Journal of Soft Computing in Civil Engineering 1-2 (2017) 89-102

Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Predicting Budget from Transportation Research Grant Description: An Exploratory Analysis of Text Mining and Machine Learning Techniques

A. Singhal[1], K. Gopalakrishnan[2*], S.K. Khaitan[3]

1. R&D, Contata Solutions, LLC, Minneapolis, Minnesota, USA
2. Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA
3. Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA

Corresponding author: *rangan@northwestern.edu*

https://doi.org/10.22115/SCCE.2017.49604

ARTICLE INFO

Article history:
Received: 12 August 2017
Revised: 25 August 2017
Accepted: 25 August 2017

Keywords:
Text mining;
Transportation research;
Natural Language Processing (NLP);
Big data;
Deep learning;
Statistical analysis;
Soft computing.

ABSTRACT

Funding agencies such as the U.S. National Science Foundation (NSF), U.S. National Institutes of Health (NIH), and the Transportation Research Board (TRB) of The National Academies make their online grant databases publicly available which document a variety of information on grants that have been funded over the past few decades. In this paper, based on a quantitative analysis of the TRB's Research In Progress (RIP) online database, we explore the feasibility of automatically estimating the appropriate funding level, given the textual description of a transportation research project. We use statistical Text Mining (TM) and Machine Learning (ML) technologies to build this model using the 14,000 or more records of the TRB's RIP research grants big data. Several Natural Language Processing (NLP) based text representation models such as the Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI) and the Doc2Vec Machine Learning (ML) approach are used to vectorize the project descriptions and generate semantic vectors. Each of these representations is then used to train supervised regression models such as Random Forest (RF) regression. Out of the three latent feature generation models, we found LDA gives the least Mean Absolute Error (MAE) using 300 feature dimensions and RF regression model. However, based on the correlation coefficients, it was found that it is not very feasible to accurately predict the funding level directly from the unstructured project abstract, given the large variations in source agencies, subject areas, and funding levels. By using separate prediction models for different types of funding agencies, funding levels were better correlated with the project abstract.

How to cite this article: Singhal A, Gopalakrishnan K, Khaitan SK. Predicting budget from transportation research grant description: an exploratory analysis of text mining and machine learning techniques. J Soft Comput Civ Eng 2017;1(2):89–102. https://doi.org/10.22115/scce.2017.49604.

1. Introduction

A key responsibility of doctoral students, researchers and university faculty members is to write proposals to secure sponsored research grants actively. In targeting potential funding sources, research proposal writers often look to recent funding levels and patterns over time for a given topic of research. In fact, it is often customary to work backward from the budget through the abstract while working on final proposal efforts. Thus, it would be advantageous to estimate the appropriate funding level for a given research topic submitted to a funding agency based on the historical trends and patterns in research funding levels. This could help researchers roughly determine the level of funding they can expect for their research topic of interest and appropriately adjust the scope of their research plan. From the agency's perspective, this can be beneficial in planning budget allocation levels distributed across its research portfolio. Fortunately, Funding agencies such as the U.S. National Science Foundation (NSF), U.S. National Institutes of Health (NIH), and the Transportation Research Board (TRB) of The National Academics make their online grant databases publicly available which document a variety of information on grants that have been funded over the past few decades. In this study, we present a novel framework to explore the feasibility of automatically estimating the appropriate funding level, given the textual description of a transportation research project, based on a quantitative analysis of the TRB's Research In Progress (RIP) online database. We investigate this approach in an attempt to aid the scientific and transportation research community to quickly gauge the estimated funding level for a research topic they are interested in exploring.

The TRB's Research in Progress (RIP) database is a public online repository that contains information on more than 14,000 (as of January 2017) current or recently completed transportation projects funded mostly by U.S. Department of Transportation (DOT), State DOTs, and U.S. DOT funded university transportation research centers [1,2]. Users of the RIP website can search the entire RIP database by various fields (keywords, title, etc.), browse subject records by subject category (administration and management, aviation, bridges and structures, construction, data, and information technology, etc.), download the records, etc. These records contain much useful information from which knowledge can be extracted using appropriate Data Science (DS) and statistical Text Mining (TM) techniques.

In the reported literature, text mining features have been used for estimating various types of important metrics such as budget, age, cost, etc. Foster et al. [3] used text mining features to estimate the price of a real estate property. Their study focused on building a regression model to predict the price of real estate from its listing, *viz.*, the property description text was used to obtain information about its sale value. Similarly, 'authorship profiling' from anonymous text based on the application of Machine Learning (ML) to text categorization is a growing field of importance owing to its forensics and security applications [4]. In the authorship profiling problem, profile dimensions such as author gender, age, native language, and personality are extracted from a given text of unknown authorship using style-based TM features [5,6]. For

instance, Nguyen et al. [7] developed a linear regression model to predict the author's age from the unknown text. Text mining features have also been useful in predicting movie revenues based on the movie reviews [8]. However, to the best of our knowledge, there is no existing study in the reported literature focusing on developing a text mining-based approach to predict the estimated total budget for scientific research projects, especially transportation research grants.

