Optimising Power Consumption of Home Appliances using Machine Learning

Shreyas S

Department of Masters of Computer Application

RNS Institute of Technology

Bangalore, Karnataka, India

hello@shreyas.im

Dr. Chetana Hegde

Department of Masters of Computer Application

RNS Institute of Technology

Bangalore, Karnataka, India

chetanahegde@ieee.org

*Abstract*—Natural Language Processing is a branch of Machine Learning that deals with the interaction of computers and humans, especially with respect to how computers analyze, understand and derive the sentiment and meaning of sentences spoken by humans. The process of computers understanding the sentiments in a sentence is known as Sentiment Analysis. There are a lot of Machine Learning Classification algorithms that is used for Sentiment Analysis whose performances vary with different sets of data. In this paper, we are going to compare the performances of six different algorithms – Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Linear Kernel Support Vector Machine, Decision Tree and Random Forest on Amazon Review’s data. We present results of each algorithm by showing in what way and by how much each algorithm is better (or worse) compared to the rest of the algorithms.

Keywords—Machine Learning; Natural Language Processing; Sentiment Analysis; Classification; Amazon Reivews;

# Introduction

Natural Language Processing is one of the dominant fields in Machine Learning which deals with how a computer program or a computer software analyzes, understands and derives the sentiment and true meaning of sentences spoken by humans.

As the amount of information and data is increasing exponentially on the internet, this human-computer interaction becomes increasingly important and enables real world applications like automatic text summarization, sentiment analysis, fighting spam, topic extraction, named entity recognition, parts of speech tagging, relationship extraction and more.

Sentiment Analysis is a great example for application of Natural Language Processing. The goal here is to identify the emotions and sentiments of statements. It is especially helpful for companies to understand what their customers think about their products.

There are a lot of Machine Learning Classification Algorithms that can be used for Sentiment Analysis. Here, we are going to look at how six different classification algorithms perform compared to each other. Performance of each algorithm is computed using Confusion Matrix, Accuracy, False Acceptance Rate (FAR), False Rejection Rate (FRR) and vertical bar graph.

Amazon Reviews are a great source of data for this project. Over 2400 reviews of different mobile phones have been used to train the Machine Learning Models and determine the performances of each Algorithm.

The paper is organized as follows: Section II contains literature survey, Section III deals with architecture and modelling, Section IV presents the implementation and pseudo code, Section V deals with Performance Analysis where every algorithm’s performance is measured and compared and Section VI concludes the work.

# Literature Survey

There are a lot of Machine Learning algorithms which can be used to work on Sentiment Analysis problems. There have been various papers on Sentiment Analysis.

Chantal Fry et al. [1] have performed a case study analysis on the performance of K-Means and Peak-Searching clustering algorithms on Amazon Product Review data. When it comes to grouping similar reviews based on topics, the results show that K-Means algorithms outperforms Peak-Searching algorithm. Maria et al. [2] have compared lexicon dictionary based approach with n-grams with three different machine learning algorithms to predict positive and negative sentiments on Amazon’s Product Reviews. Results prove that lexicon based approach outperforms the other techniques. Xing Fang et al. [3] based on Sentiment Analysis using Amazon Product Review data tackle the problem of sentiment polarity categorization. Filipe N Riberio et al. [4] have compared different sentiment analysis methods using multiple datasets originating from different data sources. Michael Crawford et al. [5] have made a comprehensive comparative study of current research on detecting spam reviews using Big Data Analytics. David Garcia et al. [6] have shown that the frequency of word use is not only determined by word length and average information content, but also by its emotional content. Kudakwashe Zvarevashe et al. [7] have proposed a framework called sentiment polarity that extracts unbiased opinions of hotel services from reviews that are automatically prepared. L Arockiam et al. [8] have proposed a framework that collects, pre-processes and analyses the sentiment present in the data. Monali Bordoloi et al. [9] have proposed an effective sentiment analysis model that uses keyword extraction and polarity assignment techniques using graph based approach. The proposed model performs better than a few models that already exist. M Sivakumar et al. [10] have proposed a new method of analysing online student feedback collected from twitter API by measuring semantic relatedness between aspect word and student opinion sentence.

In this paper, we are going to be making use of six different classification problems and see which algorithm performs the best with Amazon Review’s data. The algorithms used are Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Linear Kernal SVM, Decision Tree and Random Forest. Naïve Bayes uses a method called as Bayes theory that is based on probability and statistical knowledge. [11] K-Nearest Neighbor finds the K nearest neighbors from a given set of queries. [12] Support Vector Machine and Linear Kernal SVM are supervised learning algorithm that revolve around a notion of margin – 2 data classes separated by either side of a hyperplane. [13] Linear kernal is very similar except has more flexibility when it comes to penalties and loss functions. Decision Tree is another supervised machine learning algorithm that finds the target result by mapping observations about a data. [14] Random Forest combines a set of decision trees by selecting random samples from a given data. It relies on the values of a random vector sampled independently and distribution for all trees. [15]

# Architecture and Modelling

After scraping the review data from Amazon, we need to clean the data first so that there are no errors. Each review is cleaned using various techniques such as removing stop words, making all the words lowercase and stemming words and then the Bag of Words model is created. The data is then split into Training Set and Test Set. Machine learning model is then trained using a classification algorithm by applying it to the Training Set. After this, the model is ready for predictive analysis. The Test Set is then given to the trained model and predictions are made. Finally, results are analyzed by comparing the predictions of every algorithm.

