“Grantformatics”: An Automated Text Classifier System to Improve Research Grant Funding Opportunities

Alex Bokov, Sergio Duncan, Alfredo Tirado-Ramos

Clinical Informatics Research Division, School of Medicine

The University of Texas Health Science Center at San Antonio

San Antonio, Texas, United States of America

Abstract

In present times, where tight budgets have become a common occurrence, researchers and institutions rely heavily on funding from exterior agencies. This process typically involves utilizing an excessive amount of time and financial resources in the decision-making process that defines which opportunity will best suit their prospective research ideas and potential collaborations, which in theory may increase their chances of being funded. This paper intends to provide an innovative process by which aspiring applicants can improve the probabilities for more efficiently gaining sponsorship from funding agencies. Our model uses supervised machine learning and natural language processing methods that assign proposal abstracts to the most appropriate funding announcement (FA), consequently answering questions that would assist in further effectively targeting applications. The system utilizes a unique feature extractor carefully built to avoid reliance on idiosyncrasies of our training data in order to facilitate the incorporation of new data sets. An implementation of the system is discussed, using the Patient-Centered Outcomes Research Institute’s (PCORI) PFA mechanism as case study, and initial results are presented.

Introduction

Recent surges in the availability of mass amounts of medical data have catalyzed an opportunity for medical and scientific discovery that has not been available until today. Consequently, scientific funding agencies have acknowledged the untapped utility that these data sets could potentially have in performing large-scale disparity research and are providing researchers with the economic resources to fund their attempts to advance today’s quality of healthcare. 5 This paper does not suggest re-inventing the state of the art, but rather proposes a new application for text mining that researchers can use in order to improve their chances of achieving financial backing from these agencies.

For purposes of this paper and personal objectivity, we will be applying our model specifically for analyzing and predicting from abstracts obtained from the open repository of previously awarded abstracts from the Patient-Centered Outcomes Research Institute (PCORI) 1It should be noted that the model is not limited to PCORI, but rather applicable to any other funding agency that allows access to previously awarded text data permitting a retrospective study. The majority of PCORI applicants submit a Letter of Intent (LOI) to one of five priority categories that correspond to the agency’s funding agenda also known as the PCORI Funding Announcements (PFA) 2 Applicants can only apply to one of the five PFAs, which creates an utmost importance of deciding which one allows for the best chances of being approved. 3

The data we use for this model consist of 281 project abstracts from taken from each of the five PFAs previously funded LOIs tracing back to 2012. Our module intends to identify and classify any clearly defined clusters of similarities within each abstract. The model’s feature extractor also located words and phrases that most strongly distinguished grants that were funded by each of the PFAs by trying out various tokenization and summarization schemes to break each PFA down by common topic areas. We aim to provide an insight into the rhetorical aspect of the application process and specifically to offer a solution to three general questions:

1. How to choose between PFAs when a project looks like it is applicable to more than one?

2. How to discover hidden features and rhetorical elements that improve the success of a grant for each PFA?

3. What gaps exist in PFAs funded in the past (e.g. non-redundancy)?

State of the Art

Data scientists are often tasked with mining text datasets for rhetorical elements and contextual indicators in the hopes of correctly identifying and classifying documents. More archaically, the standard solution for most was outsourcing the tedious process to either graduate-student labor or expensive consultants who then manual code, or hand-label the corpora and using human judgment, make inferences about the data. This presented an obviously inefficient and time-consuming problem that is now being facilitated with the help of natural language processing and text mining techniques. 5

Machine learning has expanded the field of text analytics beyond human capabilities allowing researchers to systematically mine data from multiple documents without the pain-staking task of manual coding. 15 It should also be noted that although human interpretation is less time-efficient, it also presents a benefit of accuracy as a manual coder will be able to identify features of the text possibly overlooked by an automatic feature extractor. That being said, researchers have heavily favored the efficiency of accessing larger datasets and superior coding scheme performance to that of a more precise interpretation that usually only includes tokens that have minimal influence on classification. 13,15

For those who have already began implementing text mining optimization techniques, the state of the art generally proposes to identify classification systems based on their distinct type of text-timing task. The generalization is attributed to several house-hold named text mining categories like text classification and named entity recognition.