2. Dataset

The dataset used in this study comes from the TRB's RIP database, a public online repository that contains information on more than 14,000 current or recently completed transportation projects funded mostly by government funding organizations [2]. In addition to funded transportation projects in the US, the RIP database also contains curated records from the International Transportation Research Documentation Database and the Canadian Surface Transportation Research Database. As of March 1, 2017, the TRB RIP database consists of 14,184 research project records. However, not all project entries contain abstracts. For the current work, we used only those project records that had an abstract. Further, only those abstracts with a word count of at least 20 words were included. In this case, we performed a space delimited tokenization of the abstracts to obtain the word count within an abstract. In addition to text-based filtering, we also performed a funding amount based filtering. During the data cleaning process discussed elsewhere, some discrepancies related to the funding amount were identified for some project records. Therefore, we restrict our analysis to projects whose funding amount varies from USD 10,000 to 5,000,000. Filtering out projects which do not meet the criteria mentioned above reduced the records to 10,255 project descriptions (abstracts) with their corresponding funding amounts. Before proceeding with our research approach, we first study the interaction between the various source agencies and the subject areas from 2012-2016 to understand better the funding invested per subject area and the number of projects funded per subject area by the source agencies.

3. Interactions between source agencies and subject areas

In this section, we discuss the results of our experiments conducted to study agency and subject interactions from 2012 to 2016. The interactions are studied regarding (1) the amount of funding invested per subject area, and (2) the number of projects funded per subject area by the source agencies. In the interest of readability of the manuscript, we choose to present our results in this manuscript only for the interactions of type (1).

Fig. **1** and Fig. **2** show the how the top-10 funding agencies in the years 2012 and 2016 have invested across the 37 subject areas. Although similar charts were generated for years 2013, 2014, and 2015, they are not included for the sake of brevity. In Fig. **1** and Fig. **2**, we present the interaction matrix for only the top-10 funding agencies (ranked by the total funding allocated). We have not included the year 2017 in this analysis because of incomplete data for this current year. Agencies in the figure are automatically abbreviated using the 'abbreviate' package in *R*.

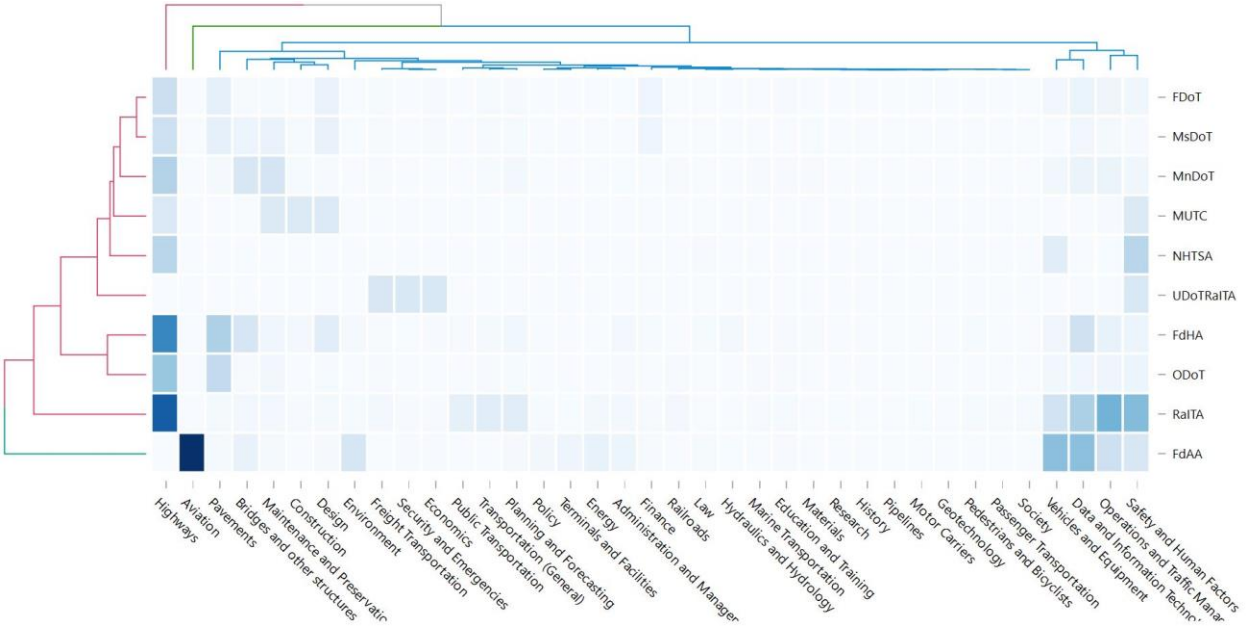


Fig. 1. Heatmap representation of the interactions between top-10 source agencies and subject areas in 2012.

