ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Predicting the Polarity of Reviews Based on Text Based Data

B.Bala Sai Charan[1], Dr. G. Sharmila Sujatha[2]

M. Tech Student, Department of Computer Science and Systems Engineering,Andhra University, Visakhapatnam, Andhra Pradesh, India[1]

Assistant Professor, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India[2]

ABSTRACT: Millions of reviews are generated on the internet about a product, person or a place every day. Because of their number and size, it is very difficult to analyze and understand such reviews. Online reviews are important assets for users to buy a product, see a movie, and make a decision. Therefore, the polarity of a review is one of the key factors for all users to read and trust the reviews about product. So our goal is to develop a classifier that predicts whether the review is positive or negative, In this paper, various classifiers such as Naive Bayes, K-nearest neighbour and Logistic regression are used to predict the polarity of reviews and compare all these machine learning classifiers to obtain the best results based on accuracy as metric.

KEYWORDS: Textclassification, Prediction, Polarity.

I. INTRODUCTION

The opinions which are given in the form of reviews are increasing in such a way that it exceeds the analysing capacity of an individual to make an informed decision regarding the product. To solve this problem of decision-making regarding opinions and sentiments. It uses Sentiment Analysis, Sentiment Analysis is natural language processing and information extraction task, Itis the process of determining whether language reflects positive or negative sentiment.

Steps involved in sentimental analysis:

1.1 Text extraction: This step involves extracting words from text-based reviews that influences outcome of the result.
1.2Text Refinement: This step involves refining reviews that involves stemming, stop word removal, tokenization.
1.3 Text Classification: This step classifies the review into one of the positive class or negative class

II. PROBLEM STATEMENT

The main objective is to determine whether a review is positive or negative given text-based reviews. The input to the classifier is text-based review and output is review is either positive or negative.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

III. DATASET

The dataset that is used for analysis consists of reviews of products of Amazon for more than 10 years which includes 500,000 reviews**,** the main aim is to obtain writer's feelings expressed in reviews, by analysing a large documents

The dataset consists of 10 attributes:

1. Id
2. ProductId - unique id for uniquely identifying the product
3. UserId - unique id for uniquely identifying the user
4. ProfileName -profile name of user
5. HelpfulnessNumerator - number of users who found the review is useful
6. HelpfulnessDenominator - number of users who indicate the review is useful or not
7. Score - a rating between 1 and 5
8. Time - time stamp for the review
9. Summary - a summary of the review
10. Text - Text of the reviews

IV. METHODOLOGY

Predicting the polarity of review requires 3 steps:

4.1 Text preprocessing:

The input to the classifiers is text-based reviews, but this text contains a lot of unnecessary characters like unusual capitalization, extra whitespaces, and punctuation, which is not useful for classification. Therefore, all of these things must be removed. Some tokens may stop words, that is, words that are very common in English and convey no information. So, we should remove these also, as they will reduce the accuracy of the classifier.

4.2 Transformation:

Once the text is cleaned, we need to convert text into feature vectors. We will use a bag of words and tf-idf representation to convert reviews into sparse vector

4.3 Classification:

To build the classifier we use three different supervised techniques: k Nearest Neighbour, Naive Bayes, and Logistic Regression.

V. IMPLEMENTATION

5.1 WORKFLOW:

(a) Sort data based on time.
(b) Convert reviews in dataset into vectors using bag of words and tf-idf
(c) Split data into train and test.
(d) Apply classification algorithm to predict polarity of train data
(e) Find accuracy of the model

5.1.1 K NEAREST NEIGHBOUR:

K Nearest Neighbour (KNN) is useful for classification problems. KNN is lazy learning algorithm which means there is no assumption for fundamental data distribution, In KNN, K is the number of nearest neighbours. The number of nearest neighbours is the deciding factor. For given input, it finds the distance from test data and gives the prediction about input.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

KNN algorithm works as follow:

1. Load the dataset
2. Find optimal k.
3. For getting the polarity of review, iterate from 1 to total number of training data points.
    a. Find the distance between test data and each row of training data. we use Euclidean distance as distance metric
    b. Sort the obtained Euclidean distances in ascending order based on their values.
    c. Select the top k rows from the sorted values.
    d. Select the most frequent class of these rows and assign to given training data point

5.1.2 NAIVE BAYES ALGORITHM:

Naive Bayes is a classification algorithm which uses the Bayes Theorem. It calculates membership probabilities for each review belongs to either positive or negative class, the class with the highest membership probability is considered as the most likely class. This is also known as Maximum A Posterior (MAP**)** for given hypothesis.