Data Acquisition

Pre-Process the Data

Create Bag of Words Model

Create Bag of Words Model

Split the dataset into Training and Test Set

Fit the Training Set to an algorithm

Predict the Test Set Results

Analyze the Results

Figure 1. Architectural Diagram

The architectural diagram for applying the classification algorithms is shown in Figure 1. The steps involved are:

## Data Collection

First step in building any Machine Learning Model is data collection. In this, all the data required to build the Model is collected and stored. Some websites like Twitter and YouTube provide API to extract information from their websites. Some websites like Amazon do not provide any API to extract information, but the data can still be extracted using web scraping techniques. For this project, we have collected data of various mobile phones’ reviews from Amazon using Python’s BeautifulSoup4 library.

## Preprocessing

The data collected in the first step will not be of the right format to process and apply algorithms out of the box. Therefore, pre-processing is a very crucial step in building any Machine Learning Model.

In this, we first take all the data and classify it into 2 datasets. One is a training set on which the machine learning model will train and the other is the test set, which the machine learning model predicts the outcome. Training Set contains 80% of the collected data and the Test Set contains 20% of the data.

Next is a process of removing all unwanted words from the reviews collected. First, the words are made into lowercase, split and each word is put into an array. Iterating over the array of words, all the unwanted stop words are removed as they do not play a role in predicting the sentiment. Stop words are words like this, a, an etc. Next we do something known as stemming words, where all the words are changed to their present tense form (for example, loved becomes love).

Finally, we will have a list of words that play a rile in detecting the sentiment. Next step is to create the Bag of Words Model. Bag of Words model is the process of converting textual data into numeric data. Text is converted into numeric data called as feature vectors. For example, consider the words ‘wow’, ‘bad’, ‘awesome’ and ‘great’. Now, we create a matrix for these words i.e., create one column for each word. Then for each review, we create a row in the matrix. And for each row, if any of the words in the column appears, its respective cell is marked as 1, else 0. This is known as sparse matrix. Finally, our data is ready to be trained and tested.

## Building the Machine Learning Model

In this, we fit different classification algorithms such as K-Nearest Neighbors, Support Vector Machine, Linear Kernel Support Vector Machine, Naïve Bayes, Decision Tree Classification and Random Forest Classification and analyze the results.

## Result Analysis

This is the final step where we compare and analyze the results of each algorithm using Confusion Matrix, Accuracy, FAR, FRR and vertical bar graph.

# Algorithm and Psuedo Code

Input: Labeled Dataset

Output: Confusion Matrix with the number of correct and incorrect predictions.

Psuedo Code of the implementation is given in the table below:

Table 1. Pseudo Code

|  |
| --- |
| 1. Import the dataset 2. Pre-Process the data   Pre-Processing ()  Remove all special characters  Make all the words lowercase  Split all the words and store them in an array  Stem the words using Porter Stemmer  Join the words back together   1. Create Bag of Words Model 2. Split the dataset into Training and Test set 3. Fit the training set to an algorithm 4. Predict the Test Set Results 5. Analyze the results using Confusion Matrix, Accuracy, False Acceptance Ratio and False Recognition Rate. |

# Performance Analysis

The results obtained via Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Linear Kernel Support Vector Machine, Decision Tree and Random Forest on Amazon Review’s data are analyzed using Confusion Matrix, Accuracy, FAR and FRR.

Accuracy is measured using the following formula:

False Acceptance Ratio (FAR) is calculated using

False Recognition Rate (FRR) is calculated using

Here is a line graph showing the accuracy of all the algorithms.

Figure 2. Accuracy Graph

Table depicitng the accuracies is given below:

Table 2. Accuracy Table

|  |  |
| --- | --- |
| **ALGORITHM** | **ACCURACY** |
| Gaussian Naïve Bayes | 37.7% |
| Support Vector Machine | 87.5% |
| Linear Kernal SVM | 87.5% |
| K-Nearest Neighbours | 79.23% |
| Decision Tree | 82.45% |
| Random Forest | 84.67% |

Confusion matrix, FAR and FRR for every algorithm have been calculated. They are given below:

1. Gaussian Naïve Bayes

Table 3. Gaussian Naïve Bayes Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Negative | Positive |
| Actual | Negative | 94 | 270 |
| Positive | 39 | 93 |