Fig. 2. Heatmap representation of the interactions between top-10 source agencies and subject areas in 2016.

Each cell in the interaction matrix denotes the total funding amount for that year between the agency and the subject area. Darker shades represent higher funding allocations. We study these interactions from the year 2012 to 2016 to observe any striking patterns from this analysis. In the current analysis, we leverage an unsupervised co-clustering algorithm to cluster the rows and columns into groups for organized information visualization. We used the d3heatmap package of *R* for producing the heatmaps and performing the co-clustering. It is a D3.js based heatmap

htmlwidget for *R*. The co-clustering was done using the hierarchical clustering algorithm. The clustering (grouping of various cells) is shown along the right and top sides of the map.

Table 1
Auto-generated abbreviations for source agencies in the RIP database [1].

S.no	Abbreviation	Full name	S.no	Abbreviation	Full name
1	FDoT	Florida Dept. of Transportation	13	ODoT	Ohio Department of Transportation
2	MsDoT	Mississippi Dept. of Transportation	14	RaITA	Research and Innovative Technology Administration
3	MnDoT	Minnesota Dept. of Transportation	15	FdAA	Federal Aviation Administration
4	MUTC	Mid-Atlantic Universities Transportation Center	16	IlDoT	Illinois Department of Transportation
5	NHTSA	National Highway Traffic Safety Administration	17	TDoT	Texas Department of Transportation
6	UDoTRaITA	U.S. Department of Transportation Research and Innovative	18	MDoT	Michigan Department of Transportation
7	FdHA	Federal Highway Administration	19	IwDoT	Iowa Department of Transportation
8	AAoSH&TO	American Association of State Highway & Transportation official	20	NYSDoT	New York State Department of Transportation
9	NCHRP	National Cooperative Highway Research Program	21	CDoT	California Department of Transportation
10	WDoT	Wisconsin Department of Transportation	22	ADoT	Arizona Department of Transportation
11	IDoT	Illinois Department of Transportation	23	GDoT	Georgia Department of Transportation
12	SCDoT	South Carolina Department of Transportation	24	NCDoT	North Carolina Department of Transportation

As shown in figure 13, the Federal Aviation Administration or FdAA invested the maximum amount in the Aviation subject area in the year 2012, and its other dominant allocations of funding have been in the areas of Vehicles and Equipment, Data and Information Technology followed by Safety and Human factors and Traffic Management. From this interaction matrix, we find that federal agencies (such as FdAA, RaiTA, and FDHA) have invested commonly in the four areas mentioned towards the right of the axis. On the other hand, state agencies such as FDoT, MsDoT, and MnDoT have clustered in the areas shows towards the left in the figure

(Pavements, Bridges, Maintenance, Construction, and design). Moreover, "Highways" is a common area where both federal and state agencies have invested significantly than other areas.

Between 2012 and 2016, the funding patterns seem to have changed significantly. In the year 2016 (Fig. **2**), 9 of the top-10 funding agencies are state DoTs. Texas DoT makes the highest investment in Highways research projects in 2016. While most state Dots concentrated their funds in the areas of Highways and Safety & Human Factors, FdHA being a federal agency spread its funds across various areas. One observation from this preliminary analysis of interactions between source agencies and subject areas is that funding allocations tend to vary with subject areas over the years depending on geopolitical factors (ex., passing of highway bill) and policies. Between 2012 and 2016, Highways is one subject area that has received the reasonably consistent level of funding from various state and federal agencies.

4. Research approach

In this section, we discuss the proposed research approach as shown in Fig. **3**. The proposed approach consists of 3 main steps: (1) Text to vector conversion module, (2) Training supervised regression models, and (3) Predicting funding amount using trained regression models. These steps are described in the following sections.