The MAP for a hypothesis is:

MAP(H)=max(P(H|E))
=max((P(E|H)*P(H))/P(E))
= max(P(E|H)*P(H))

P(E) is evidence probability, used to normalize the result.

Naive Bayes classifier assumes that all the features are unrelated to each other. Every feature is independent, In review dataset, we test a hypothesis for given multiple features. we use feature independence approach to 'uncouple' multiple evidence and treat each as an independent one.

P(H|Multiple Evidences) = P(E1| H)* P(E2|H) ……*P(En|H) * P(H) / P(Multiple Evidences)

5.1.3 LOGISTIC REGRESSION:

Logistic Regression is a Classification algorithm which is used mostly for binary classification problems, The hypothesis (hø) of logistic regression maintains the cost function between 0 and 1.

$$0 < h\emptyset(x) < 1 \qquad (1)$$

Eq.1 Hypothesis of logistic regression

We use the sigmoid function to maps any real value into another value between 0 and 1. we use sigmoid function to map predictions to probabilities.

When we pass the inputs to prediction function, it returns probability score between 0 and 1 using a cost function

The Cost function of Logistic Regression is defined as

$$cost(h\emptyset(x), y) = -\log(h\emptyset(x)) \quad if\ y = 1 \qquad (2)$$

$$cost(h\emptyset(x), y) = -\log(1 - h\emptyset(x)) \quad if\ y = 0 \qquad (3)$$

These techniques were chosen as they are simple to understand and implement, they run relatively quickly, and they have historically given good results for text classification. In each classifier, we use 70% of the data for training and the rest for testing.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

VI. RESULTS

Predicting the polarity of reviews is performed using K Nearest Neighbours with bag of words and tf-idf representation, Naive Bayes with a bag of words and tf-idfrepresentation and Logistic Representation with bag of words and tf-idf and find the accuracy of each classifier using accuracy as metric

ACCURACY=(Tp+Tn)/(Tp+Tn+Fp+Fn)

Where Tp = True Positive
Tn = True Negative
Fp = False Positive
Fn = False Negative

6.1The result showing accuracy of various classifiers:

Table. 6.1. Result

SNO	CLASSIFIERS	ACCURACY
1	K NEAREST NEIGHBOURS USING BAG OF WORDS VECTOR REPRESENTATION	85
2	K NEAREST NEIGHBOURS USING TF-IDF VECTOR REPRESENTATION	86
3	NAIVE BAYES USING BAG OF WORDS VECTOR REPRESENTATION	84
4	NAIVE BAYES USING TF-IDF VECTOR REPRESENTATION	84
5	LOGISTIC REGRESSION USING BAG OF WORDS VECTOR REPRESENTATION	91
6	LOGISTIC REGRESSION USING TF-IDF VECTOR REPRESENTATION	89

Table 6.1 displays accuracy of various classifiers in predicting the polarity of text-based reviews,

The results showing that logistic regression using bag of words vector representation given best accuracy compares to remaining classifiers in predicting the polarity of reviews

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

VII. CONCLUSION

In this paper, a set of techniques are discussed and compared for predicting the polarity of product reviews based on natural language processing and classification algorithms. The objective is to predict polarity for customer reviews of a product sold online. The experimental results indicate that the techniques used are very promising in performing their tasks. Predicting the reviews is not only useful to customers, but also product manufacturers to improve their product quality and features thus helping then in marketing.

REFERENCES

[1] Jason Jong, "Predicting Rating with Sentiment Analysis on Yelp Dataset," Stanford Univ., Stanford, CA,2011
[2] Chen Li and Jin Zhang, "Prediction of Yelp Review Star Rating using Sentiment Analysis," Stanford Univ
[3]JayashriKhairnar, MayuraKinikar "Machine Learning Algorithms for Opinion Mining and Sentiment Classification"International Journal of Scientific and Research Publications,
[4] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis" Amit Gupte et al,
[5] analyticsvidhya/blog/2018/08/k-nearest-neighbor-introduction-regression-python/
[6] machinelearningplus/predictive-modeling/how-naive-bayes-algorithm-works-with-example/
[7] towardsdatascience/understanding-logistic-regression-step-by-step-704a78be7e0a/
[8] medium/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428