NAME OF THE PROJECT

Rating Prediction Project

Submitted by:

Varsha Vasant Shinde

ACKNOWLEDGMENT

We first would like to thank god for giving us the opportunity to be here and make this project. Secondly, we thank our families for their unwavering support and great sacrifices in order to help us reach this moment. We would also like to thank the faculty of Mohd Kashif sir and FLIP ROBO for their work and efforts to give us the best education possible in order to reach our full potential, especially our SME Mr.Mohd Kashif for his guidance and assistance in not only this but Throughout our entire journey

INDEX

INTRODUCTION

In today's digital world, most of the consumers prefer e-commerce, because of the lucrative offers, but primarily because they have a review and feedback system to judge the product. It is also now a common practice amongst customers to post reviews about a product they purchase, be it positive or negative. Such reviews provide valuable feedback on these products, which may further be used by potential customers to find opinions of existing users before deciding to purchase a product.

They are also used by product manufacturers to identify strengths and problems in their products and to find competitive intelligence, such as its potential worth in the market. It makes them susceptible to change based on customer feedback, such as, product with good and a high feedback will result in high sales, thus, intimating the manufacturers of providing a slightly lesser discount to enhance sales. Products with slightly negative or apathetic reviews will result in a slump of sales, making the manufacturers provide a slightly higher discount.

Due to the reason of profit or fame, spammers promote or demote a target product. They post fake reviews or malicious opinions, which misleads the potential buyer product manufacturers and market researchers. In this paper, we make an attempt to detect spam and fake reviews, and filter out expletives, vulgar and curse word, by incorporating sentiment analysis. We match the rating published with the calculated ratings of each review, by producing a sentiment score with the help of an in-house dictionary.

The reviews, whose difference in both the ratings is higher than a threshold value, are considered to be spam. Since there are millions of products and customers, and a large review data set, it becomes difficult for the customers and company personnel to track each review and identify the customer sentiment. So, we categorically divide the product reviews based on the products' features. Each feature and sub-feature is categorized as per how negative or positive it is, and gives the company personnel an indication.

Business analysts analyze the categories and deduce which features contribute more to the sales of the products, and which features add to its declination. Finally, we graphically show and analyze the different features of the product which adds to its popularity or demotion.

Many e-commerce web sites enable their customers to write product reviews and feedback in the form of ratings. This gives the company personnel an indication about their products' standing in the market, while also enabling fellow customers to form an opinion and help purchase a product. However, due to the reason of profit or fame, many target products are promoted or demoted in the form of spam.

It may contain fake reviews or malicious opinions, which is misleading. In this paper, we make an attempt to detect spam and fake reviews, and filter out reviews with expletives, vulgar and curse words, by incorporating sentiment analysis. Other studies solve this by using just the ratings as a parameter.

This paper, however, by taking upon consideration Amazon dataset, matches the posted rating with the calculated ratings of each review, by producing a sentiment score with the help of an in-house dictionary. Finally, we graphically show and analyze the different features of the product which adds to its popularity or demotion.

Motivation for the Problem Undertaken

Product reviews help consumers understand whether a product's quality and specifications will meet their expectations. Reviews can also give consumers a clearer picture of what to expect when their product arrives, thereby helping to reduce the rate of returns.

Analytical Problem Framing

Data Gathering

The opinions expressed in product reviews provide valuable information to online retailers about their effective gain or loss, and standing worth in the market. For our paper, we take a dataset of Amazon Electronic product reviews. To retrieve the dataset, we have used three information retrieval techniques-

a. Using the tool WebHarvy web scraper,
b. Crawl data by developing a code,
c. Collecting data sets available open-source online.

Pre-Processing and Abusive Reviews Filtering

Online informal text of product reviews may involve expletives and curse words. It requires more sophisticated methods to clean noise in raw text. We have created a 'bag of words', containing around 350 words of expletives, and some other non-relevant words mentioned below. This is known as product feature extraction.

Sentiment Lexicon Analysis

We are crawling Amazon reviews dataset wherein different electronic products are rated on a scale of 1 to 5. We then develop a dictionary of our own with words which might define sentiment in a review and have some weightage

Other previous studies just used SentiWordNet. Our study however, uses an in-house developed Dictionary. We divide the review sentence into individual tokens.

Sentiment Score

Sentiment Score:

- ✓ The sentiment score shows a review's sentiment polarity.
- ✓ That is, the degree of how good or bad a review is.
- ✓ We calculated the sentiment score by incorporating the sentiment weightage of each word in the review.
- ✓ The score will be in the scale of -1 to 1.
- ✓ Eg- consider the review"good camera and awesome battery life". Here, "good" and "awesome", both are the adjectives showing some sentiment.

The sentiment score is calculated by using the equation:-

$$\text{Score}(r) = \sum_{i=1}^{i=N} \{(F) * (W)\} / L$$

Where, N= number of different Sentiment words in the review.

