Sentiment Analysis of Movie Reviews using Machine Learning Classifiers

Mamtesh
National Institute of Technology
Kurukshetra, India

Seema Mehla
National Institute of Technology
Kurukshetra, India

ABSTRACT
In today's world, it has become customary to collect opinions and reviews from people through various surveys, polls, social media platform and analyse them in order to understand the preferences of customers. So, in order to understand the sentiments of customers and their view on the services offered by producers, there comes the need for an accurate and canonical mechanism for speculating and anticipating sentiments which possess the ability to fabricate a positive or negative impact in the market and thus making this kind of analysis important for the pair of producers and consumers. In this paper, the main focus is to anatomize the reviews conveyed by viewers on various movies and to use this analysis to understand the customers' sentiments and market behaviour for better customer experience.

This paper intends to analyse the reviews of customers on various movies by implementing three algorithms namely K Nearest Neighbours, Logistic Regression and Naive Bayes and provides conclusive remarks.

Keywords
K Nearest Neighbours, Logistic Regression and Naive Bayes

1. INTRODUCTION
In today's world there is a lot of impact being made by technology in daily life. There has been an tremendous growth in adoption of new technologies research and development. Technology has been developed to such an extent where it has become a part in our lives. The advancements in Web were made to such a large extent, as a matter of which there is an enormous increase in the volume of sentimental content accessible in the Web. Such variety of information is found day in day out in the social web / websites / public network in the semblance of movie reviews or product ratings, customer statements, testimonials, critiques in discussion forums etc. By using these kind of information collected from web in a proper way using an appropriate technology, one can bring a huge change in the market by understanding the market trends and companies or producers can customize their product as preferred by customer. This type of analysis is called Sentiment Analysis [4][5].

Sentiment Analysis refers to the use of natural language processing (NLP) and text analysis to schematically study, identify, extract, and evaluate opinions within text. It builds systems that try to identify whether the statement makes a positive, negative impact or doesn't create an impact at all i.e., neutral impact. Sentiment Analysis also called as Opinion Mining, besides identifying the opinion and its emotion, also extract attributes of the statement such as

- Polarity : The opinion expressed by speaker is positive or negative
- Opinion holder : who is expressing the opinion (the person, or entity)
- Subject : what is the product that is being discussed

Since sentiment analysis has many practical applications, there has been a tremendous growth of interest in research and development of various Analysis and Prediction techniques[7][9]. The input for this mechanism is the large number of texts expressing opinions which are available publicly and privately from various review sites, forums, blogs, and social media platforms such as Facebook, Twitter, Linkedin, Reddit, Quora etc.,

Sentiment analysis can be done at distinct levels as follows-

- **Document level:** Sentiment analysis that extracts the sentiments of a complete paragraph or document.
- **Sentence level:** Sentiment analysis which analyse a single line and draws out the sentiments from that sole line.
- **Sub-sentence level:** Sentiment analysis that draws out sentiments of sub-expressions within a sentence [2].

All the algorithms that are being used into Sentiment Analysis systems may be broadly classified into following three types[1][10]

- **Rule-based systems** perform sentiment analysis on the basis of a set of manually designed rules.
- **Automatic systems** that emphasize on machine learning algorithms to learn from data.
- **Hybrid systems** combine the features from both rule-based systems and automatic systems in order to improve the precision and accuracy.

Rule-based systems
To identify the subjectivity, polarity, or the topic of an opinion, a rule-based system follows manually crafted rules in some kind of scripting language.

Many types of inputs may be used by rules

- Traditional NLP techniques, such as stemming, tokenization, parsing and part of speech tagging
- Other resources, such as lexicons (i.e. lists of words and expressions).

Automatic systems
Contrary to rule-based systems, automatic systems do not use manual rules. Instead these systems implement the machine learning techniques The task of sentiment analysis can be modeled into classification problem, which in turn, can be solved using machine learning classifiers, like K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Decision Tree & Random Forest [8].

2. LITERATURE REVIEW

According to Liu [1], sentiment computing was domain of study which analysed people's opinions, sentiments, appraisals, emotions, attitudes and evaluation towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes.

Nakov et. al [2] merged word-level and character-level models into one model, and they also got a great result. The text sentiment processing could be divided into three levels: word, sentiment and document. For word level sentiment analysis, we try to identify words which can be mapped with semantic lexicons. Li et. al [3] compared the performance between lexicon based classifier and a statistical machine learning-based classifier. They also chose NTUSD as their base sentiment lexicon, which included 8276 and 2810 words of positive and negative sentiments respectively.

