Improve the predictability of SmartFund

Yiwei Dong
ydong2@nd.edu
Supervisor: Meng Jiang, Qingkai Zeng
{mjiang2, qzeng}@nd.edu

June 2020

Introduction

- This project aims at improving the predictability of SmartFund, which is a model used for research outcomes prediction developed by Alvin Alaphat and Dr. Jiang. SmartFund leverages the past project data (award amount, investigators, university, and abstract) from the National Science Foundation (NSF) website, to predict the number of papers and corresponding citations a new project might produce, and thus it is expected to help NSF fund projects efficiently and cost-effectively. However, the performance of SmartFund model is not so satisfying, with coefficient of determination 0.348 for #papers prediction and 0.188 for #citations prediction. Therefore, it is important to enhance the predictability of SmartFund to make it useful in the real sense.

Motivitions

- Textual features can be better extracted by SciBERT model

Textual features hidden in the abstract can be extracted more amply by SciBERT, the state-of-the-art natural language processing model.

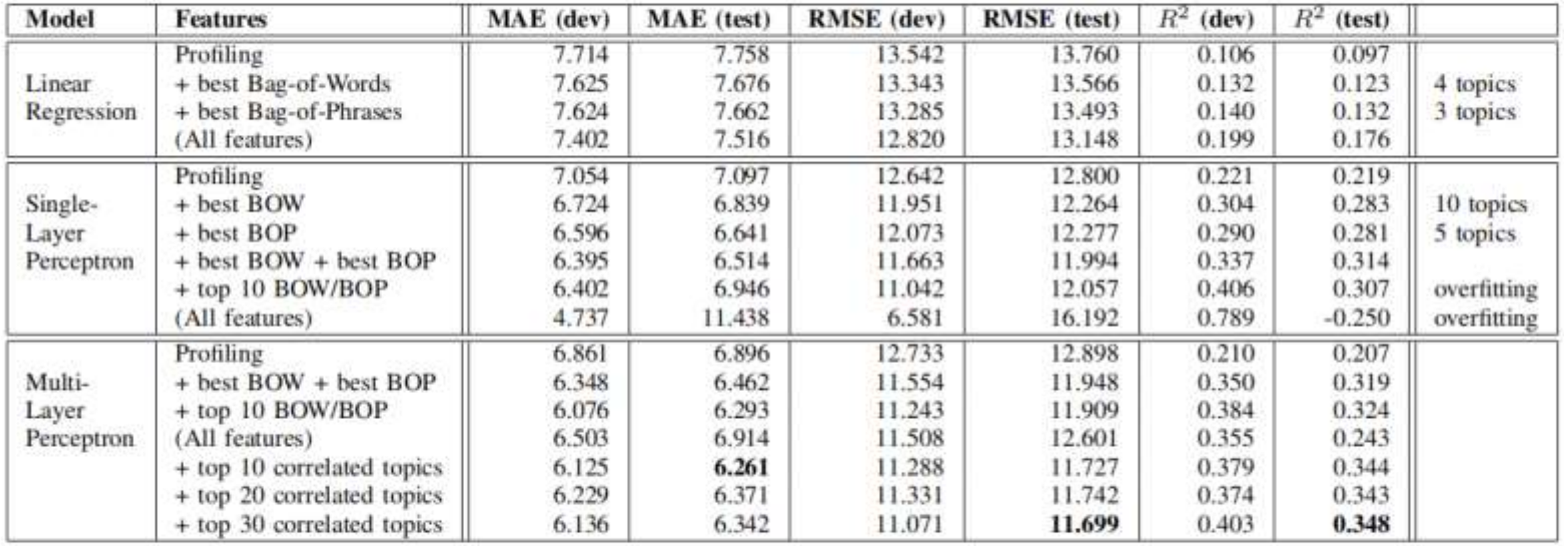
Work of Alvin Alaphat and Dr. Jiang: LDA + regression Models

- Extract data from NSF[1] and Open Academic Data[2]
- Feature extraction
    - Bag-of-Words model
    - Latent Dirichlet Allocation (LDA)
    - Term Frequency-Inverse Document Frequency (TF-IDF)
    - AutoPhrase model: upgrading vocabulary from words only to words and phrases
- Feature selection
- Regression models

