Mitigate the problem of ***unpredictability*** in CS graduate admissions to Elite Universities – MIT, Stanford, Harvard, University of California, Berkley etc (Acceptance rate < 10%).

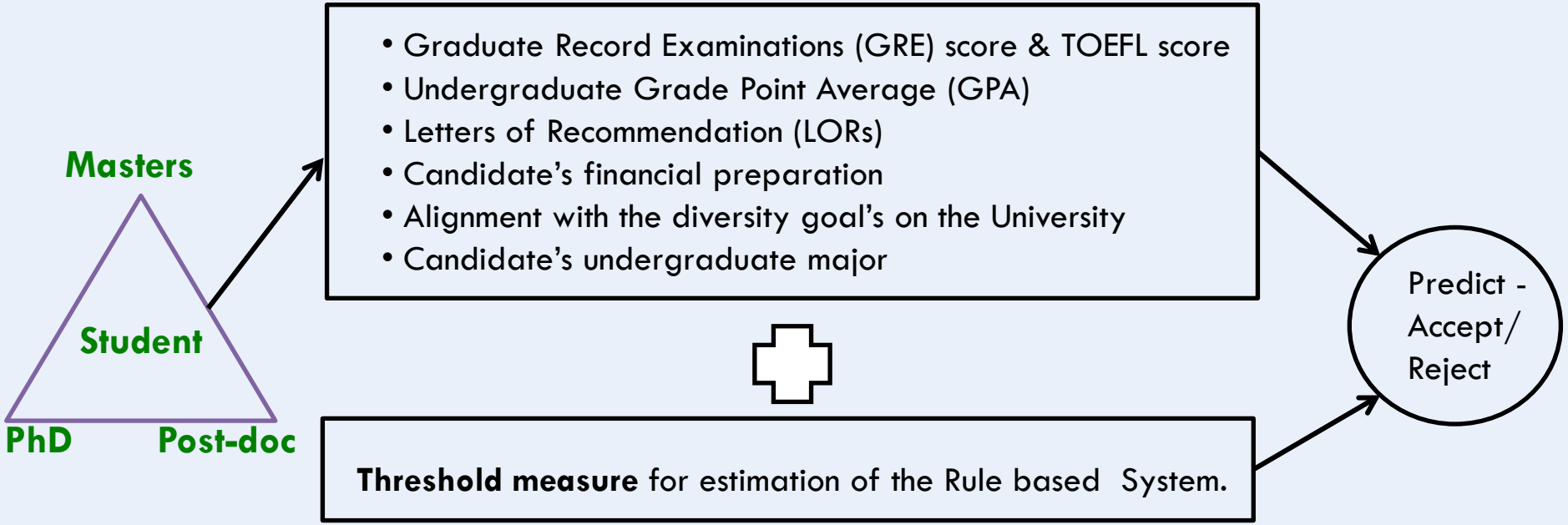
2. Propose ***Novel Solution*** – Traditional solution are qualitative in nature, while ours is quantitative based on Machine learning and Deep learning.

3. Bolster application of ***Document Similarity.***

4. Build perspective for my own graduate application! ☺

Literature Survey

1. ***Ward (2006)*** discusses a rating based qualitative model with identified factors :



Literature Survey

2. ***Raghunathan (2010),*** having been a member of the admission's committee at Stanford, identifies that SOP is one of **"the trickiest component"** of an overall application.
    - He notes that **verbose** SOPs deter the chances of candidate's selection.

3. Text Similarity and related measures have been extensively studied – ***Choi et al. (2010); Adomavicius and Tuzhilin, (2005); Gomaa and Fahmy, (2013***), **BUT**
    - To the best of our knowledge, there is no reported study which evaluates SOPs based on the features identified by us, **OR**
    - Use ML and DL based techniques of this kind, at the time of submission.

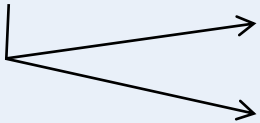
Dataset Creation

Sources:

1. Acquaintances from IITB.
2. Publicly disclosed SOPs from personal websites.
3. Admission consultancy blogs.

Total:

- Modest size of 50 manually verified SOPs.
- Each containing about 1000-1500 words.
- Equally spit



Accepted SOPs to Elite University
(acceptance rate <15%)

Rejected SOPs

1. Had EXTENSIVE spelling mistake.
2. Too short
3. Other minor details (next few slides)

- All SOPs were written in English.

Methodology

(For Machine Learning experiments)

