SENTIMENT ANALYSIS ON RESTAURANT REVIEWS

Monalisha Ojha
Department of Mathematics
Birla Institute of Technology
Mesra, 835215
monalishaojha974@gmail.com

Arjita Basu
Department of Mathematics
Birla Institute of Technology
Mesra, 835215
arjita.basu@gmail.com

Ankit Tewari
Artificial Intelligence Engineer
Knowledge Engineering and Machine Learning Group
ankit.tewari@estudiant.upc.edu

August 4, 2019

ABSTRACT

The project is on sentiment analysis on a data set retrieved from kaggle. For example, Zomato, one of the largest online food delivery sites receives orders and reviews in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This project provides a way of sentiment analysis using data mining techniques which will process the huge amount of restaurant review data faster. We are going to work on this dataset of reviews and apply an algorithm to extract a meaningful result. The following literature survey illuminates the meaning of sentiment analysis, the various current methodologies of implementation, and the future scope in this domain.

***K*eywords** Machine Learning · Data Analytics · Sentiment Analysis · TF-IDF · Classification · Logistic Regression

1 Introduction

The purpose of this analysis is to build a prediction model to predict whether a review on the restaurant is positive or negative. To do so, we will work on Restaurant Review dataset, we will load it into predicitve algorithms Multinomial Naive Bayes, Bernoulli Naive Bayes and Logistic Regression. In the end, we hope to find a "best" model for predicting the review's sentiment.

Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. It is the branch of machine learning which is about analyzing any text and handling predictive analysis.

Sites such as Zomato have a star rating system that lets users easily see what the general opinion about a particular establishment is without having to read all the reviews for that particular restaurant. However, there is an abundance of user generated information online from social media platforms like twitter, facebook or blog posts where users express their sentiment about a particular product or business. Since the reviews data on the Yelp dataset provides both the star rating and the text for each review, I wanted to know if it would be possible to train a machine learning model with this data in order to identify if a particular business review was positive or negative based only on its text.

A Naive Bayes algorithm was used to build a binary classification model that would predict if the review's sentiment was positive or negative. A Naive Bayes classifier assumes that the value of a particular feature is independent of the value of any other feature, given the class variable. It uses the training data to calculate a probability of each outcome based on the features. One important characteristic of the Naive Bayes algorithm is that it makes naive assumptions about the data. It assumes that all the features in the dataset are independent and equally important.

2 Dataset

In this paper, we have used RRestaurant devieatwsaset.The data is consisted of 1000 observations and 3 features. For the initial pre-processing, we has inspected each feature of the dataset to 1) remove features with frequent and irreparable missing fields or set the missing values to zero where appropriate and 2) remove irrelevant or uninformative features or duplicate features . The team has split the data into train, validation, and test sets. Since the dataset is relatively large ,first 10 data was deemed sufficient for testing and validation sets. Consequently, several feature selection techniques were used to find the features with the most predictive values to both reduce the model variances and reduce the computation time.

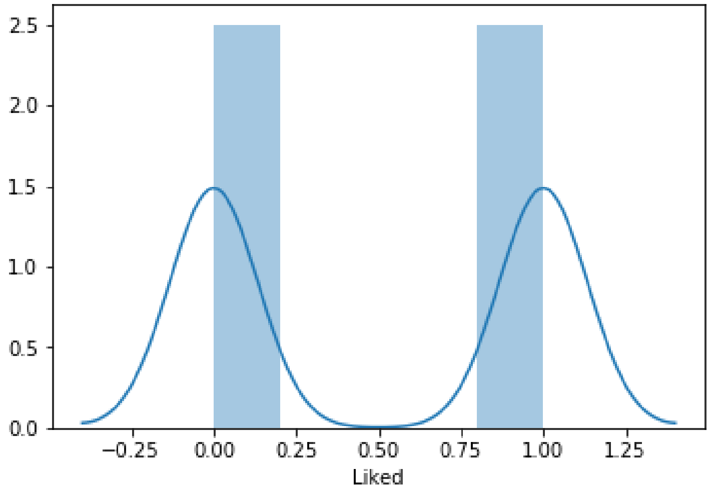


Figure 1: Dataset

3 Feature Selection and Data Preprocessing

Here, we will discuss various steps used to clean the data and understand the relationship between variables and use this understanding to create better features.In order to build my classifier, it was necessary to identify the features that the model would use. Since I will be training the model only on review text, my features will be the words or sequence of words of these reviews. I decided to explore different models with different combinations. I would first try a model in which each particular word would be a feature. I also used n-gram combinations. An n-gram is a contiguous sequence of n items from a given sequence of text or speech.

3.1 User Reviews

We first looked at the distribution of user reviews and found that there were 50 records where the user reviews were null values. Since, user reviews cannot be null we removed such instances from the data.

3.2 Sentiment

There were 10 negetive values which needs to be removed from the dataset in order to do sentiment analysis.

4 Sentiment Analysis

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

Precursors to sentimental analysis include the General Inquirer, which provided hints toward quantifying patterns in text and, separately, psychological research that examined a person's psychological state based on analysis of their verbal behavior.

Subsequently, the method described in a patent by Volcani and Fogel, looked specifically at sentiment and identified individual words and phrases in text with respect to different emotional scales. A current system based on their work, called EffectCheck, presents synonyms that can be used to increase or decrease the level of evoked emotion in each scale.

