NATURAL LANGUAGE PROCESSING TO PREDICT KICKSTARTER PROJECT SUCCESS

Authors
Marissa Lafreniere / 57812456 / mlafreni@uci.edu
James Kim / 49475860 / leeseokk@uci.edu
Jasmine Pham / 71784204 / thaotp3@uci.edu
Sa Yang / 85326121 / say1@uci.edu

ABSTRACT

Kickstarter is a crowdfunding platform that monetarily supports creative projects. Records show that, as of July 2022, the platform has received $6.67 billion U.S. dollars from more than 20 million backers to fund 221,669 projects, making a success rate of 39.78%[1]. To understand what kinds of projects are favored by investors, we used deep learning based models to predict the possibility of success based on each project's short description. Applied models include LSTM and Bidirectional Encoder Representations from Transformers (BERT). We believe our work will provide empirical and authentic references for project owners and improve their chance of being successfully funded.

1 INTRODUCTION

Kickstarter is a crowdfunding platform that supports creative independent projects. On the website, each project is exhibited with a cover video or image, a short description, a deadline and their funding goal. Since the platform applies an "all or nothing" protocole, projects that failed to reach their targeted amounts of money will not be funded. As a result, unsuccessfully funded projects took up the majority. Although the successful rate of a project depends on many factors, we believe the description of the project is the key to success.

To understand what kind of description stands higher possibility for investment, we applied key transformer models in Natural Language Processing (NLP) to capture the structure of natural language as well as its quality. Compared to traditional pre-deep-learning approaches, which focus on sentence length, word of choice, etc, our approach could capture more complex features in the sentences using deep neural networks. Our project involves Natural language processing(NLP), which is a combination between computer science, linguistics and AI, where natural language data is used to find the pattern in the dataset to perform some repetitive tasks [1]. It can be considered as a way of communicating and interacting between humans and computers. By NLP, computers can read texts, analyze, and interpret it after learning a sufficient set of data. Our team will utilize NLP to analyze the information contained in burbs of the project and use that to predict.

It is easy for a person to read the introduction or blurbs of a project and understand the concept of the project so that they will decide if a project is practical and contributes to the community. And with their understanding of the project, they will have their decision on whether to fund that project or not. But it was a challenge for the computer to perform the same task. That is why it needs advanced technique(NLP) in order to break down the blurbs and analyze the sentence's patterns and categorize them into useful or not-useful information. And after the whole analysis, the computer can predict the chances that the project will be successful.

Our approach to this project is that we will start with separating the words in the blurbs, count them and use those with significant contribution to a successful project to predict other projects' possibility of being chosen to fund. Therefore, we chose *Term Frequency - Inverse Document Frequency (TF-IDF), Machine learning models using Gensim preprocessing, Long short-term memory(LSTM)* and *Bidirectional Encoder Representations from*

[1] https://www.kickstarter.com/help/stats

Transformers (BERT) to build our model. With this approach, we think the main contribution to the accuracy of this project is the words that are used in the blurbs of projects. If strong and effective words were used, it is more likely that the project will be successful.

2 RELATED WORKS

2.1 SUICIDE NOTE CLASSIFICATION USING NATURAL LANGUAGE PROCESSING

[2] discusses approaches and techniques used in classifying suicide notes as authentic or fake using natural language processing. A semi-supervised approach was used due to the small dataset size. The types of models explored fall into 5 categories: decision trees, classification rules, function models, lazy learners, and meta-learners. In the end, they produced a model that correctly classified notes 78% of the time which was better than untrained individuals who performed with 49% accuracy and mental health professionals with 63% accuracy. They theorize this is because humans tend to focus on the content of the letters while the machine algorithms tended to focus on structure.

2.2 MEDICAL SUBDOMAIN CLASSIFICATION OF CLINICAL NOTES USING A MACHINE LEARNING-BASED NATURAL LANGUAGE PROCESSING APPROACH

Medical notes have different subdomains such as cardiology or neurology. The research group in [3] constructed a Machine Learning Natural Language processing pipeline to automatically classify notes. In their model they experimented with “the clinical NLP system, clinical Text Analysis and Knowledge Extraction System (cTAKES), the Unified Medical Language System (UMLS) Metathesaurus, Semantic Network, and learning algorithms". They were tested using two different data sets to see how well they performed independently as well as on each other, their portability. In the end, the convolutional recurrent neural network with neural word embeddings trained-medical subdomain classifier performed the best with average area under receiver operating characteristic curve (AUC) of 0.983 and average F1 score of 0.8575 between the two data sets.