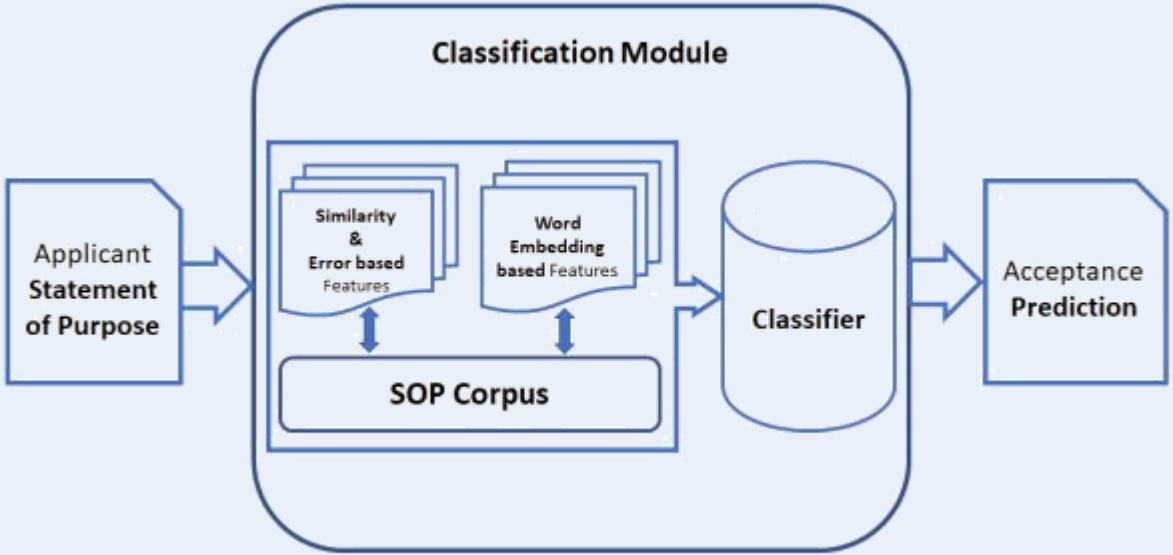
1. **Data Preprocessing Pipeline** – Tokenization, Stemming and Stopping.
2. **Feature Identification.**
    1. Textual Features
    2. Word Embedding based Features
    3. Similarity Score based Features
    4. Error based Features
3. **Binary Classification** using SVM, Logistic Regression and Random Forest Decision Tress.
4. **Ablation Test** for best feature set combination identification.
5. Reporting **Precision, Recall and F-Score.**

Methodology

(For Deep Learning experiments)

1. **Multi Layer Perceptron and Feed Forward Neural Network**
    1. Input layer which takes in each word in the corpus.
    2. One hidden layer with 100 neuron, simple matrix multiplication operation.
    3. Output layer with a single neuron for binary classification

System Architecture



Input - PDF/.docx file, which gets converted to a textbased corpus.

Classification Module - Calculates feature values for features.

Output - Predicts an accept or reject based on the classification model.

Features Used

A. Textual Features:

1. **PoS Ratios** - Ratio of nouns, adjectives, adverbs, and verbs to the entire text, obtained using NLTK7 (Loper and Bird, 2002).
2. **Discourse Connectors** - It is the number of discourse connectors in the essay computed using a list of discourse connectors.
3. **Count of Named Entities** - Number of named entities in the essay. We tried using this as a feature but this drastically lowered the F-scores, and had to be avoided in the final set of reported experiments.
4. **Readability** - The Flesch Reading Ease Score (FRES) of the text (Flesch, 1948).
5. **Length features** - Number of words in the sentence, number of words in the paragraph, and average word length.
6. **Coreference Distance** - Sum of token distance between co-referring mentions.
7. **Degree of Polysemy** - Average number of WordNet (Fellbaum, 2010) senses per word.

Features Used

B. Document Similarity Score and Error based Features:

1. **Cosine Similarity -** Cosine Similarity Score of an SOP with the corpus of accepted essays dataset, where we ensure that the SOP being compared is not a part of the accepted essay corpus.
2. **Similarity-based features using GloVe -** The similarity between every pair of content words in adjacent sentences. The similarity is computed as the cosine similarity between their word vectors from the pre-trained GloVe word embeddings (Pennington et al., 2014). We calculate the mean and maximum similarity.
3. **Spell Check Errors** *(count)* – We use PyEnchant9 to embed a spell checker and count the number of errors in each document.
4. **Out of Vocabulary Words** *(count)* – We use the pretrained Google news word embeddings and find out word vectors for every token in the document. The tokens which do not return any vector are either rare words or in all probability out of vocabulary words.

Features Used

C. Word Embeddings based Feature:

1. **Average Word Vector Scores** - Average of word vectors of each word in the statement calculated using pre-trained Google News word vectors (Mikolov et al., 2013).

Evaluation Metrics – Precision, Recall and F-Score

Classifier	P_{acc}	P_{rej}	$\mathbf{P_{avg}}$	R_{acc}	R_{rej}	$\mathbf{R_{avg}}$	F_{acc}	F_{rej}	$\mathbf{F_{avg}}$
RFDT	0.86	0.79	0.83	0.76	0.88	0.82	0.81	0.83	0.82
LR	0.69	0.83	0.76	0.88	0.60	0.74	0.77	0.70	0.74
SVM	0.89	0.96	**0.92**	0.96	0.88	**0.92**	0.92	0.92	**0.92**
Neural Network Based									
Multilayer Perceptron (Train-Test Split)	-	-	0.82	-	-	0.82	-	-	0.82
Feed Forward NN (FFNN) (Train-Tune-Test Split)	-	-	0.36	-	-	0.60	-	-	0.45

SVM with 92% F-score outperforms :

1. Random Forest Decision Trees (RFDT) with a margin of 9%
2. Logistic Regression (LR) with a margin of 18%

Ablation Test

(A total of 317 features were ablated)

	Individual Feature Sets (N-fold)			
Features	2-F	5-F	10-F	50% Split
T [14]	54	46	44	40
WE [300]	48	78	40	44
SE [3]	48	56	56	49
	Combination of Feature Sets			
T + WE [314]	56	62	62	52
T + SE [17]	48	50	38	30
SE + WE [303]	**90**	**92**	**92**	**92**
T + WE + SE [318]	52	50	53	43

While Word Embedding features independently do not contribute significantly, but when combined with Similarity Score and Error Based (SE) feature set form our best reported model forms our final model.

Note - [X] in the table indicates X number of dimensions of that feature combination

Future Scope

1. Integrate Parts-of-speech (POS) based similarity measures.

2. Experiment with Recurrent Neural Networks (RNNs) which have been shown to work well with textual data.

3. An open source web application which would allow prospective applicants to evaluate their SOPs with our system.

Question?

Thank you!