An Application of Text Classification to Check a Product Review

A Thesis Proposal

Submitted to the Institute of Information Technology, Jahangirnagar University, Savar, Dhaka in partial fulfillment of the requirements for the 1st semester, Masters of Science degree in IIT.

By

Shuvashish Paul Sagar

(Exam Roll: 180194)

Supervised By

MD. FAZLUL KARIM PATWARY

Professor
Institute of Information Technology
Jahangirnagar University

Institute of Information Technology

Jahangirnagar University

April 29, 2019

DECLARATION

This project proposal is submitted to the Institute of Information Technology, Jahangirnagar University, Savar, Dhaka in partial fulfillment of the requirements for having the Masters in IIT. This is also needed to certify that the project work is under the Masters of Science course of the Information Technology.

So, I, here by, declare that this project proposal has not been submitted elsewhere for the requirement of any kind of degree, diploma or publication.

Shuvashish Paul Sagar

(Exam-roll: 180194)

Certificate

This is to certify that the thesis on “An Application of Text Classification to Check a Product Review” is a confide record of thesis work done by Shuvashish Paul Sagar (Exam Roll - 180194) for partial fulfillment for the requirements to the Institute of Information Technology of “M.Sc in Information Technology

Abstract

Consumer reviews on online products plays a vital role in selection of a product. In recent years the online marketing is rapidly growing. The customer reviews are the measurement of customer satisfaction. This review data in terms of text can be analyzed to identify customer's sentiment and their demands too. Many researches over sentiment analysis have been proposed in recent years. The main purpose of this work is to help the developers identifying whether a review is a positive or negative one. In this work we will propose a way to collect data set from various e-commerce sites and classify them with the most accurate classifying algorithm. Then we would like to classify an unknown text with the known data set and comment the outcome. The comment would be a boolean set – either true or false.

Keywords: Data mining, Text mining, Sentiment analysis, Classification

Table of Contents

List of Figures

List of Abbreviations

KNN	**K- Nearest Neighbor**
DM	**Data Mining**
SVM	**Support Vector Machine**
SA	**Sentiment Analysis**

CHAPTER 1

Introduction

1.1 Generic

The rapid growth of internet technology results in an increased number of data set. People are more likely to purchase products from online business sites now-a-days. They also post their comments and satisfaction in internet. User's satisfaction analysis involves in data mining process and techniques to identify the sentiment or likelihood to purchase a product. It would be very necessary for the e-commerce business sites to know user's satisfaction level. Now-a-days people are more likely to post their expression about anything in internet. Online statement presents the actual mind state of a person. Analyzing this will be our main goal.

1.2 Limitation

In some e-commerce sites people do not post texts in English while they prefer the local language. This would be difficult to analyze with this proposed system. So, all the texts we are dealing with is English.

In some case there is no review in a product. This would be also a limitation of us. We can not comment on the product if there is no review. It can be a new product added to the site or such a kind of product where users are not likely to put comments about the product.

1.3 Motivation

Data mining is one of the most interesting topics in research. The lots of data can be used for further work with data mining.

Our text mining research work will be able to re-cycle unused data to work with it in supervised learning mechanism. With this research work we can use the unused data of user review section. With this dataset we will enrich our dataset and thus classifier model. This purpose attracts the researchers much.

1.4 Objective

In this paper, we will propose a system that would collect data from various e-commerce sites using a web crawler bot. This collected data would be used as a data set. Then we would like to classify them based on the most accurate classification algorithm. Based on the best classifier we will check a given product whether it is good or not.

Our main objective of this paper work is to propose a model that would classify a given text. We will apply it into an e-commerce site's products and check the reviews. Then we will comment on the product whether it is positive or negative. We will also collect the predicted data and save it to our dataset. In this way we will enrich our dataset thus machine. So this is a machine learning process where we would update our dataset gradually.

CHAPTER 2

Literature Review

Researchers are very much interested in text mining as the use of e-commerce business have been growing rapidly in recent years. In this section we are providing an overview of the studies regarding text mining and sentiment analysis.

Some researches are based on classifying different datasets with different suitable algorithm [i]. J. Xia, F. Xie, Y. Zhang, and C. Caulfield discussed on some algorithms and its application in their research work entitled as “Artificial intelligence and data mining: algorithms and applications” [ii]. In the research paper of W.H. Inmon there is a process to store the data set in the data warehouse [iii].

A lot of researches are done with classification algorithms and its applications. They also discussed on some regression and correlation matches within different data sets [v][vi][viii][ix][x].

Working with different datasets are comparatively more common work in research field. Some researches deal with sentiment analysis. These works collect user’s posts from different social media and classify them. Then they can analyze the new data matching with the known data set [vii].

CHAPTER 3

Methodology

3.1 Proposed Methodology

In this paper work we will discuss on building a web crawler bot that will collect review data from different e-commerce sites. We will use this data as a data set for our classification algorithm. We will choose an e-commerce site to extract user reviews on particular type of products.