[1] NSF Award Search: https://www.nsf.gov/awardsearch/download.jsp

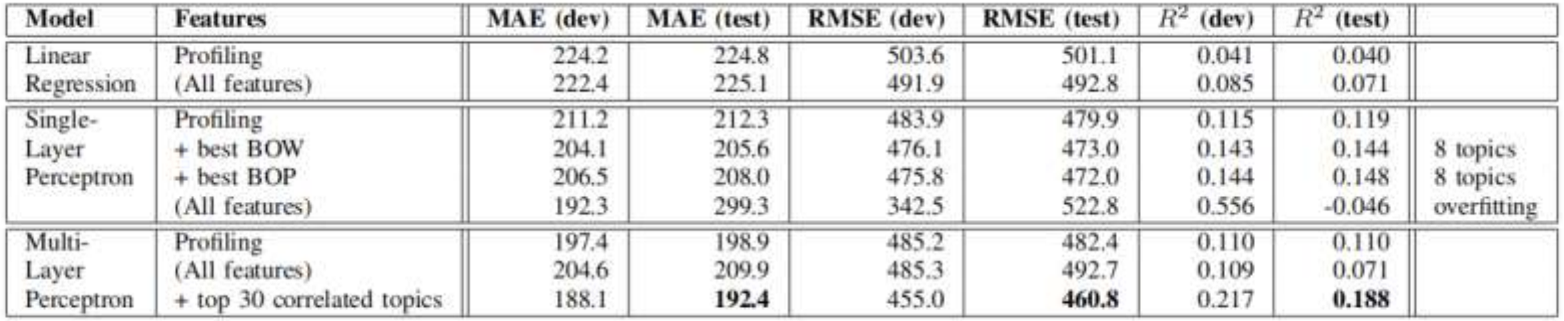
[2] Open Academic Graph: https://www.openacademic.ai/oag/

Results of Topic Modeling + Regression Models



Model	Features	MAE (dev)	MAE (test)	RMSE (dev)	RMSE (test)	R^2 (dev)	R^2 (test)	
Linear Regression	Profiling	7.714	7.758	13.542	13.760	0.106	0.097	
	+ best Bag-of-Words	7.625	7.676	13.343	13.566	0.132	0.123	4 topics
	+ best Bag-of-Phrases	7.624	7.662	13.285	13.493	0.140	0.132	3 topics
	(All features)	7.402	7.516	12.820	13.148	0.199	0.176	
Single-Layer Perceptron	Profiling	7.054	7.097	12.642	12.800	0.221	0.219	
	+ best BOW	6.724	6.839	11.951	12.264	0.304	0.283	10 topics
	+ best BOP	6.596	6.641	12.073	12.277	0.290	0.281	5 topics
	+ best BOW + best BOP	6.395	6.514	11.663	11.994	0.337	0.314	
	+ top 10 BOW/BOP	6.402	6.946	11.042	12.057	0.406	0.307	overfitting
	(All features)	4.737	11.438	6.581	16.192	0.789	-0.250	overfitting
Multi-Layer Perceptron	Profiling	6.861	6.896	12.733	12.898	0.210	0.207	
	+ best BOW + best BOP	6.348	6.462	11.554	11.948	0.350	0.319	
	+ top 10 BOW/BOP	6.076	6.293	11.243	11.909	0.384	0.324	
	(All features)	6.503	6.914	11.508	12.601	0.355	0.243	
	+ top 10 correlated topics	6.125	**6.261**	11.288	11.727	0.379	0.344	
	+ top 20 correlated topics	6.229	6.371	11.331	11.742	0.374	0.343	
	+ top 30 correlated topics	6.136	6.342	11.071	**11.699**	0.403	**0.348**	

Table 1: Performance on predicting the number of papers produced by an unseen project[1]