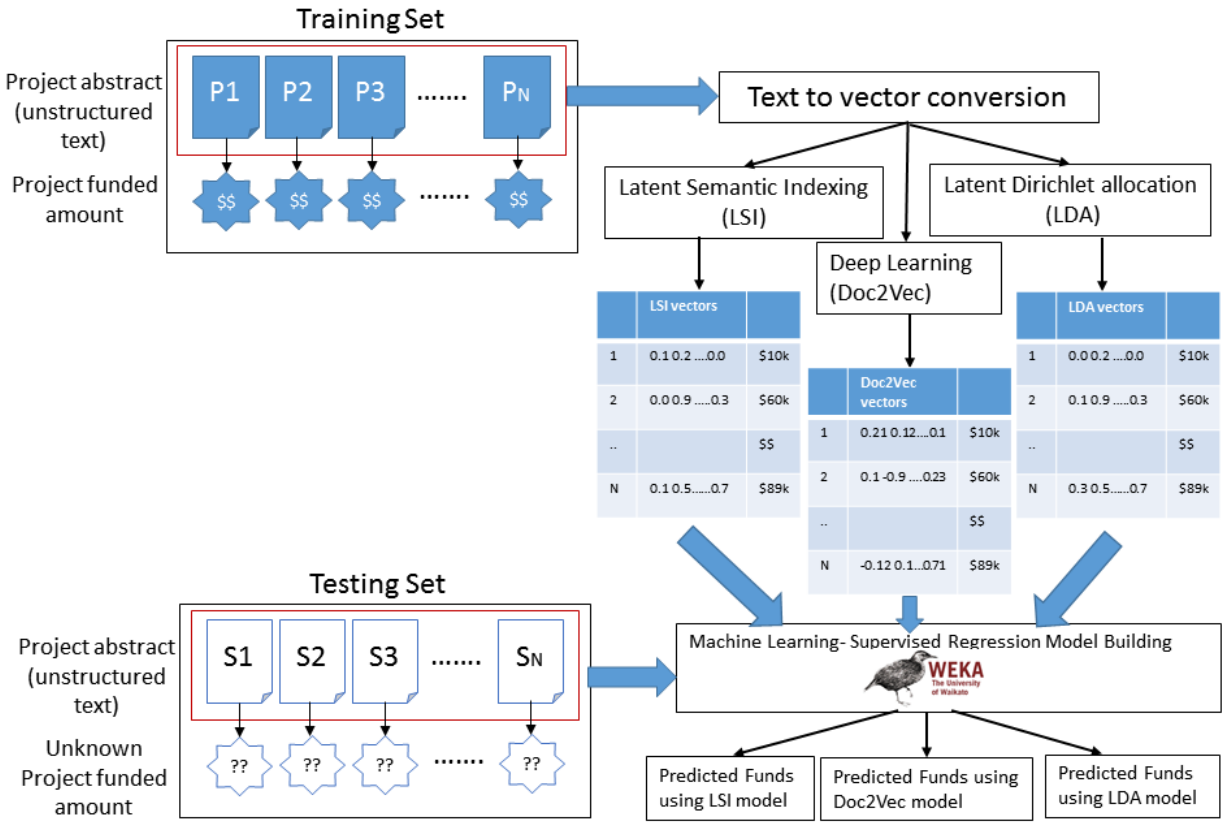


Fig. 3. A schematic of the proposed research approach.

4.1. Text to vector conversion

The abstracts in the TRB's RIP database summarize the project information in the form of unstructured text. We assume that the text in the project abstract captures the essential ideas of the project. However, for using this unstructured text for training a regression model and for estimating the funding level or total budget for a new project, the text documents need to be converted to numerical features describing the document.

We use various vector space models to convert a text document to a variety of vector representations. In vector space models, a document is represented through a fixed dimensional vector where each dimension of the vector represents a feature of the documents [9,10]. In this work, we explore the use of latent vector space models. Latent vector space models aim to identify latent features of the document instead of the generic word count (as in Bag-of-Words model) or weighted word frequency (as in Term Frequency-Inverse Document Frequency or TF-IDF). For a very large corpus, the Bag of Word and the TF-IDF models produce very high-dimensional vector representations for documents. High dimensional data is not appropriate for a regression task. Therefore, we leverage latent vector space models which reduce very-high dimensional document vectors to significantly reduced dimensions. The latent vector space models used in this study are described below:

1. Latent Dirichlet Allocation (LDA) [11]: It is a generative probabilistic model to represent documents in a corpus as a finite mixture over an underlying set of topics. Per this model, each document can be represented as a mixture of various topics, where a distribution over words characterizes each topic.
   Given a document text D from corpus C, we represent it with k-dimensional vector obtained from LDA modeling. In this case $k << |C|$, where $|C|$ is the number of unique text tokens in the corpus.
2. Latent Semantic Indexing (LSI) [12]: It is a mathematical model used to determine the relationship between terms and "hidden" concepts in content. Starting with a TF-IDF or bag-of-words representation, LSI transforms the term-document matrix to term-concept and concept-document matrix using Singular Value Decomposition (SVD).
   Given a document text D from corpus C, we represent it with an l-dimensional vector obtained from LSI modeling. In this case $l << |C|$.
3. Doc2Vec [13]: It is an unsupervised framework that learns continuously distributed vector representations from pieces of texts. The texts can be of variable-length, ranging from sentences to documents. The name Paragraph Vector (popularly known as Doc2Vec) is to emphasize the fact that the method can be applied to variable-length pieces of texts, anything from a phrase or sentence to a large document.
   The approach for learning paragraph vectors (document vectors) is inspired by the methods for learning the word vectors. The inspiration is that the word vectors are asked to contribute to a prediction task about the next word in the sentence. So, even though the