To build a web crawler bot we need to specify a particular site first. All the sites are not designed in the same pattern. We need to find out the pattern of the site we have specified. The e-commerce sites are basically designed in a standard pattern. There is a page where we can find all the available categories of products of the site. We need the URL of each category. After this step we can store all the available category links in a database. Then we will crawl the pages one by one. While crawling the category pages we will find the product pages. Our main concern of this research work is the product pages. Her we will extract the user review section. We collect the user reviews and store the user reviews against the category and the product.

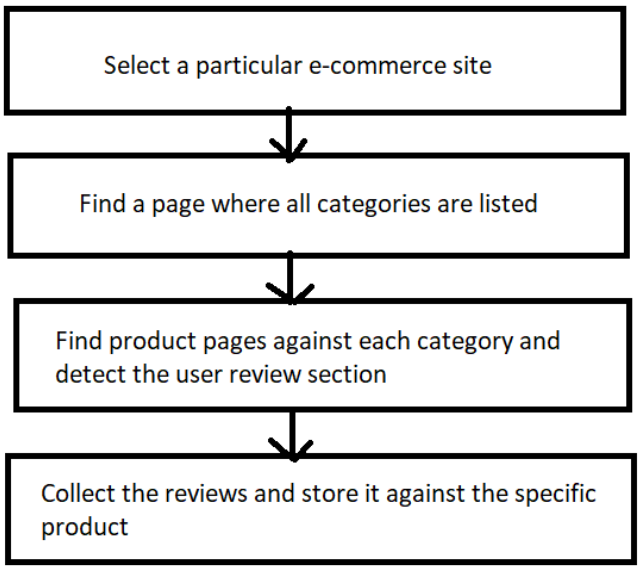

Fig 3.1.1: Steps to collect the user review data set

Now with this dataset we need to run some classification algorithms and find out the accurate one with our dataset.

Classification is basically a model that separates data into different distinct classes. To build this model we need huge set of training data. Our collected data set will be used to build the model. This model will be then used to predict other reviews. In this step basically we will build the model.

In this section we will make the following assumption:

- We will use our collected dataset as training dataset
- Omit any pre-processing of the dataset
- Ignore the presence of categorical features as it would be turned into an integer value later
- Also ignore the null values
- Finally, we will use conventional training mechanism to build the model

3.2 Algorithms

Each single row of a dataset is classified according to its similarities. Among so many data mining methods classification is the best known and popular. The aim of classification is to predict an unknown object with some known parameters of related other objects. In this section we will discuss some algorithms needed to classify a dataset.

Naïve Bayes: Naive Bayes is a probabilistic technique for constructing classifiers. The characteristic assumption of the naive Bayes classifier is to consider that the value of a particular feature is independent of the value of any other feature, given the class variable.

Despite the oversimplified assumptions mentioned previously, naive Bayes classifiers have good results in complex real-world situations. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification and that the classifier can be trained incrementally.

Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities for each of K possible outcomes or classes.

$$p(Ck|x1,.....,xn)p(Ck|x1,.....,xn)$$

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it simpler. Using Bayes theorem, the conditional probability can be decomposed as −

$$p(Ck|x)=p(Ck)p(x|Ck)p(x)p(Ck|x)=p(Ck)p(x|Ck)p(x)$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is −

$$p(Ck|x1,.....,xn)=1Zp(Ck)\prod i=1np(xi|Ck)p(Ck|x1,.....,xn)=1Zp(Ck)\prod i=1np(xi|Ck)$$

where the evidence Z = p(**x**) is a scaling factor dependent only on x_1, …, x_n, that is a constant if the values of the feature variables are known. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label y^=Cky^=Ck for some k as follows –

$$y^=argmaxp(Ck)\prod i=1np(xi|Ck)$$

K-nearest neighbor: KNN is used for both classification and regression predictive problems. Here 3 aspects are considered:

- Ease to interpret output
- Calculation time
- Predictive Power

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

Decision Tree: Decision trees are a type of model used for both classification and regression. Trees answer sequential questions which send us down a certain route of the tree given the answer. The model behaves with "if this than that" conditions ultimately yielding a specific result. This is easy to see with the image below which maps out whether or not to play golf.

