IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 4, April 2019

Comparing Machine Learning Techniques for Sentiment Analysis

Saransh Jitendra Sachdeva[1], Raj Abhishek[2], Dr. Annapurna V.K[3]

Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, India[1,2,3]

Abstract: Sentiment Analysis uses Natural Language Processing (NLP) and text analysis to systematically identify and extract subjective opinion of a document. There are several ways to evaluate the polarity of document. This paper gives insights on various machine classifiers used. Each Classifiers are evaluated separately using predefined metrics to find the best classifier for correctly determining the polarity of document.

Keywords: Sentiment Analysis, Polarity, Machine Learning Classifiers, confusion matrix, K-fold cross validation

I. INTRODUCTION

Sentiment analysis is the contextual mining of text which identifies and extract subjective information of text. People are always curious about what other people think, or what their opinion is. Social media has become a platform to express their opinions on a given topic. Dataset is analyzed and presented in such a way that identifies the online mood of public towards that topics as positive or negative. It can also be used to identify the sentiment spike on a product, to identify social media influencers or product supporters. It helps to identify any potential negative thread emerging on a topic at early stage and deal with it faster. In recent years Tech giants started providing sentiment analysis as a service.
In machine learning, classification is a supervised learning approach where the computer program learns from the data input given to it and then uses this learning to classify new observations. This data set can be bi-class or multi-class. Classifier is an algorithm that maps the input data to a specific category.

This work compares existing machine learning classifiers on Restaurant review dataset of 1000 reviews. Different classifiers used are Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest and Logical Regression. For each classifiers dataset goes through preprocessing stage where the data is cleaned. Each classifier is evaluated using K-fold cross validation and confusion matrix. Mean accuracy and mean standard deviation are calculated and compared to find the best existing classifier for sentiment analysis. The remainder of this paper is organized as follows; Section 2 describes related works performed by other researchers in this field. Section 3 describes the methodology to perform the experiment which includes Data pre-processing module, Classification algorithms and comparison module. Section 4 describes the primary results obtained from the experiment. Section 5 concludes the result obtained from the experiment.

II. RELATED WORK

Sentiment classification of reviews has been the focus of research since it was first introduced in 2003. It has been attempted in different domains such as movie reviews, product reviews, and customer feedback reviews. It has been used to classify tweets on the basis of polarity and hence deciding the stock market predictions.

The Naive Bayes Classification and Support vector machines were discussed in [1]. It gives insight about definitions of Opinion Mining, few machine learning techniques on the basis of their usage and importance for the analysis evaluation of Sentiment classifications.

Two class problem i.e. dividing text into positive and negative sentiments were researched in [2]. The accuracy of sentiment analysis prediction were improved by using deep neural network.

Stanford library was used to classify the data further and insights on NLP (natural language processing) was given in [3]. The dataset was obtained from Twitter API. This paper used RNN (Recurrent neural Network) algorithm to determine the polarity of data and analysed its implementation, challenges and advantages over traditional algorithms.
The survey of various machine learning techniques is represented in [4]. Comparison between different classifier was carried out. On comparing, 85% more accuracy was obtained by using supervised machine learning which is higher than that of unsupervised learning technique.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 4, April 2019

III. METHODOLOGY

The methodology used to carry out the sentiment analysis of a product review or service is by classifying the sentiments in two categories: positive and negative. For this, a sample dataset of 1000 restaurant reviews is taken. It is a bi-class dataset. The dataset undergoes Data Pre-processing steps before being fed to classifier for training purpose.

A. Data Preprocessing Module

Tokenization breaks up a sequence of strings into tokens which are pieces such as words, keywords, phrases, symbols. In this process unwanted characters like punctuation marks, whitespaces are discarded. All the remaining characters within continuous strings are part of token and comprised alphanumeric characters only [5]. The tokens become the input for another process parsing and text mining.