- **F**= frequency of the sentiment word in a review
- **W**= weightage of the sentiment word in a review
- **r** = rth review
- **L**= square root of length of unique tokens

Negation of Feature Word Score:

Words such as adjectives and verbs are able to convey the opposite sentiment with the help of negative prefixes. So, there is a need to check the presence of any negative word before the sentiment word and then change its polarity to get the correct score. Else, it would consider it a good review and give it a positive score. For this problem, we have implemented an algorithm as-

- ✓ Check the index number of the sentiment word in the review string,
- ✓ Store two indexes before the index of the sentiment word,
- ✓ Check the word store in two indexes with the negative words dictionary,
- ✓ If the word is present, negate the polarity of the sentiment word,
- ✓ Repeat this step for each sentiment word in the review string.

In an instance review like “The built-in speaker also has its uses but so far nothing revolutionary”, the token “revolutionary” signifies a positive sentiment. However, the prefix “nothing” negates this sentiment, thus, making it negative. Therefore, identifying negative phrases is very important.

Model/s Development and Evaluation

The train-test split is a technique for evaluating the performance of a machine learning algorithm.

It can be used for classification or regression problems and can be used for any supervised learning algorithm.

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

- **Train Dataset**: Used to fit the machine learning model.
- **Test Dataset**: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

This is how we expect to use the model in practice. Namely, to fit it on available data with known inputs and outputs, then make predictions on new examples in the future where we do not have the expected output or target values.

The train-test procedure is appropriate when there is a sufficiently large dataset available.

Machine learning aims to develop an algorithm in order to optimize the performance of the system by using example data. The solution that machine learning provides for sentiment analysis involves two main steps.

1. The first step is to "learn" the model from the training data
2. The second step is to classify the unseen data with the help of the trained model Machine learning algorithms can be classified in deferent categories:

- ✓ supervised learning
- ✓ semi-supervised learning
- ✓ unsupervised learning

Supervised learning

The process where the algorithm is learning from the training data can be seen as a teacher supervising the learning process of its students. The supervisor is somehow teaching the algorithm what conclusions it should come up with as an output. So, both input and the desired output data are provided. It is also required that the training data is already labelled. If the classifier gets more labelled data, the output will be more precise.

The goal of this approach is that the algorithm can correctly predict the output for new input data. If the output were widely different from the expected result, the supervisor can guide the algorithm back to the right path. There is however some challenges involved when working with supervised. The supervised learning works fine as long as the labelled data is provided. This means that if the machine faces unseen data, it will either give wrong class label after classification or remove it because it has not ”learnt” how to label it .

Unsupervised learning

In difference with supervised learning is trained on unlabeled data with no corresponding output. The algorithm should find out the underlying structure of the data set on its own. This means that it has to discover similar patterns in the data to determine the output without having the right answers. One of the most important methods in unsupervised learning problems is clustering. Clustering is simply the method of identifying similar groups of data in the data set .For sentiment classification in an unsupervised manner it is usually the sentiment words and phrases that are used. This means that the classification of a review is predicted based on the average semantic orientation of the phrases in that review. This is obvious since the dominating factor for sentiment classification is often the sentiment words. This technique has been used in Turney’s study (2002).

Semi-supervised learning

Which has the benefit of both supervised and unsupervised learning refers to problems in which a smaller amount of data is labelled, and the rest of the training data set is unlabeled. This is useful for when collecting data can be cheap but labelling it can be time consuming and expensive. This approach is highly favourable both in theory and practice because of the fact that having lots of unlabeled data during the training process tends to improve the accuracy of the final model while building it requires much less time and cost. In a semi-supervised learning was experimented where they used 2000 documents as unlabeled data and 50 randomly labelled documents.

Naive Bayes

Naive Bayes is another machine learning technique that is known for being powerful despite its simplicity. This classifier is based on Bayes theorem and relies on the assumption that the features (which are usually words in text classification) are mutually independent. In spite of the fact that this assumption is not true (because in some cases the order of the words is important), Naive Bayes classifiers have proved to perform surprisingly well. The first step that should be carried out before applying the Naive Bayes model on text classification problems is feature extraction.

CONCLUSION

In this paper, we incorporate sentiment analysis of reviews techniques into the review detection. First we have made our own dictionary having sentiment words along with the weight given to the word according to its polarity.

Then a method is proposed to calculate the sentiment score of the reviews from the natural language text by a shallow dependency parser. A set of discriminative rules are presented through intuitive observation.

The discriminative rules are combined with the time series method to find the spam and fake reviews. Then the experimented case study and dataset demonstrate the efficiency of our proposed method. In future we would try to improve the method of calculating the sentiment score of the reviews.

We would also try to update our dictionary containing sentiment word. We would try to add more words in our dictionary and update the weights given to those words to get more accurate calculated score of the reviews.