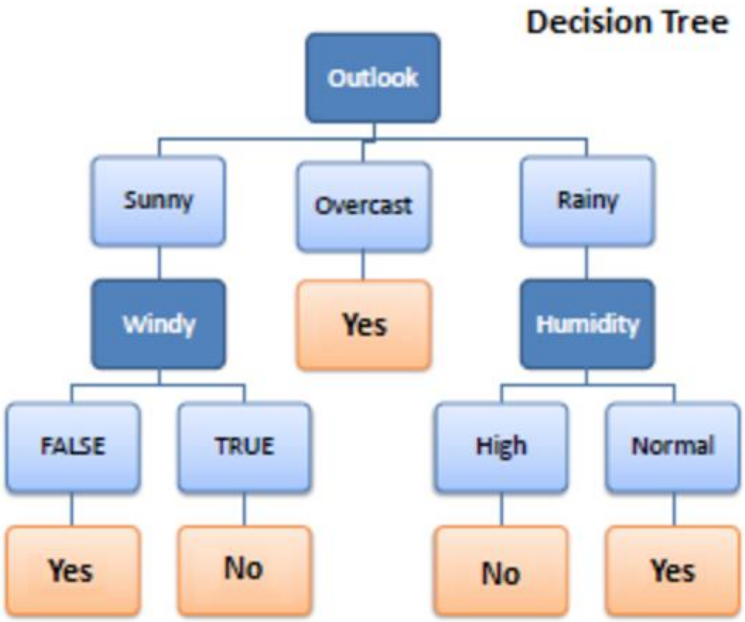

Fig 3.2.1: An example of decision tree

The flow of this tree works downward beginning at the top with the outlook. The outlook has one of three options: sunny, overcast, or rainy. If sunny, we travel down to the next level. Will it be windy? True or false? If true, we choose not to play golf that day. If false we choose to play. If the outlook was changed to overcast, we would end there and decide to play. If the outlook was rainy, we would then look at the humidity. If the humidity was high, we would not play, if the humidity is normal, we would play.

Support Vector Machine (SVM): A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

CHAPTER 4

Expected Outcome

The value of the criteria would be calculated with a confusion matrix.

		Precision class	
		A	b
Actual class	a	TP	FN
	b	FP	TN

Fig 4.1: Confusion Matrix

TP = true positive

FN = false negative

FP = false positive

TN = true negative

Accuracy (AC): The percentage of correct predictions. It can be calculated as follows according to the confusion matrix.

$$AC = (TN+TP)/(TP+FN+FP+TN)$$

Precision (P): the fraction of correctly predicted positive observations among the total predicted positive observations.

$$P = TP/(TP+FP)$$

Recall(R): The correctly predicted observation among all observations. It can be calculated as:

$$R = TP/(TP+FN)$$

F-measure: The Precision and Recall criteria can be interpreted together rather than individually. To accomplish this, we consider the F-Measure values generated by the harmonic mean of the Precision and Recall columns, as the harmonic mean provides the average of two separate factors produced per unit. Therefore, F provides both the level of accuracy of the classification and how robust (less data loss) it is:

$$F\text{-measure} = 2*P*R / (P+R)$$

ROC area: the ROC field curve determines the predictive performance of the different classification algorithms. The area under the ROC curve is one of the essential evaluation criteria used to select the best classification algorithm. When the area under the curve is approaching 1, it indicates that the classification was carried out correctly.

Calculating the statistical terms, we would find which of the classification algorithm suits the model most.

CHAPTER 5

Conclusion & Future Work

We will collect the dataset first through the proposed API. Then we would build the model. Using this model, we will classify the unknown comments come from the users. Hopefully this will accurately. This is our text mining proposed system. In this research work we will work with different models and find the accuracy of each one.

In future we will work with the Bengali text detection and hopefully other local languages too. We will work the slang or bad words detection too. It would be great if we can find that which are fake comments or which are real.

References

i) R. Arora and S. Suman, “Comparative analysis of classification algorithms on different datasets using WEKA,” International Journal of Computer Applications, vol. 54, no. 13, pp. 21–25, 2012.

ii) J. Xia, F. Xie, Y. Zhang, and C. Caulfield, “Artificial intelligence and data mining: algorithms and applications,” Abstract and Applied Analysis, vol. 2013, Article ID 524720, 2 pages, 2013.

iii) W. H. Inmon, Building Data Warehouse, QED/Wiley, Hoboken, NJ, USA, 2005.

iv) J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Burlington, MA, USA, 2012.

v) Ghosh, M., & Sanyal, G. (2018). Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis. Applied Computational Intelligence and Soft Computing, 2018.

vi) D. Donko and A. Dzelihodzic, “Data mining techniques for credit risk assessment task,” Recent Advances in Computer Science and Applications, pp. 105–110, 2013.

vii) M.Rambocas, and J. Gama, “Marketing Research: The Role of Sentiment Analysis”. The 5th SNA-KDD Workshop’11. U nivers ityof Porto, 2013

viii) Jiang, M.; Liang, Y.; Feng, X.; Fan, X.; Pei, Z.; Xue, Y.; Guan, R. Text classification based on deep belief network and software regression.

ix) S. Belkıs, Credit Rating, Capital Market Licensing Registration and Training Organization, Turkey, 2016.

x) A. Wang, L. Yong, W. Zeng, and Y. Wang, “The optimal analysis of default probability for a credit risk model,” Abstract and Applied Analysis.