Project Kindle

Devoja Ganguli: M.A. Student, Department of Economics
Nazila Shafiei: Ph.D. Student, Department of Linguistics
Suji Yang: M.A. Student, Department of Linguistics

May 4, 2019

1 Introduction

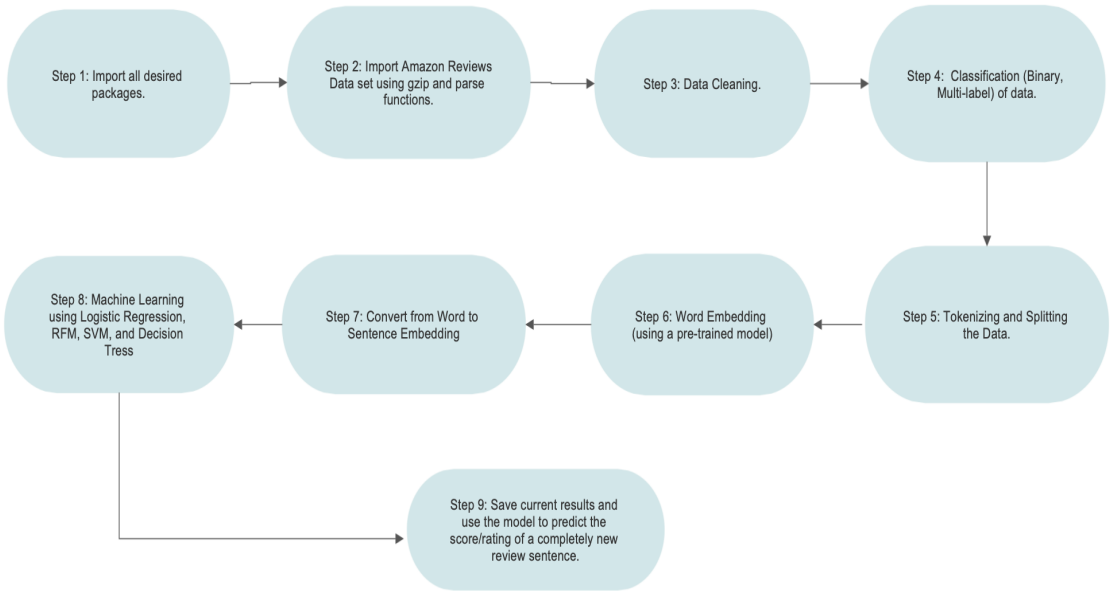
Increasing dependence of mankind on the internet has users relying on product reviews to decide whether or not to purchase a product. While product reviews are important to users who would like to make the decision to purchase, it is also useful to the company selling it. Due to the vast amount of opinion text available on websites, it is impossible for anyone to come to a conclusive opinion about the product by simply reading through the information available. To overcome this hurdle, one can use 'Sentiment Analysis', which is a useful tool for categorizing the data (review text) into positive and negative sentiments. Sentiment analysis or *opinion mining* is currently a popular research topic in the field of Natural Language Processing (NLP) and Linguistics (Liu 2012).

In this project we use Amazon Kindle Reviews (McAuley, UCSD) to perform sentiment analysis and build a model that predicts the reception of the e-book amongst its readers. This kind of a model could in the future be used by Amazon to decide which e-books they should continue to list and which ones they can take off the website. At a broader level, the model could also be used to curate a list of suggestions to a user based on his/her reviews of past preferences and purchases.

2 Overview

The main package/library used in this project is the scikit-learn library of Python for supervised machine learning. Other libraries used includes `gzip`, `json`, `Numpy`, `Pandas`, `Scipy`, `nltk`, and `gensim`. The original Amazon Reviews data file was available in a zipped file format which required the use of the library `gzip`. Once imported, the file was converted into a data frame to allow for easier analysis of data. The library `json` was also used in the process of importing the data into the python script. `Numpy` and `Pandas` were used throughout the code to perform various operations

on the data frame, including the creation of empty vectors to store modified variables. The library `nltk` helped with filtering stop-words from the text reviews to achieve a higher accuracy level of the model. `Gensim` library provided a pre-defined `word2vec` model that we implemented in our code to obtain word embeddings for the reviews.