word vectors are initialized randomly, they can eventually capture semantics as an indirect result of the prediction task.

In Doc2Vector framework, every document text (paragraph) is mapped to a unique vector, represented by a column in matrix D and every word is also mapped to a unique vector, represented by a column in matrix W. The document vector and word vectors are averaged or concatenated to predict the next word in a context.

Paragraph vectors also address some of the key weaknesses of bag-of-words models. First, they inherit an important property of the word vectors: the semantics of the words. In this space, "powerful", for instance, is closer to "strong" than to "Paris." The second advantage of the paragraph vectors is that they take into consideration the word order, at least in a small context, in the same way, that an n-gram model with a large n would do.

Given a document text D from corpus C, we represent it with an m-dimensional vector obtained from Doc2Vec modeling. In this case $m << |C|$.

4.2. Predictive modeling using regression

As shown in Fig. **3**, the vector representations of the documents are used as features in a regression model. A regression model learns the relationship between one or more dependent variable (denoted by Y or Y1, Y2, Y3, etc.) and a series of other changing variables (known as independent variables). In the proposed approach, the budget or the funds allotted to a transportation research project is the target variable or dependent variable. Various regression models are used to learn the relationship between projects' total budget and the associated text description (represented in vector form as described in the previous step).

Predictive modeling using regression analysis is accomplished in two phases: training and testing. In the training phase, we use a set of transportation research project records and their corresponding budget to train the model about the relationship between the text features and the budget. In the testing phase, given a research project description represented by text features, its budget is estimated as the prediction of the regression model.

5. Experiments

The experiments are carried out using a combination of Python, R, and WEKA [14] frameworks. Wikato Environment for Knowledge Analysis (WEKA) is a suite of open-source ML software developed at the University of Waikato, New Zealand and is written in Java.

For all the experiments, we use a 10-fold cross-validation approach to validate the performance of the proposed approach. In k-fold cross-validation, the entire dataset is randomly partitioned into k equal size subsamples, where a single subsample is used as the validation data for testing the model and the remaining k-1 subsamples for training the model. This process is repeated k times while ensuring that each of the k subsamples is used exactly once as validation data. k-fold

Cross validation ensures the separation of training and testing sets and removes the bias of splitting data into training and test sets.

In this work, we used three popular regressions models, namely, Random Forest (RF) [15], Decision Stump (DS) [16] and Linear Regression (LR) [17]. The default parameter settings available in the WEKA machine learning suite were employed for each of these regression models.

Quantitative assessments of the degree to how close the models could predict the actual outputs are used to provide an evaluation of the models' predictive performances. A multi-criteria assessment with two different goodness-of-fit statistics was performed using all the data vectors to test the accuracy of the trained models. The criteria that are employed for evaluation of the models' predictive performances are the Mean Absolute Error (*MAE*) and Root-Mean-Squared Error (*RMSE*) between the actual and predicted values.

We use two metrics to report and compare the performance of various regression models in the next section:

- Mean Absolute Error (MAE): It is the mean of the absolute difference between the predicted value and the actual value of the dependent variable. In the present case, the dependent variable is the budget of the project.
- Root Mean Squared Error (RMSE): It is the root of the mean of the squared values of the difference between the predicted value and the actual value of the dependent variable.

6. Results and discussion

In this section, we discuss the results of the performance evaluation of the three text mining feature generations approaches. We compare their performances in the following ways:

1. ***Comparing feature length:*** In this analysis, we compare the performances of our regression approach by varying the feature length of text vector. As shown in Fig. **4** (a-f), the length is varied from 10 dimensions to 300 dimensions of the latent features. The figures show the evaluation of two metrics MAE and RMSE. As shown in Fig. **4**(a-b), for RF regression, Doc2Vec feature generation approach has the lowest error at ten dimensions and error increases as the dimensions are increased. For LDA, we observe that both MAE and RMSE decrease as the number of dimensions are increased from 1 to 300. For the DS regression, we find that the impact of increasing the dimensions fluctuates with the number of dimensions (Fig. **4** (c-d)). With the LR model, we find that errors (MAE and RMSE) decrease as the number of dimensions increase (for LDA and LSI) (Fig. **4** (e-f)).

