Machine Learning Engineer Nanodegree

Capstone Proposal

by

Arunprakash Nagarajan

Yelp review sentiment Analysis

Domain Background

Yelp is one of the popular application which provides users with restaurant information. It has got significant amount of data about the restaurants, food, etc. Yelp is also very famous for the user comments. Yelp's business plan revolves around users commenting about the restaurant and food and the ratings which they give. So, the amount of data collected is significantly large.

One of the important reasons for me to try sentiment analysis is mainly because of the scope of the analysis part in such natural language project. This is a part I feel was not much discussed in the Nanodegree and apart from using the ML algorithms, this would give me a good scope to analyze, clean the data which would be generally very dirty.

Doing such a project to analyze with sentiment and loads of reviews, two research papers were immensely helpful. One of them is "Opinion Mining and Sentiment Analysis" (Reference 1) which talks extensively about different ways to handle the text data especially which talks about what others feel. This paper gave me a good insight on what kind of analysis can be made with the data that I have got and how to tell a story with it. Second paper is by Columbia University's Department of Computer science "Sentiment Analysis of Twitter Data" (Reference 2) which takes a specific example of twitter data, its cleaning process and especially how they managed to get its sentiment and its value. With these research references it was very helpful to get an idea on how to go about the problem that I have considered.

One reference that is always helpful for sentiment analysis is the Stanford NLP resources (Reference 3), which discussed broad range of Natural Language Processing techniques.

Problem Statement

One challenge any system which relies on user comments is the sentiment of the review being posted. Not every user has got the time to rate the food and restaurant and write a review so generally users write a feedback or review on the overall experience from which it can be analyzed if it is a positive or a negative review.

It can be ideally treated as a classification problem where model can be trained using

Machine Learning to identify the words used generally for a negative or a positive review. Putting it quantitatively, if a review comes in the system can classify it as either 5 rated which is positive or 1 rated which is negative.

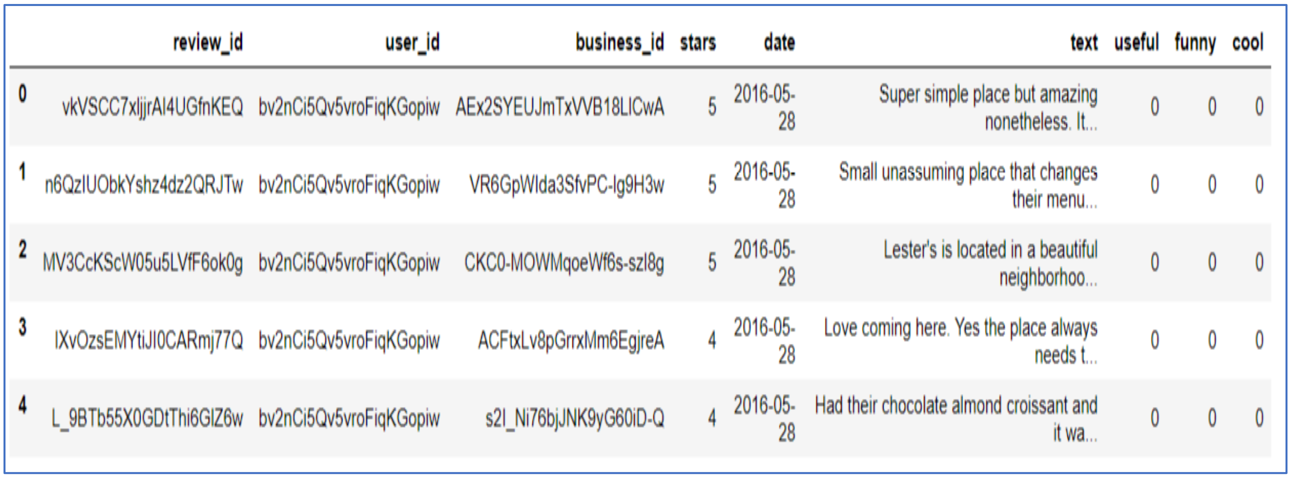
This project is not just about analyzing the sentiments but trying to use the other columns as well for analysis to answer questions like how character length is for different ratings, what are the top words generally used in different ratings, etc.

Dataset and Inputs

I am using dataset which was available in Kaggle (reference 4) by Yelp. I have also uploaded the same in the Dropbox (reference 5) to make things easier. As mentioned before the dataset is extremely huge with more than 5 million records and 9 columns. The main issue with the project is the size of the dataset. So, I have planned to consider only random 10000 records from it and would use records with only the ratings 1(Negative) and ratings 5(Positive) for the sentiment analysis which would be like 4100 records.The dataset description and columns are as below,