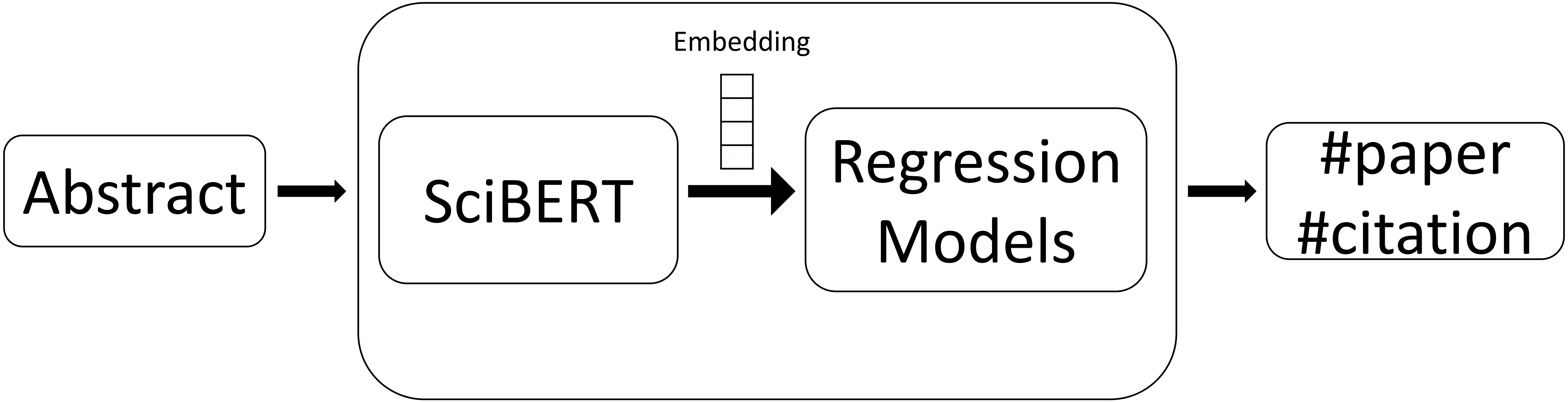
Model	Features	MAE (dev)	MAE (test)	RMSE (dev)	RMSE (test)	R^2 (dev)	R^2 (test)	
Linear Regression	Profiling	224.2	224.8	503.6	501.1	0.041	0.040	
	(All features)	222.4	225.1	491.9	492.8	0.085	0.071	
Single-Layer Perceptron	Profiling	211.2	212.3	483.9	479.9	0.115	0.119	
	+ best BOW	204.1	205.6	476.1	473.0	0.143	0.144	8 topics
	+ best BOP	206.5	208.0	475.8	472.0	0.144	0.148	8 topics
	(All features)	192.3	299.3	342.5	522.8	0.556	-0.046	overfitting
Multi-Layer Perceptron	Profiling	197.4	198.9	485.2	482.4	0.110	0.110	
	(All features)	204.6	209.9	485.3	492.7	0.109	0.071	
	+ top 30 correlated topics	188.1	**192.4**	455.0	**460.8**	0.217	**0.188**	

Table 2: Performance on predicting the number of citations the produced papers can have[2]

[1,2] from Alvin and Dr. Jiang

SciBERT Model + Regression Models

- Using SciBERT Model to get sentence embeddings of abstracts
- Feed the sentence embeddings to regression models



Data Processing

- Delete items with none abstract
- Tokenize abstract & Cut to 490 tokens (BERT model can tackle at most 512 tokens at one time. Here, we regard each abstract as one sentence, and take the sentence embedding as its feature)
- Split train, dev, and test

Use SciBERT to Embed All The Abstracts

ID	Abstract	#paper	#citation
8854199	This is a workshop on the use of mathematical and computer models in biological resource conservation. It is designed to...	0	0
9012033	The distribution of Ba in the ocean is similar to the refractory components, silica and alkalinity. Therefore reconstructions of Ba in ancient...	1	90
1536005	The research objective of this project is to investigate how performance can be made comparable across a wide range of construction project sizes and types. Performance...	0	0
	...	...	...

➡

Sentence embeddings			
0	1	...	767
-0.6507	0.0876	...	-0.8717
-0.6902	-0.3954	...	-0.8150
0.1398	-0.3006	...	-0.5552
...	...	...	...

Feedforward Neural Network

- Pass the obtained sentence embedding as the input to the feedforward neural network
- Train & Save the model with the least MSE on validation set

Future Work

- Optimize the SciBERT + Feedforward neural network model
- Test the reasonability of each step(e.g. If it is reasonable to cut each long abstract to 490 words? If it is better to cut those abstracts backward so that we can get a better result?)

Conclusion

In this project:

- We propose to improve the predictability of SmartFund, to make it function better
- We propose to use SciBERT to extract the information contained in abstracts

Feel free to contact me at ydong2@nd.edu should you have any questions!