Fig. 4. Comparing various text mining feature generation models by varying the feature length.

2. ***Comparing Regression models***: In this section, we compare the various regression models (Random Forest, Decision Stump, and Linear Regression) using the MAE metric. As shown in Fig. **5**, for Doc2Vec features, Random Forest regression gives the highest errors whereas LR consistently gives lower errors.

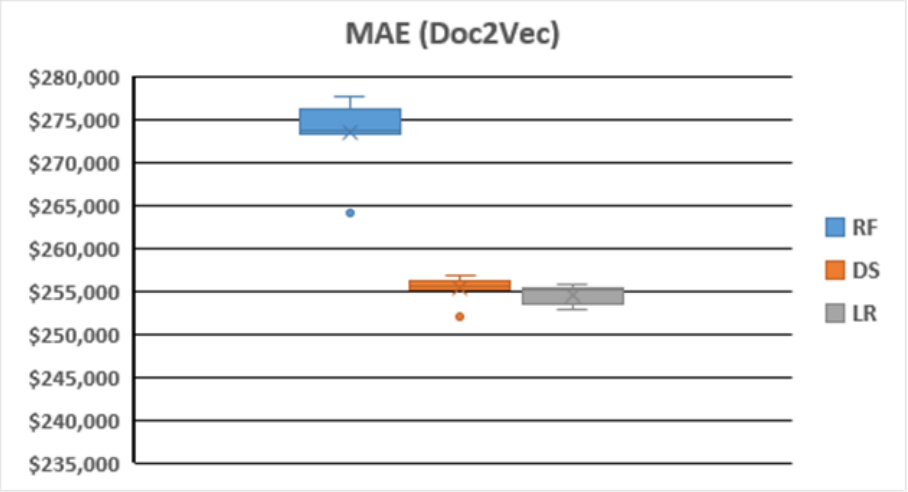


Fig. 5. Box plot showing MAE for Doc2Vec Features for all Dimensions.

For LDA features, we find that DS and LR consistently give higher error compared to RF (Fig. **6**). We observed an opposite phenomenon with the Doc2Vec features (discussed above).

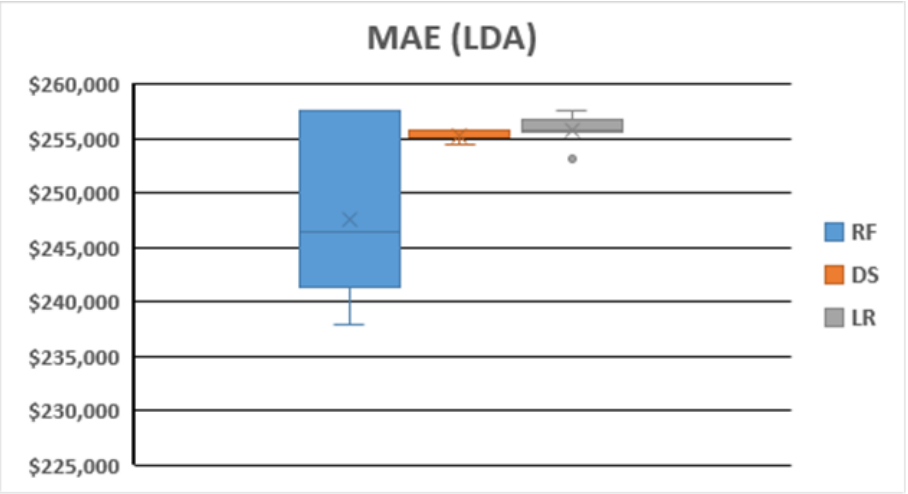
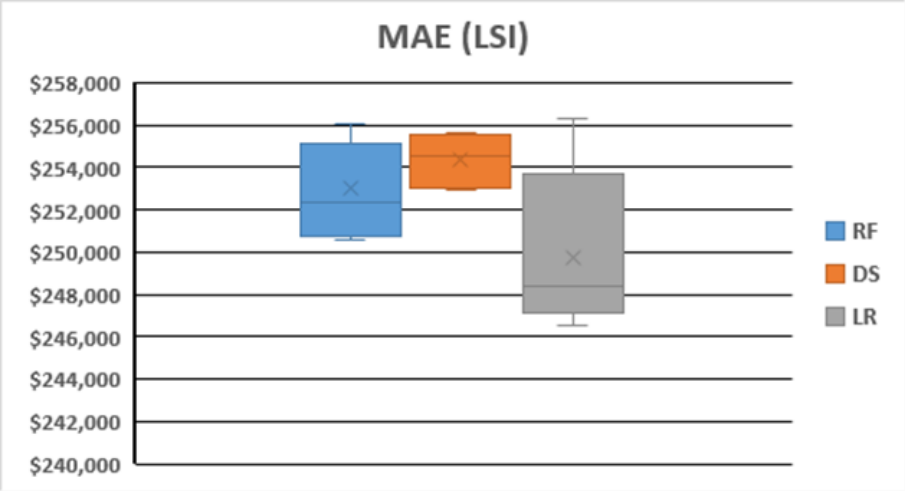


