DSCI 644 Project Proposal

Rochester Institute of Technology

Michael Kogan
Michael Kitching
Akanksha Arora
Mohammadreza Shojaei Kol Kachi
Muhammad Fazalul Rahman

Sept 13, 2020

Table of Contents

Introduction

The purpose of this document is to outline the problem, project purpose, preliminary approach, and key requirements. This document will serve as a starting point for this project and will drive the remainder of our work. The requirements outlined in this document will be continuously revisited to ensure that they are being met as the project moves forward.

Your most unhappy customers are your greatest source of learning - Bill Gates

Today's era is about reviews, customer satisfaction and ensuring high product quality. Every business wants to be a leader in it’s own field and strives to expand into others. Customer feedback is a popular and effective way to measure business performance. Amazon, which is one of the largest companies in the world, uses a simple review approach for their applications. Ratings on a pre - defined scale with associated text reviews provided by the users help gauge how well each application is received. There are many ways of analyzing these reviews, and the currently implemented multiclass classification model is only able to attain an accuracy of about 60%. Furthermore, it does so by predicting the most common class, which is a very rudimentary machine.

Project Purpose

Our customers need a robust and high - accuracy model that is able to interpret text reviews. The purpose of this project is to revisit the currently implemented approach and develop it into a high - quality model. In order to accomplish this, we will redefine the problem as a binary classification problem, classifying reviews as good or bad. This is a coarser scale which should allow much higher accuracies while still allowing the client to draw conclusions about text reviews.

We will develop a sentiment analysis model that will determine the probability of a given text review being positive. The model will be built in Azure ML, and will need to attain an accuracy of at least 90% on test data. A project website will be created to act as a gateway to the project, containing all relevant details and links to all documents. This website will be hosted on GitHub. The project will follow the OpenUP process and will be broken down into four sprints, which will be managed through Trello.

Preliminary Plan

Our initial approach to this is to look at areas where the current model is lacking and improve on them. From preliminary analysis we have identified three areas for improvement:

1. Text pre - processing
2. Text representation
3. Modelling

Our plan is to improve on each of those areas to attain an improved model for this dataset.

Text Pre - Processing

Currently, the pre - processing of the text includes steps which may be unnecessary and may actually hurt the accuracy of the model. One of the key focuses here will be the selection of stopwords. Not all stopwords in the default set should be removed. For example, the word “not” is considered a stopword but can be really valuable, especially with N - Gram representation. There is a significant difference between “good” and “not good”, and removing the word “not” makes those two statements identical.

Other stop words should be included but are not in the default set. Namely, these are the words “full” and “review”, which appear at the end of every review and therefore offer no information. Further research will be done to identify the correct pre - processing steps for our model.

Text Representation

Currently, the text is represented using Latent Dirichlet Analysis (LDA), meaning the entire set of reviews is broken down into topics and each text is represented as a vector of percentages, where the percentages correspond to how much the text aligns with each identified topic. This may not be the ideal representation of the text, and as such we will explore other representations, including the famous bag - of - words representation.

Modelling

Currently, the output is generated by a multi - class neural network. We believe that a better analysis of sentiment is a model that generates the probability of a review being positive or negative. As such, we will focus on binary classification with probability output. We will investigate several models, including XGBoost, logistic regression, and neural networks. These models can each use different text representations.

Another weakness of the current model is that it only uses a fraction of the available data for training. This is an issue, especially for a neural network which must tune many parameters and therefore needs a large amount of data. We intend to use the entirety of the dataset to develop our model.

Project Management

The project will be managed using OpenUp methodology. In this process, there are four phases with key deliverables during each phase. The first phase, the Inception Phase, will outline the project proposal and key requirements, noted in this document. The second phase, the Elaboration phase, will deliver a Github repository, Trello board, a project webpage, and any related architecture which the team will use going into the next phase. This next phase, the Construction phase, will include the design of and deliverance of the working product. In the final phase, the Transition phase, will present the product, assess its success per the user requirements and wrap up the project.

The team will manage the project using formal communication in slack and project management in Trello. The deliverables will be stored in the Github repository and linked to the project webpage. The final product will be delivered per the user requirements below.

Requirements

Requirement	Title	Description
FR1	Enhanced Model	The provided model needs to be re - visited and enhanced to meet the performance requirements.
FR2	Model Input	The model shall receive as input only a text review. No other input variables should be considered.
FR3	Model Output	When given a text review, the model will output the probability that the review is positive. A review is considered positive if the associated rating is > 0.5.
FR4	Model Evaluation	The model will be evaluated using 0 - 1 loss on the test set.
FR5	Data Set	The solution must use the entirety of the provided dataset: “AppReview.csv”.
FR6	Data Exploration	The solution should examine various pre - processing and text representation approaches.

FR7	Project Website	A website needs to be created that will host the project and act as a gateway to all deliverables.
FR8	Website Content	Project details, including the project's purpose, requirements, goals, and architecture, should be included on the website. In addition, all project documentation and the Azure ML Studio model need to be linked.
FR9	Project Documentation	Each sprint is to be capped by a report which will include all deliverables associated with that sprint, which are listed in the "Team Project Instructions" document. These will be made available on the website.
PR1	Prediction Latency	A single prediction should take no more than 0.1 seconds.
PR2	Model Accuracy	The model should achieve test error of less than 10%.
ImpR1	GitHub	The project website must be hosted on GitHub.
ImpR2	Trello	The project must be managed through Trello.
ImpR3	Azure ML Studio	The model must be implemented in Azure ML Studio.
ImpR4	Project Lifecycle	The project is to follow the OpenUP process and be broken down into four sprints: Initiation, Elaboration, Construction, and Transition.