Many other subsequent efforts were less sophisticated, using a mere polar view of sentiment, from positive to negative, such as work by Turney, and Pang who applied different methods for detecting the polarity of product reviews and movie reviews respectively. This work is at the document level. One can also classify a document's polarity on a multi-way scale, which was attempted by Pang and Snyder among others: Pang and Lee expanded the basic task of classifying a movie review as either positive or negative to predict star ratings on either a 3- or a 4-star scale, while Snyder performed an in-depth analysis of restaurant reviews, predicting ratings for various aspects of the given restaurant, such as the food and atmosphere (on a five-star scale).

First steps to bringing together various approaches—learning, lexical, knowledge-based, etc.—were taken in the 2004 AAAI Spring Symposium where linguists, computer scientists, and other interested researchers first aligned interests and proposed shared tasks and benchmark data sets for the systematic computational research on affect, appeal, subjectivity, and sentiment in text.

5 Architecture

I cleaned the data by removing numbers, punctuation characters, stopwords (extremely common words that have little value in helping with classification) and separators from the corpus. I also removed some terms that I identified would confuse the model, words like beer, chicken, burger, cookie, drink, cafÃ©, dessert would not be relevant for classification so I eliminated them from the corpus.

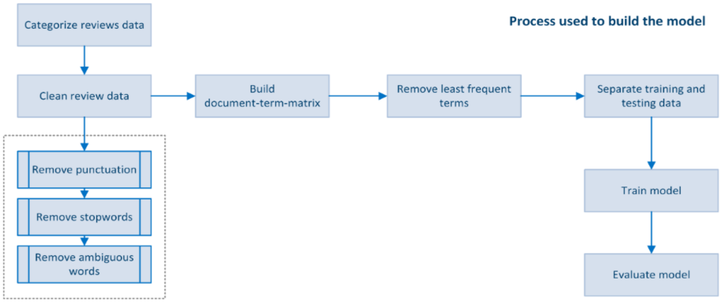


Figure 2: Model Architecture

6 Methods

Linear Regression was set as a baseline model on the dataset using all of the features as model inputs. After selecting a set of features using Lasso feature selection, several machine learning models were considered in order to find the optimal one. All of the models were implemented using scikit-learn library.

6.1 KNN

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (for e.g. distance function). The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, i.e. the value of k, accuracy might increase.

6.2 Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Mathematically, logistic regression estimates a multiple linear regression function defined as:

reg.png

$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

When selecting the model for the logistic regression analysis, another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase the amount of variance explained in the log odds (typically expressed as (R)).

6.3 Naive Bayes

Naive Bayes is a simple but powerful algorithm for predictive modelling.

Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

th.PNG

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

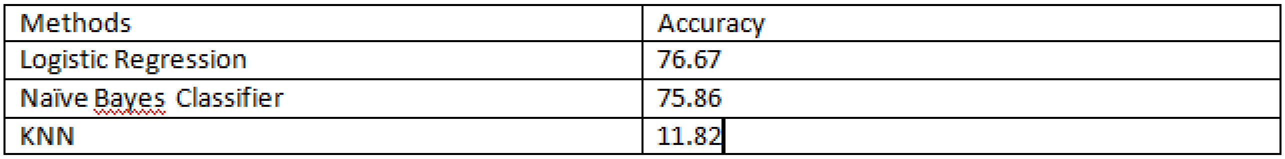
Where

- P(A|B) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(B|A) is the probability of data d given that the hypothesis h was true.
- P(A) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(B) is the probability of the data (regardless of the hypothesis).

We are interested in calculating the posterior probability of P(A|B) from the prior probability p(A) with P(B) and P(B|A).

7 Results

Using Logistic Regression and NLP , we have done sentiment analysis on restaurant reviews to predict if the restaurant is good or bad. The following table represent the accuracy score.

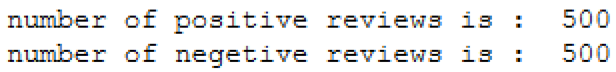


Methods	Accuracy
Logistic Regression	76.67
Naïve Bayes Classifier	75.86
KNN	11.82

Figure 3: Result Table

8 Conclusions and Future work

Considering what is and what is not accounted for in the models built in this study, their predicting results are fairly accurate. To further improve the sentiment analysis accuracy, more variabilities need to be considered and modeled.



```
number of positive reviews is :  500
number of negetive reviews is :  500
```

Figure 4: Result

This project attempts to come up with the best model for sentiment analysis based on user reviews, rantings . Machine learning techniques including Logistic Regression, Tree-based models , k nearest neighbors, Random Forest Classifiers along with feature importance analyses are employed to achieve the best results in terms of Mean Squared Error, Mean Absolute Error, and R2.The initial experimentation with the baseline model proved that the abundance of features leads to high variance and weak performance of the model on the validation set compared to the training set. This level of accuracy is a promising outcome given the heterogeneity of the dataset and the involved hidden factors , which were impossible to consider.

We have identified a couple of area where we can make improvements. Currently there are some steps that we need to perform manually. The future works on this project can include (i) studying other feature selection schemes such as Random Forest feature importance, (ii) further experimentation with neural net architectures .

9 References

[1] Sentiment Analysis - https://www.kaggle.com/apekshakom/sentiment-analysis-of-restaurant-reviews [2] Python | NLP analysis of Restaurant reviews https://www.geeksforgeeks.org/python-nlp-analysis-of-restaurant-reviews/

[3] http://cs229.stanford.edu/proj2018/report/96.pdf

[4] https://scikit-learn.org/stable/, scikit-learn Machine Learning in Python.

[5] NLP for Review https://www.kaggle.com/monalisha/nlp-for-review/edit