Text mining provides an untapped source of data and opportunity for research that with recent additions in text-mining toolkits like R’s RTextTools or Python’s NLTK are openly available for exploration. Most text-mining that is performed by data scientists aims to answer a scientific question related to their research. Our approach was not to invent new software or change pre-existing packages but provide researchers with an innovative and practical example of using text mining tools to improve personal infrastructure and grant development.

Methods

[add intro text]

Obtaining and Prepping Data

We began by locating and accessing the data from PCORI’s website. Once determined that there is sufficient data for classification, we proceed to extracting it for pre-processing and formatting purposes. With the help of a batch script, the scraped web data is cleaned of unwanted HTML tags and other mark-up then formatted to reflect one proposal abstract per file within its awarded PFA category.

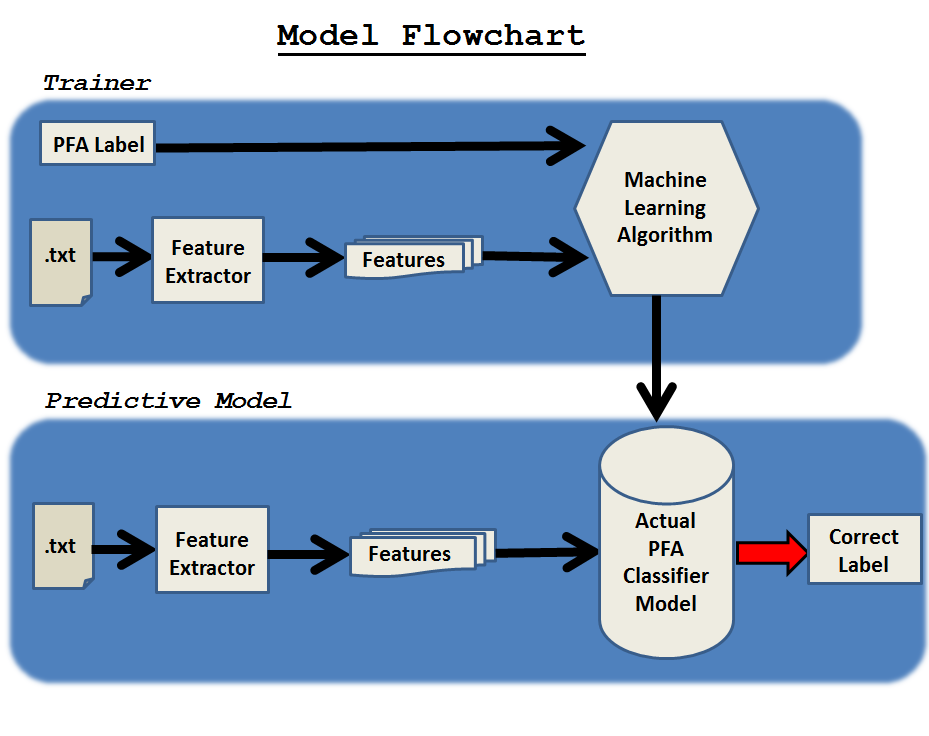
Build Matrix

We then create a function that will run the new corpus through R’s tm module to remove all stop words, punctuation, whitespace, etc. The next step is to build the document matrix which allows for the text data to be converted into quantitative form and concatenates the successfully award abstract name to its PFA category. A TDM variable is then created to represent terms on one axis and the name of PFA awarded on another. (Returns result as list where name of abstract is 1, tdm 2 and output is the result.

[add pseudo-code]

Hold-out Trainer

After verifying that the outputted matrices correlate accurately with desired file ids, the matrix’s content is then fed into the machine learning algorithm consisting of a hold-out trainer and actual classification model. The hold-out trainer interprets each input in order to teach the subsets by withholding the appropriately labeled PFA, and comparing output to actual results. The next function, almost identical to the previous one, accepts the input from the trainer and returns the abstract accurately classified to its awarded PFA.