Text normalization helps to reduce tokens as much for efficient creation of Bag of words model. It requires being aware of what type of text is to be normalized and how it is to be processed afterwards. It converts all characters to lowercase and get rid of unwanted characters and numbers such as punctuation marks and white spaces. Stop words which does not add any meaning to sentence are removed. Stemming is applied to all the words which removes affixes from word and stem words are only taken into consideration.

The bag-of-words model makes a unigram model of the text by keeping track of the number of occurrences of each word. This can later be used as a features for Text Classifiers. In this only individual words are taken into account and give each word a specific subjectivity score. It is calculated using conditional probability from the bi-class dataset.

Each review taken in dataset is manually classified before going through pre-processing stage. The accuracy of the model can be measured by splitting data into 2 parts. A training set is fit and transformed for a model. Test dataset just fits into the model which is already transformed by the training set. It creates predictions and evaluate the model. The training process finds hidden dependencies and patterns in the data which are analysed. The data set is split in 80:20 ratio to avoid over-fitting. It is important to cover as many words as possible that express sentiment and represent the lexicon used in the target texts.

B. Classification Algorithms

Classification predicts the class of given data points, termed as targets/ labels or categories.

Classification predictive modelling approximates mapping function (f) from input variables (X) to discrete output variables (y).

$$F(X) = y;$$

Where X is Pre-processed sentence (data points) and y=Polarity of sentence (Class/label)

The following algorithms have been used for this paper:

Naïve Bayes

Naive Bayes classification algorithm is used in binary (two-class) and multi-class classification problems. It does not calculate the values of each attribute P(d1, d2, d3|h), and are assumed conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on. This approach performs extremely well on data where this assumption does not hold.

Only statistical view of data is considered to calculate the probability The starting point is that the probability of a class C is given by the posterior probability P(C|D) given a training document D. Here D refers to all of the text in the entire training set. It is given by D = (d_1, d_2, .., d_n), where d_i is the i_{th} attribute (word) of document D.

$$P(C = c_i|D) = \frac{P(D|C=c_i) \cdot P(C=c_i)}{P(D)}$$

Equation 1: Bayes Formula

Random Forest

Random Forest algorithm is an ensemble based supervised classification algorithm. Results from multiple random uncorrelated decision trees are observed to predict the best response.

Support Vector Machine

A Support Vector Machine (SVM) classifier is defined by a separating hyper plane. Using given labelled training data it creates an optimal hyper plane which help in categorizing new dataset. For bi-class classifiers a linear hyper plane is created such that distance between hyper plane and support vectors is maximum.

Decision Tree

A decision tree classifier uses tree structure for classification. Each attribute is represented by an internal node whereas outcome is represented by leaf nodes. This classifier uses recursive partitioning to divide the tree on the basis of the attribute value.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 4, April 2019

Logistic Regression
Logistic Regression classification algorithm predicts the probability of a binary categorical dependent variable.
It describes data and explains the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

C. Comparison Module

Validation quantifies numerical results and its relationship between variables of data.

K-Fold Cross Validation
Overfitting happens when learning the parameters of a prediction function and tests it on the same data. A model gives correct results for its training dataset but its performance reduces drastically on new dataset.
K Fold cross validation divides data into k subsets to avoid overfitting. The process is repeated k times where each subset is once chosen as test set and other rk-1 subsets are chosen as training set. Average of k trials are taken to get mean error estimation which reduces variation and standard deviation on the dataset. Data points that lie more than one standard deviation from the mean can be considered unusual and are usually eliminated.

Confusion Matrix
A confusion matrix describes the performance of a supervised classification model on a set of test dataset. It helps in visualization of the performance of an algorithm. The count of correct and incorrect predictions are summarized and allotted to each class. It gives insight into the errors being made by a classifier and also its type.