Fig. 6. Box plot showing MAE for LDA Features for all Dimensions.

For LSI, as shown in Fig. **7**, the variance of error is large for all the regression models. However, the errors are lower for the LR model compared to RF and DS regression.

3. ***Comparing feature generation models (Doc2Vec, LDA, LSI):*** Finally, we compare the three feature generation approaches (Doc2Vec, LDA, and LSI) on their best performances

in project budget prediction. In



4. Fig**. 7.** Box plot showing MAE for LSI Features for all Dimensions.
5. **Table 2**, we show the various parameters such as the feature dimension and the regression model which give the lowest MAE and RMSE for the feature generation models.

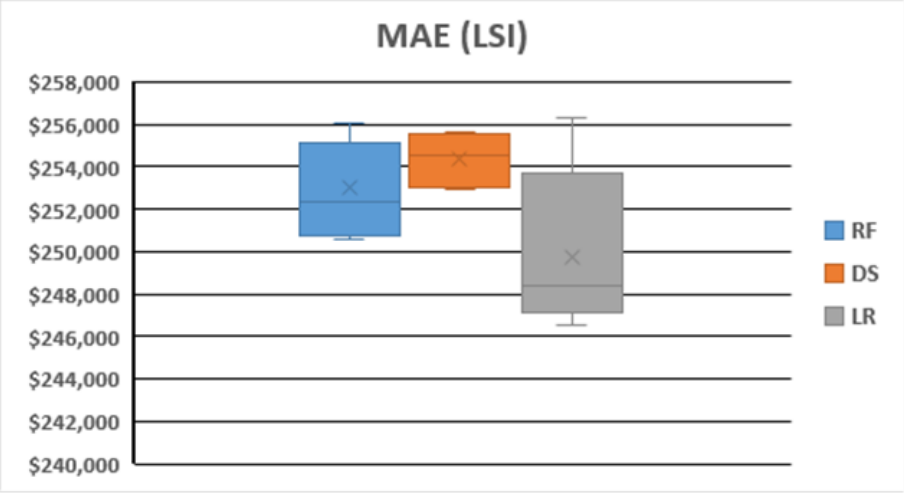


Fig. 7. Box plot showing MAE for LSI Features for all Dimensions.

Table 2
Comparing the best performances of Doc2Vec, LDA, and LSI.

	Doc2Vec			**LDA**			**LSI**		
	Value	**Dim**	**Model**	**Value**	**Dim**	**Model**	**Value**	**Dim**	**Model**
Min MAE	252,110	300	DS	237,884	300	RF	246,500	100	LR
Min RMSE	471,199	300	LR	461,678	300	RF	455,152	300	LR

6. ***Using pre-trained models with transfer learning:*** We also record the significance of using pre-trained Doc2Vec models (models trained on external corpuses) with a method called *transfer learning*. The pre-trained models have learned to document embedding from the different corpus that includes words and concepts similar to the transportation document corpus we have used. In this experiment, we have used two pre-trained models, namely, Wikipedia based and AP-News based. The performance of the Doc2Vec pre-trained prediction model is shown in the table below. The models were pre-trained to represent a document in 300-dimension space only. As shown in the table, the RMSE is lowered by using the pre-trained models and linear regression estimator.

Table 3
Performance evaluation of Doc2Vec using pre-trained models with transfer learning.

	Wikipedia			AP-news		
	RF	**DS**	**LR**	**RF**	**DS**	**LR**
MAE	267,234	256,765	255,303	269,009	254,996	254,642
RMSE	471,915	483,100	468,097	472,428	482,271	468,895

7. ***Budget estimation using filtered dataset:*** The previous experiments showed that it is not very feasible to accurately predict the funding level directly from the unstructured project abstract, given the large variations in source agencies, subject areas, and funding levels ranging from USD 10,000 to 5,000,000. In this experiment, we conducted our budget estimation for a filtered set of funding agencies. We selected Utah DoT and Oklahoma DoT funding agencies and all the projects funded by these. There were a total of 200 projects funded by these. Additional filters were imposed, such as funding levels were limited between USD 10,000 USD to USD 5,000,000; abstract word count of greater than 10. Table 4 summarizes the performance of LSI-based budget estimation models using the filtered dataset. As shown in Table 4, the LSI model using LR gives a correlation (R) value of 0.635. The MAE and RMSE are minimum for LSI using RF. Other text mining feature generation models (LDA and Doc2Vec) did not perform well in this experiment, and therefore the results are not included. In comparison to the results from previous experiments, we find a significant improvement in the prediction performance of the LSI model. Thus, using a separate model for different types of funding agencies seems to be the better approach for budget estimation.