Our entire project can be easily mapped out as a flowchart (included above). The subsequent sections describe in detail each of the steps in the flowchart.

3 Data

We chose to perform our analysis on Amazon Kindle e-book reviews. The complete dataset is available for free by Dr. McAuley (UCSD), and includes 142.8 million Amazon product reviews spanning between May 1996 - July 2014. The dataset contains not only reviews, but also ratings and helpfulness votes. For the purpose of this study, we use only a small subset of the data, focusing only on those reviews pertaining to the Kindle e-book. We also concentrate on the product code, the actual review and the 'overall' rating of the e-book. Other product specific characteristics like the 'helpfulness vote', 'reviewer ID' , 'summary' , and timestamps were discarded to create less confusion during the analysis. As already mentioned, the data was available in a zipped format which required the use of the `gzip` library and parse functions, following which we converted the file into a data frame. After the data was converted into a data frame, we proceeded to step 3, Data Cleaning.

Cleaning basically means getting rid of unnecessary data and keeping what matters. Dealing with a data frame requires the use of `pandas` library. After removing all the undesirable columns from our data frame, we stripped the text reviews of any punctuation marks. All uppercase letters were converted to lowercase letters, white spaces and empty lines were removed. A filter for stop words was also applied to remove generic, reoccurring words like 'i', 'me', 'her', 'so', 'do', etc. Removal of stop-words is important because they add little value to the model, and often create a dimensionality problem in the model.

4 Classification

The dataset already had a classification of scores based on the text review. However, we decided to further classify the reviews in a binary fashion to create our own 'score' column to avoid the dimensionality problem and prevent any confusion by the model. To do so, we used number 3 as a cut-off point, meaning that we consider ratings 3 and below as negative (bad) and ratings of 4 and 5 as positive (good). To be able to feed this into the model, we assign number 1 to positive reviews and number 0 to negative values. The classification was done by creating dummy variables for each of the ratings. This gave us 5 new dummies. Next, we added the dummy variables representing ratings 4 and 5. As a result we got a binary score column with ratings 4 and 5 represented by 1 and any other rating represented by 0.

asin	reviewText	overall	1.0	2.0	3.0	4.0	5.0	score
B000FA64PK	Another well written eBook by Troy Denning, bu...	3.0	0	0	1	0	0	0
B000FA64PK	This one promises to be another good book. I h...	5.0	0	0	0	0	1	1
B000FA64PK	I have a version of "Star by Star" that does n...	4.0	0	0	0	1	0	1
B000FA64PK	Excellent! Very well written story, very excit...	5.0	0	0	0	0	1	1
B000FA64QO	With Ylesia, a novella originally published in...	2.0	0	1	0	0	0	0

Later on, we re-run our python script without using the binary classification, simply using the original ratings, and compare the results between the two forms of classification.

5 Tokenizing and Splitting

Step 5 of the process is to tokenize and split the data into lists. Using the function `str.split().tolist()` we are able to convert each individual review into a list of strings. All the lists are appended to an empty list, and the result is list of lists.

6 Word Embedding and Sentence Embedding

For the data to be accessible by the algorithm, the data needs to be in numeric form (Bengfort, Bilbro & Ojeda 2018). This means that we need to convert our text format to numeric format. This is done via vectorization, which allows us to represent words and sentences with an array; and is done in two steps, word embedding and sentence embedding. Word embeddings can be done in a number of ways, either by creating one's own model , or with the help of a pre-trained model. Because of our limited time and skills as well as limited computational power on our systems, we opted to use a pre-trained model. The library `gensim` has an in built `word2vec` model, which as the name suggests converts words to vectors of 300 numbers. We used a pre-trained Google news dataset that uses `word2vec` to create this vectorization. Once our tokenized data was vectorized, we normalized the vectors by their magnitudes. Next we created an empty array in which the word embeddings were added to one another to create sentence embeddings for each of the reviews. This allowed us to have a vector of similar length for all the reviews, making the comparison of different reviews much easier. Creating the sentence embeddings this way also took care of the dimensionality problem that may have arose.