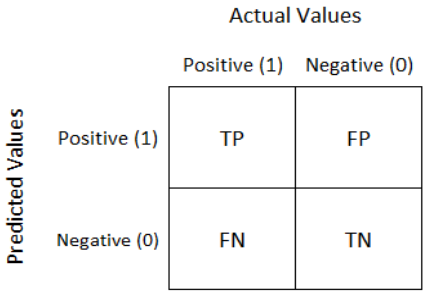

Fig 1: Confusion Matrix

- **True Positives (TP):** These are the cases in which the prediction and results are both positive.
- **True Negatives (TN):** These are the cases in which the prediction and results are both negative.
- **False Positives (FP):** These are the cases in which the prediction was positive but result is negative. This is also termed as Type-I error
- **False Negatives (FN):** These are the cases in which the prediction was negative but result is positive. This is also termed as Type-II error.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 2: Formula of Accuracy

IV. RESULTS

A. K-Fold Cross Validation

This paper evaluates the performance of machine learning classifiers using K-Fold Cross Validation, where **k=10**, and compares the accuracy and standard deviation obtained by them in the following table. Standard deviation tells about the concentration of the data around the mean of the data set. Standard deviation is inversely proportional to the concentration of the data around the mean with high concentration, the standard deviation will be low, and vice versa. It cannot be negative.

Table 1: Evaluating classifier performance using K-Fold Cross Validation

Classifier	Accuracy	Standard Deviation
Naïve Bayes	67.36 %	0.049
Random Forest	72.63 %	0.038
Support Vector Machine	70.65 %	0.058
Decision Tree	70.50 %	0.045
Logistic Regression	75.37 %	0.041

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 4, April 2019

B. Confusion Matrix

The dataset has been split into two parts, training dataset and test dataset in the ratio of 8:2. Each Machine Learning Classifier is evaluated using confusion matrix.

Naïve Bayes

Below is the confusion matrix for the naïve Bayes classifier in our project. The classifier has obtain accuracy of **73%.**

Table 2: Confusion Matrix of Naïve Bayes

	0	1
0	55	42
1	12	91

Random Forest

Below is the confusion matrix of the performance of the Random Forest. The accuracy of this model comes out to be 72%.

Table 3: Confusion Matrix of Random Forest

	0	1
0	87	10
1	46	57

Support Vector Machine

Below is the confusion matrix of the performance of the Support Vector Machine. The accuracy of this model comes out to be 68.5%.

Table 4: Confusion Matrix of SVM

	0	1
0	71	26
1	37	66

Decision Tree

Below is the confusion matrix of the performance of the Decision Tree. The accuracy of this model comes out to be 71%.

Table 5: Confusion Matrix of Decision Tree

	0	1
0	74	23
1	35	68

Logical Regression

Below is the confusion matrix of the performance of the Logical Regression. The accuracy of this model comes out to be 75.5%.

Table 6: Confusion Matrix of Logical Regression

	0	1
0	74	23
1	26	77

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 4, April 2019

V. CONCLUSION

This paper compares different machine learning classifiers for sentiment analysis. Each model is individually evaluated using K-fold Cross Validation. After comparing the mean accuracy and standard deviation of all the classifiers, this paper concludes that Logistic Regression and Random Forest classifiers gives the best result in correctly predicting the polarity of the given dataset.

REFERENCES

[1]. Manoj Kumar Das, Binayak Padhy and Brojo Kishore Mishra, "Opinion Mining and Sentiment Classification: A Review", International Conference on Inventive Systems and Control (ICISC-2017), 16 October 2017.
[2]. Adyan Marendra Ramadhani and Hong Soon Goo. "Twitter Sentiment Analysis using Deep Learning Methods", 2017 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 16 October 2017.
[3]. Dipti Mahajan and Dev Kumar Chaudhary, "SENTIMENT ANALYSIS USING RNN AND GOOGLE TRANSLATOR", 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 23 August 2018.
[4]. Bhavitha Bk, Anisha P Rodrigues , Niranjan N Chiplunkar, comparative study of machine learning techniques in sentimental analysis, 978-1-5090-5297-4 IEEE, 2017.