2.3 NATURAL LANGUAGE PROCESSING VERSUS RULE-BASED TEXT ANALYSIS: COMPARING BERT SCORE AND READABILITY INDICES TO PREDICT CROWDFUNDING OUTCOMES

This study believes that traditional writing structure analysis could not thoroughly evaluate a writing. Thus, the researchers use Bidirectional Encoder Representations from Transformers (BERT) to quantify the value. The result shows that the quality of writing is not the sole feature that decides the chance a crowdfunding will be favored by backers.

3 DATASET

Our dataset consists of 215,513 kickstarter projects from 2017[2]. The dataset is stored as a csv file, containing three fields, index, short descriptions and their corresponding state being either “failed” or “succeeded”. The first field is an index from 1 to 215513. The second field is a “blurb” which is a high level description of the project written by the project creators to entice users to donate to their project being completed. The last column is the label for the projects which are either “successful” or “failed. In the scope of this project, “successful” is defined as a kickstarter project that received all the funding they were asking for and “failed” is defined as a project that did not.

4 METHODS

4.1 TF-IDF (TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY) WITH SIMPLE LOGISTIC REGRESSION

[2] https://www.kaggle.com/datasets/oscarvilla/kickstarter-nlp

The first model created was the simplest starting with a semi-supervised logistic regression model. First all of the data (both training and testing) was put through a sklearn TF-IDF Vectorizer. Term Frequency - Inverse Document Frequency is a technique used in information retrieval to weigh the importance of different words in a document. Then a simple linear regression model is fit to data.

4.2 PREPROCESSING DATA WITH GENERATE SIMILAR(GENSIM)

Gensim processed unstructured data using unsupervised language algorithms (such as Word2Vec, FastText, Latent Semantic Indexing, etc) to capture the relationship between words within the training document and discover its semantic structure. By preprocessing our dataset with gensim, we can ensure our data performance. (Since more complicated model still did not work well, I used logistic regression and naive bayes to predict the data)

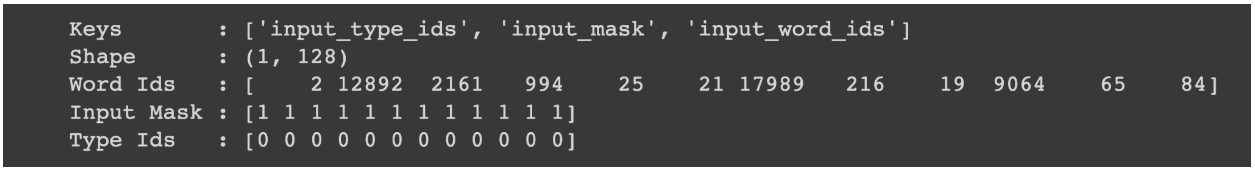
4.3 LONG SHORT-TERM MEMORY (LSTM)

LSTM is an updated version of Recurrent Neural Network(RNN) to overcome the vanishing gradient problem. The architecture of LSTM has a memory cell at the top which helps to carry the information from a particular time instance to the next time instance in an efficient manner. After processing data, we convert text tokens into a sequence of integers. We implemented a LSTM model with embedding, LSTM and dense layers, then used splitted data to train and evaluate the model.

4.4 BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

The $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ language models are transformer models built by Google pre-trained using data retrieved from BooksCorpus and the English Wikipedia. It can be further tuned to perform a more specialized task by a smaller set of data. The first model will utilize DistilBERT which is a light weight and faster version of BERT.

BERT model has its special preprocessing procedure, the first preprocessing step is to convert each blurb in the dataset into a tuple containing the blurb's input ids and attention masks, which are an optional argument used when batching sequences together.. Figure 1 shows a representation of the preprocessed text.



```
    Keys       : ['input_type_ids', 'input_mask', 'input_word_ids']
    Shape      : (1, 128)
    Word Ids   : [    2 12892  2161   994    25    21 17989   216    19  9064    65    84]
    Input Mask : [1 1 1 1 1 1 1 1 1 1 1 1]
    Type Ids   : [0 0 0 0 0 0 0 0 0 0 0 0]
```

Figure 1: Preprocessed Text

The dataset states were changed to binary values and the blurbs were tokenized using DistilBertTokenizer. The model only went through 2 epochs of fitting the data as there were issues with system resources.

5 RESULTS

We would like to use a variety of methods to evaluate the performance of our models. The simplest is the raw percentage, it is the ratio of how many correct predictions were made divided by the total number of predictions made. Log loss is the negative average of the log of corrected predicted probabilities. Lastly, area under receiver operating characteristic curve (AUC) is the area under a receiver operating characteristic curve (ROC) which shows the performance of a classification model at all classification thresholds.