Table 4
Performance of LSI-based budget estimation models using filtered dataset.

	LSI		
Model	**Correlation (R)**	**MAE**	**RMSE**
RF	0.5831	338,159	626,732
LR	0.635	382,606	691,207
DS	0.5033	388,325	660,978

7. Conclusions

We proposed a fully automated machine learning approach to predict the approximate budget for a research project, given a short description of the project. For this, we investigated latent feature generation models such as Doc2Vec, LDA, and LSI to convert project text description to numeric features which are later used in machine learning regression models to predict project budget. We tested the proposed approach TRB database containing approximately 14,000 project entries and their approved budget information, ranging from USD 10,000 to 500,000,000. Finally, the proposed approach was quantitatively validated using various experiments. The following are the significant findings:

- Out of the three latent feature generation models, we found LDA gives the least MAE using 300 feature dimensions and Random Forest regression model.
- LSI gives the least RMSE using 300 feature dimensions and Linear regression model.
- However, based on the correlation coefficients, it was found that it is not very feasible to accurately predict the funding level directly from the unstructured project abstract, given the large variations in source agencies, subject areas, and funding levels.
- By using separate prediction models for different types of funding agencies, funding levels were better correlated with the project abstract.
- The proposed approach provides a framework which can be built into a tool to help the research community to estimate the budget or funding level for their research projects.

References

[1] Daly J. TRB Webinar: Learning About and Using the Research in Progress (RiP) Database 2016:14. http://www.trb.org/ElectronicSessions/Blurbs/174599.aspx.

[2] Gopalakrishnan K, Khaitan SK. TEXT MINING TRANSPORTATION RESEARCH GRANT BIG DATA: KNOWLEDGE EXTRACTION AND PREDICTIVE MODELING USING FAST NEURAL NETS. Int J TRAFFIC Transp Eng 2017;7. doi:10.7708/ijtte.2017.7(3).06.

[3] Foster DP, Liberman M, Stine RA. Featurizing Text: Converting Text into Predictors for Regression Analysis. Whart Sch Univ Pennsylvania, Philadelphia, PA 2013.

[4] Argamon S, Koppel M, Pennebaker JW, Schler J. Automatically profiling the author of an anonymous text. Commun ACM 2009;52:119–23.

[5] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS One 2013;8:e73791.

[6] Rosenthal S, McKeown K. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. 1, Association for Computational Linguistics; 2011, p. 763–72.

[7] Nguyen D, Smith NA, Rosé CP. Author age prediction from text using linear regression. Proc. 5th ACL-HLT Work. Lang. Technol. Cult. Heritage, Soc. Sci. Humanit., Association for Computational Linguistics; 2011, p. 115–23.

[8] Joshi M, Das D, Gimpel K, Smith NA. Movie reviews and revenues: An experiment in text regression. Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist., Association for Computational Linguistics; 2010, p. 293–6.

[9] Singhal A, Kasturi R, Srivastava J. Automating Document Annotation Using Open Source Knowledge. 2013 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol., vol. 1, IEEE; 2013, p. 199–204. doi:10.1109/WI-IAT.2013.30.

[10] Singhal A, Srivastava J. Research dataset discovery from research publications using web context. Web Intell 2017;15:81–99. doi:10.3233/WEB-170354.

[11] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993–1022.

[12] Landauer TK. Latent Semantic Analysis. Encycl. Cogn. Sci., Chichester: John Wiley & Sons, Ltd; 2006. doi:10.1002/0470018860.s00561.

[13] Le Q, Mikolov T. Distributed representations of sentences and documents. Int. Conf. Mach. Learn., 2014, p. 1188–96.

[14] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016.

[15] Breiman L. Random Forests. Mach Learn 2001;45:5–32. doi:10.1023/A:1010933404324.

[16] Holte RC. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Mach Learn 1993;11:63–90. doi:10.1023/A:1022631118932.

[17] Lai T., Robbins H, Wei C. Strong consistency of least squares estimates in multiple regression II. J Multivar Anal 1979;9:343–61. doi:10.1016/0047-259X(79)90093-9.