7 Machine Learning And Results

The data is now prepared to be fed into ML models. We split the data frame into training (70%) and testing (30%) and applied Logistic Regression, SVM, RFM and Decision Trees. The following table gives a comparison of the results of the models on the binary classification of our data. The table reports the weighted results for precision, recall and f1-score.

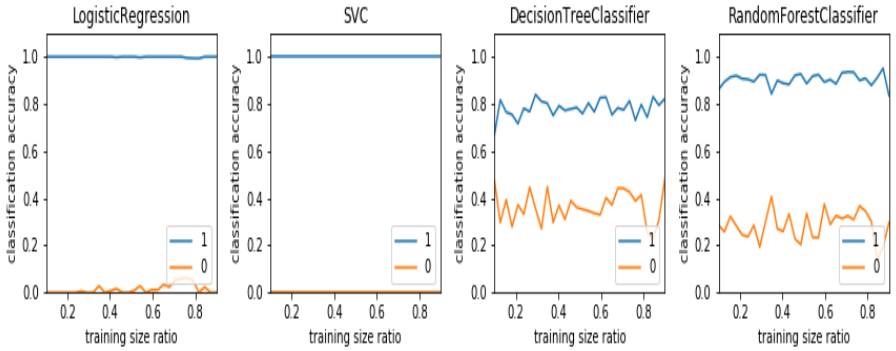
Model	Precision	Recall	f1-score	Accuracy
Logistic Regression	0.83	0.77	0.68	0.77
SVM	0.59	0.77	0.67	0.77
Decision Trees	0.69	0.69	0.69	0.69
Random Forest	0.74	0.77	0.74	0.77

Figure 1: Model Results for Binary Classification

As we can see Logistic Regression offers the highest precision, where 'the precision is intuitively the ability of the classifier not to label as positive a sample that is negative'. For our model, higher precision means that the positive reviews are predicted as positive and not negative, and vise versa. However, Logistic Regression has one of the lowest f1-scores. SVM has a consistently low Precision and f1-score. The recall and accuracy scores are similar for 3 of the 4 models used.

The following figure graphs how different training and testing sizes and how the accuracy of

classifiers fluctuates. The plots show extreme contradictions between the accuracy for negative versus positive comments.



The following table shows the performance statistics of the model once multi-label classification is used.

Model	Precision	Recall	f1-score	Accuracy
Logistic Regression	0.37	0.51	0.37	0.51
SVM	0.26	0.51	0.34	0.51
Decision Trees	0.51	0.47	0.48	0.47
Random Forest	0.42	0.50	0.45	0.50

Figure 2: Model Results for Multi-label Classification

Logistic Regression techniques seem to have lost a lot of its precision level. SVM however continues to give the baseline results.

Moreover, as once can observe, binary classification gives a better prediction in general. However, none of the models stands out in comparison to others.

8 Conclusion

All the models perform more or less similarly within the chosen classification category. This is probably because some of the information is lost when we combine the data to binary classes and also to sentence embedding. In case of multi-label classification, it could be the large dimension of data that creates some level of confusion.

We finally use our developed model to predict the sentiment of a completely new review. We do this using binary random forest model as it is gives us a better prediction than the others. The `test_new_data` function takes a string/review , tokenizes and vectorizes it followed by which it uses the `saved_model.pkl` to predict the the rating of this new review.

Some of the limitations of the project include: a) we did not estimate the results multiple times using cross-validation on our data; b)having limited computational power prevented us from run-

ning the model on the whole dataset; c) last but not least, we plotted different training and testing sizes and compared the accuracy of classifiers. However, the plots show extreme contradictions between the accuracy for negative versus positive comments. Moreover, the accuracy of the models do not show positive correlation with the size of training sample, rather they show fluctuation. For these reasons, finding the perfect training size seems impossible.

Selected References

1. Bengfort, Benjamin, Rebecca Bilbro, and Tony Ojeda. 2018. *Applied Text Analysis with Python: Enabling Language-aware Data Products with Machine Learning.* O'Reilly Media, Inc.

2. He, Ruining, and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web, pp. 507-517. International World Wide Web Conferences Steering Committee*

3. Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5 (1)* : 1-167.

4. McAuley, Julian. Amazon product data.