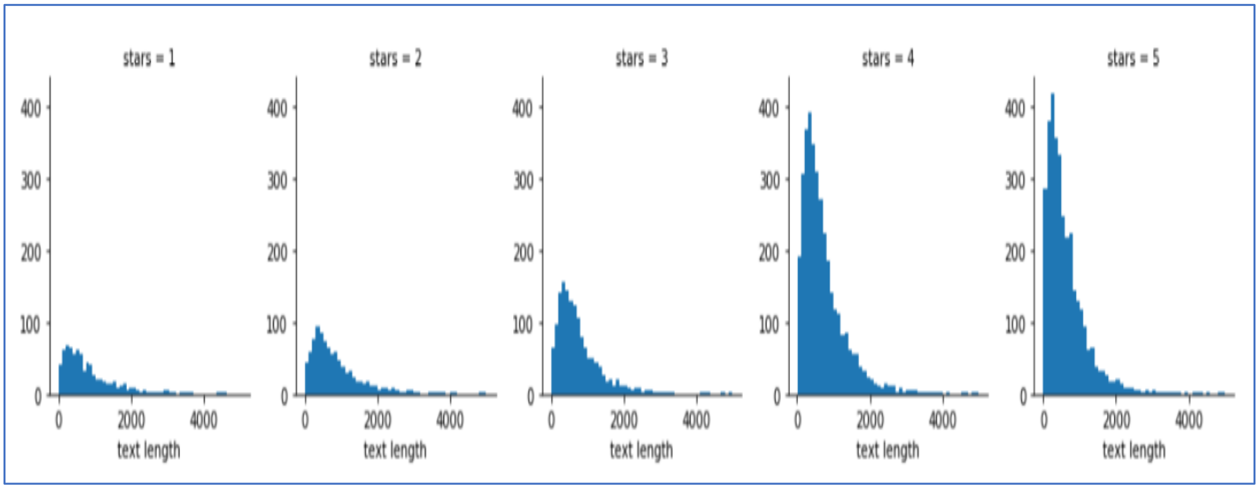
S.No	Column Name	Column Description
1.	“review_id”	ID posted for the review
2.	“user_id”	ID of the user who posted
3.	“business_id”	ID of the business being reviewed
4.	“stars”	Rating stars being given
5.	“date”	Date the review was given
6.	“text”	Actual review text
7.	“useful”	Comments on the review given by the user
8.	“funny”	Comments on the review given by the user
9.	“cool”	Comments on the review given by the user

The below is a sample row from the dataset.



	review_id	user_id	business_id	stars	date	text	useful	funny	cool
0	vkVSCC7xljjrAI4UGfnKEQ	bv2nCi5Qv5vroFiqKGopiw	AEx2SYEUJmTxVVB18LlCwA	5	2016-05-28	Super simple place but amazing nonetheless. It...	0	0	0
1	n6QzIUObkYshz4dz2QRJTw	bv2nCi5Qv5vroFiqKGopiw	VR6GpWIda3SfvPC-lg9H3w	5	2016-05-28	Small unassuming place that changes their menu...	0	0	0
2	MV3CcKScW05u5LVfF6ok0g	bv2nCi5Qv5vroFiqKGopiw	CKC0-MOWMqoeWf6s-szl8g	5	2016-05-28	Lester's is located in a beautiful neighborhoo...	0	0	0
3	IXvOzsEMYtiJI0CARmj77Q	bv2nCi5Qv5vroFiqKGopiw	ACFtxLv8pGrrxMm6EgjreA	4	2016-05-28	Love coming here. Yes the place always needs t...	0	0	0
4	L_9BTb55X0GDtThi6GlZ6w	bv2nCi5Qv5vroFiqKGopiw	s2I_Ni76bjJNK9yG60iD-Q	4	2016-05-28	Had their chocolate almond croissant and it wa...	0	0	0

I also calculated initial text length and it gives an idea of how this considered data is distributed as well with respect to the target column which would be the stars.



SOLUTION STATEMENT

The main issue with the project is the size of the dataset. So, I have planned to consider only the ratings 1(Negative) and ratings 5(Positive) for the sentiment analysis. Although I am planning to put up a deep analysis of the other columns and visualization like the characteristics of the reviews and other columns relationship with the reviews. Analysis in the sense of how text length is for various stars and how it differs for positive and negative reviews. Putting out various plots like heat maps to get a relationship between the different columns would be a very good start in the analysis of the data.

BENCHMARK MODEL

Linear regression model will be implemented to predict if the reviews are either negative (1 rating) or positive (5 rating). This model will be evaluated using accuracy score as the evaluation metric. This model will be used as a benchmark model.

EVALUATION METRICS

Accuracy score will be used for evaluating all the models developed in this project. The confusion matrix and classification report which are available as a part of the sklearn package as well for the model can be executed to get the accuracy score. As seen in the above target class distribution I believe trying out various models and since it going to be either 1 or 5 rating this is a classic classification problem and accuracy would be a key performance metric for this.

Project Design

Below are the steps which I have planned for the project:

1. Import the dataset and do the initial analysis like checking the description and other statistics.
2. Analyze the count of reviews for 1,2,3,4,5 rating stars and store separately.
3. Calculate the character length variation, word cloud and most used words for each of those above stored reviews.
4. Consider only the two ratings data for sentiment analysis which are 1 rated and 5 rated reviews to significantly reduce the dataset size for easier computational purpose.
5. Clean the dataset for sentiment analysis. Cleaning involves, removing punctuations, making uniform case for the letters, performing lemmatization and vectorizing the same.
6. Test train split the data for training and testing set.
7. Choose different algorithms for getting better accuracy and print the classification report.

References

1. *http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf*

   (Opinion Mining and Sentiment Analysis by Bo Pang and Lillian Lee)

2. *http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf*

   (Sentiment Analysis of Twitter Data by Apoorv Agarwal ,Boyi Xie ,Ilia Vovsha ,Owen Rambow, Rebecca Passonneau)

3. *https://stanfordnlp.github.io/CoreNLP/*

4. *https://www.kaggle.com/yelp-dataset/yelp-dataset*

5. *https://www.dropbox.com/s/jmsvpq6sg5ufyee/yelp.csv?dl=0*