5.1 TF-IDF (TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY) WITH SIMPLE LOGISTIC REGRESSION

At first, the TF-IDF model did not perform particularly well with a log loss of 0.591 and score of 0.3192. However, since this is a binary classification problem the predictions were able to be inverted. This resulted in the same log loss, but a percent and AUC score of 0.6807 which is much closer to the performance of the other models.

5.2 PREPROCESSING DATA WITH GENERATE SIMILAR(GENSIM)

After processing raw data with Gensim, we used logistic regression and naive bayes model to predict data. The result is more than what we expected, especially for Naive Bayes (with 0.7589 log loss and 67% accuracy score).

5.3 LONG SHORT-TERM MEMORY(LSTM)

We expected the result of the deep learning model would be higher than the machine learning model. However, after the process of cleaning data and fitting the LSTM model with different values of hyperparameters, although the training accuracy is good enough (around 73%, AUC - 0.8114), the evaluation accuracy score is 67.08% and AUC is 0.7337

5.4 BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

At this point in our project, we do not have enough preliminary results to report on the accuracy of the BERT model. However, based on related research papers as seen in [4] and [5], we expect the performance to far exceed that of the TF-IDF model.

Table 1: Summary of Model Performance

Model	**Log Loss**	**Raw Percent**	**AUC**
TF-IDF with Logistic Regression	0.591	0.6807	0.6807
DistilBERT	0.6225	0.6617	0.6618
Naive Bayes, pre- processing with Gensim	0.7589	0.672	0.672
Logistic Regression, preprocessing with Gensim	0.6213	0.6661	0.6662
LSTM	0.6229	0.6708	0.7337

6 FURTHER WORKS

A Kickstarter project provided investors with more than just a blurb, there are also introduction videos, funding goals and intended deadlines that project would be finished. Thus, although blurbs account for the most percentage in the successful rate, there are many other factors that would contribute to the success of a project. We cannot depend solely on the introductory descriptions to get a high accuracy success rate. But in a short period of time, it is not efficient for us to investigate all of these features. Therefore, for further work, we can investigate more on other aspects of the projects such as project categories, introduction video, funding goals and deadlines to see their effects on the success of a project. Moreover, we can also evaluate the importance of each factor regarding its effects.

7 CONCLUSION

This project will help a lot in providing project creators a general overview of how their project description affects the chance of success. From the result, we can see that descriptions play an important role in predicting the success of a project. We implemented both machine learning and deep learning models, and got results around 68%. With the idea in future works, we hope to improve the predicting result with more complex datasets and models, including multimedia classification, creator behavior classification, and with consideration of various other parameters.

REFERENCES

[1] Hirschberg and Manning, 2015 Julia Hirschberg, Christopher D. Manning Advances in natural language processing Science, 349 (6245) (2015).

[2] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide Note Classification Using Natural Language Processing: A Content Analysis. Biomedical Informatics Insights. https://doi.org/10.4137/BII.S4706

[3] Weng, WH., Wagholikar, K.B., McCray, A.T. et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 17, 155 (2017). https://doi.org/10.1186/s12911-017-0556-8

[4] Mitra, T., & Gilbert, E. (2014). The language that gets people to give. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. https://doi.org/10.1145/2531602.2531656

[5] Chan, C. R., Pethe, C., & Skiena, S. (2021). Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes. Journal of Business Venturing Insights, 16, e00276. https://doi.org/10.1016/j.jbvi.2021.e00276

[6] Mitra, T., & Gilbert, E. E. (2014, February 1). *The language that gets people to give: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM Conferences. Retrieved June 9, 2022, from https://dl.acm.org/doi/abs/10.1145/2531602.2531656

ACKNOWLEDGEMENTS

Marissa Lafreniere

- Research different projects and datasets
- Worked on abstract
- Researched and discussed Suicide Note classification and Medical Subdomain classification in related works section
- Implement TF-IDF with logistic regression model and discuss in methods and results section
- Assist with creating final presentation and presenting project in front of class
- Write description of data set

James Kim

- Assisted with introduction and report
- Worked on model descriptions
- Worked on model results
- Researched *Natural language processing versus rule-based text analysis* in related works
- Introduced future ideas for implementations and different parameters to further consider
- Worked on the BERT Model
- Worked on presentation slides

Jasmine Pham

- Assisted with Abstract.

- Worked on Introduction, Conclusion and further works for the report.
- Worked on LSTM models (code, method and result)
- Worked on Naive Bayes and logistic regression model using gensim preprocessed data (code, method and result).
- Participated in preparing presentation slides.

Sa Yang

- Assisted with Introduction
- Assisted with Abstract
- Assisted with Dataset
- Assisted with Method (BERT)
- Tuned BERT base model