****

**Figure 1.** General overview of supervised text classification system. [explain the flow in the picture]

Results

Finally, we have to interpret our model’s accuracy and classification results.

Accuracy

# Clusters

#Results

1. How to choose between PFAs when a project looks like it is applicable to more than one?

2. Discover hidden features and rhetorical elements that improve the success of a grant for each PFA?(i.e. things those reviewers favor even though PCORI hasn't told them to do so)?

3. What gaps exist in what each PFA has funded in the past (PCORI wants non-redundant topics)?

Discussion

We built a model that uses supervised machine learning and natural language processing methods to aid in assigning proposal abstracts to the most appropriate funding announcement. Our model proves, in this implementation, the successful incorporation of innovative text mining techniques to advancing research efforts. By using minimal parameters and original HTML mark-up, our test implementation provided consistent 60-63% accuracy when asked to classify historical abstracts with their appropriate PFAs. Improving the accuracy of the model is a matter of machine learning trial and error-based training, by applying an array of tokenization and parsing methods, all the while remaining cautious of overfitting the feature extractor and limiting the model to only predicting from our specific agency’s data sets. For future work, the model will continue to be trained with alternative data sets increasing accuracy and further enabling general application. For this purpose we plan to begin implementing data sets from a multitude of funding agencies other than PCORI. For the next version of the implementation we will create an application that allows researchers to submit an abstract and directs them to the funding agency announcement that might provide the best probabilities of approval and ultimately successful funding. As with the majority of machine learning systems, there is no golden ticket that allows for flawlessly accurate and out-of-the-box models for the specific data researchers plan to mine. Consequently, researchers should expect to have to implement their own feature extractor parameters in order to tailor the accuracy of the model accordingly.

**References**

1. About Us | PCORI [Internet]. [cited 2015 Mar 12]. Available from: http://www.pcori.org/about-us
2. National Priorities and Research Agenda | PCORI [Internet]. [cited 2015 Mar 12]. Available from: http://www.pcori.org/content/national-priorities-and-research-agenda
3. FAQs for Applicants | PCORI [Internet]. [cited 2015 Mar 12]. Available from: http://www.pcori.org/content/faqs-applicants
4. tm.pdf [Internet]. [cited 2015 Mar 12]. Available from: http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
5. Jurka TP, Collingwood L, Boydstun AE, Grossman E, van Atteveldt W. Rtexttools: A supervised learning package for text classification [Internet]. 2011 [cited 2015 Mar 12]. Available from: http://rjournal.github.io/archive/2013-1/collingwood-jurka-boydstun-etal.pdf
6. TextMining\_L\_JDS\_Jan2014.pdf [Internet]. [cited 2015 Mar 12]. Available from: http://www.unt.edu/rss/class/Jon/Benchmarks/TextMining\_L\_JDS\_Jan2014.pdf
7. Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. Journal of Statistical Software. 2008;25(5):1–54.
8. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005 Mar 1;6(1):57–71.
9. Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. Journal of Statistical Software. 2008;25(5):1–54.
10. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in. Bioinformatics. 2007 Oct 1;23(19):2507–17.
11. Moschitti A, Basili R. Complex Linguistic Features for Text Classification: A Comprehensive Study. In: McDonald S, Tait J, editors. Advances in Information Retrieval [Internet]. Springer Berlin Heidelberg; 2004 [cited 2015 Mar 13]. p. 181–96. Available from: http://link.springer.com/chapter/10.1007/978-3-540-24752-4\_14
12. Forsyth RS, Holmes DI. Feature-Finding for Text Classification. Lit Linguist Computing. 1996 Dec 1;11(4):163–74.
13. Kosala R, Blockeel H. Web Mining Research: A Survey. SIGKDD Explor Newsl. 2000 Jun;2(1):1–15.
14. Namburu SM, Tu H, Luo J, Pattipati KR. Experiments on Supervised Learning Algorithms for Text Categorization. 2005 IEEE Aerospace Conference. 2005. p. 1–8.
15. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Comput Surv. 2002 Mar;34(1):1–47