FAR = 0.7438

FRR = 0.10743

1. Support Vector Machine

Table 4. Support Vector Machine Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Negative | Positive |
| Actual | Negative | 98 | 27 |
| Positive | 35 | 336 |

FAR = 0.0743

FRR = 0.09641

1. Linear Kernal SVM

Table 5. Linear Kernal SVM Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Negative | Positive |
| Actual | Negative | 98 | 27 |
| Positive | 35 | 336 |

FAR = 0.07438

FRR = 0.09641

1. K-Nearest Neighbour

Table 6. K-Nearest Neighbour Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Negative | Positive |
| Actual | Negative | 38 | 8 |
| Positive | 95 | 334 |

FAR = 0.02203

FRR = 0.2617

1. Decision Tree

Table 7. Decision Tree Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Negative | Positive |
| Actual | Negative | 85 | 39 |
| Positive | 48 | 324 |

FAR = 0.10743

FRR = 0.13223

1. Random Forest

Table 8. Random Forest Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Negative | Positive |
| Actual | Negative | 104 | 47 |
| Positive | 29 | 316 |

FAR = 0.12947

FRR = 0.82369

# Conclusion

In this paper, we compared 6 classification algorithms of machine learning with Amazon Review Data. It is clear that Support Vector Machine and Linear Support Vector Machine outperformed Gaussian Naïve Bayes, K Nearest Neighbor, Decision Tree and Random Forest with an accuracy of 87.5%. Further, we can remove a few words from the stop words list and check if the same algorithms perform better with the same dataset.

# References

1. Chantal Fry and Sukanya Manna, “Can we Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms”, *IEEE Tenth International Conference on Semantic Computing (ICSC),* pp. 376 – 376, March 2016
2. Afshan Ejaz, Zakia Turabee, Maria Rahim and Shakeel Khoja, “Opinion mining approaches on Amazon product reviews: A comparative study”, *2017 International Conference on Information and Communication Technologies (ICICT),* pp. 178 – 178, March 2018
3. Xing Fang and Justin Zhan, “Sentiment analysis using product review data”, *Journal of Big Data 2015*, January 2015
4. Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves and Fabrício Benevenuto, SentiBench – “A benchmark comparison of state-of-the-practice sentiment analysis methods”, *EPJ Data Science 2016,* February 2016
5. Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter and Hamzah Al Najada, “Survey of review spam detection using machine learning techniques”, *Journal of Big Data 2015,* June 2015
6. David Garcia, Antonios Garas and Frank Schweitzer, “Positive words carry less information than negative words”, *EJP Data Science 2012,* January 2012
7. [Kudakwashe Zvarevashe](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Kudakwashe%20Zvarevashe.QT.&newsearch=true) and [Oludayo O. Olugbara](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Oludayo%20O.%20Olugbara.QT.&newsearch=true), “A framework for sentiment analysis with opinion mining of hotel reviews”, [*2018 Conference on Information Communications Technology and Society (ICTAS)*](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8363089), March 2018
8. A Jenifer Jothi Mary and L Arockiam, “Jen-Ton: A framework to enhance the accuracy of aspect level sentiment analysis in big data”, *2017 International Conference on Inventive Computing and Informatics (ICICI),* pp. 456-456, November 2017
9. Monali Bordoloi and S K Biswas, “E-commerce sentiment analysis using graph based approach”, *2017 International Conference on Inventive Computing and Informatics (ICICI),* pp. 575-575, November 2017
10. M Sivakumar and U Srinivasalu Reddy, “Aspect based sentiment analysis of students opinion using machine learning techniques”, [*2017 International Conference on Inventive Computing and Informatics (ICICI)*](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8360186)*,* pp. 726-726, November 2017
11. M Mayilvaganan and D Kalpanadevi, “Comparison of classification techniques for predicting the performance of students academic environment”, *2014 International Conference on Communication and Network Technologies*, pp. 113-118, December 2014
12. Jihyun Ryoo, Meena Arunachalam, Rahul Khanna and Mahmut T Kandemir, “Efficient K nearest neighbor algorithm implementations for throughput-oriented architectures”, *2018 19th International Symposium on Quality Electronic Design (ISQED),* pp. 144-144, March 2018
13. Thuy Nguyen Thi Thu and Voun Dang Xuan, “Using support vector machine in FoRex predicting”, *2018 IEEE International Conference on Innovative Research and Development (ICIRD),* pp. 1-1, May 2018
14. Kushan Ajay Choksi and Sachin K Jain, “Pattern matrix and decision tree based technique for non-intrusive monitoring of home appliances”, *2017 7th International Conference on Power Systems (ICPS),* pp. 826-826, December 2017
15. Heng-Yi Su and Tzu-Yi Liu, “Enhanced-Online-Random-Forest Model for Static Voltage Stability Assessment Using Wide Area Measurements”, *IEEE Transactions on Power Systems,* pp. 2-2, June 2018