Jiang et. al[4] proposed an improved KNN algorithm for text categorization having combination of one pass clustering algorithm and KNN algorithm. K nearest neighbours and Support Vector Machine had much better performance than other classifiers. However, KNN was the sample-based learning technique, which made use of all the training documents in order to predict the labels of the test document and had very huge text similarity computation. As a result, we couldn't use it widely in real-world applications.

Aston et. al [5] proposed algorithms in which perceptrons with best learning rate, and vote perceptron to classify the sentiments of tweets in data stream. In this paper the subjectivity or objectivity of the tweet with much lower error rate was determined.

Hodeghatta Umesh Rao [6] presented the results of work performed on Twitter to classify Twitter messages of Hollywood movies as positive, negative, and cognitive sentiment statements.

3. CONTRIBUTION

Lexical stratagem is the primary pathway to ascertain contrariety in order to quantify large amount of words with some sentiment score and deduce the collective contrariety and apply few algorithms upon them.

3.1 Data Collection

XYZDATA.csv files which consists of movie reviews. The purpose is to predict whether a given review is positive or negative. To accomplish this, an algorithm is trained using the reviews and their classifications in train_data.csv, and later on make predictions on the reviews contained in test_data.csv. Then errors are computed by making use of the actual classifications of test_data.csv, and observe to what extent were our predictions accurate.

A) Dataset Pre-processing

The construction of the dataset preliminary to application of any algorithm on it is involved in the pre-processing of the document. This is done to remove the unwanted words or symbols. These words / symbols do not affect the outcome but can slow down the processing of algorithm. Our dataset pre-processing involves the following steps:

Stopping: A technique to remove most repeating words considering a stop-word list as the basis so as to reduce size of document is called stopping. Commonly known stop words includes a, an, the, this, to, for etc.

Porter Stemming: The process of removing regular or frequent morphological endings from English words is called Porter Stemming. It stems the words to root words. For example, hot, hotter, hottest are stemmed to root word, 'hot'.

Moreover, words with frequency greater than 80% of the dataset are ignored, as they are likely to be a stop-word. Likewise, words having very low frequencies should also be ignored.

B) Train Test Splitting

The whole dataset is divided into two parts namely training data (X_train, y_train) and test data (X_test, y_test). Test data will be used later in the process of computing the efficiency of various classifiers after being learn from training data. A ratio of 80:20 is chosen to break the original dataset.

This step outputs the following lists:

- X_train: training review/features
- y_train: training sentiments/output (1 for positive, 0 for negative)
- X_test: test review/features
- y_test: test sentiments/output (1 for positive, 0 for negative)

3.2 APPLYING ML CLASSIFIERS

The problem of sentiment analysis is also viewed popularly as a 0/1 classification problem. So, popular supervised machine learning classifiers like KNN, Logistic Regression, Naïve Bayes, Support Vector Machine can be used here.

A) K Nearest Neighbours Classifier

KNN is perhaps the simplest and most widely used machine learning algorithm. It can be applied to classification problems as well as regression problems. For smaller datasets, it outperforms most of the other classifiers. Implementation of KNN involves the following two major steps:

i. Finding a group of k objects which are nearest to test object, and

ii. Assigning a label based on predominance of a class in the neighbourhood of the test object.

The step (1), is to calculate the distance from test object to each of training object. Distance may be Euclidian distance, Cosine similarity etc. Euclidian distance is calculated as;

$$E(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

While, cosine similarity is calculated as;

$$similarity = \cos\theta = \frac{A.B}{||A||\,||B||}$$

The step (2), is to sort the vector after calculating the distance from training objects, and pick first k values. Predominance of a class decides the label to be assigned to that test object.

Choice for value of k: There is no direct way to compute the optimal value of k. It solely depends upon problem in hand and the dataset. One simple way to compute optimal k value is to try for different values of k and pick the one having highest accuracy percentage. Also, value of k should not be very large, as it may make the model over fit. Besides from the

obvious advantages of being simple and efficient, KNN suffers a major disadvantage of being computationally expensive. Thus, it is not favoured for large datasets.

B) Logistic Regression Classifier

Logistic Regression is categorized as a classification algorithm. Mostly, it is used to predict a binary outcome (like 0 / 1, False / True, No / Yes, Wrong / Right) when a set of independent variables is given. In order to represent the binary outcome or categorical outcome, we make use of certain dummy variables.

- Hypothesis => Z = WX + B
- $h_\Theta(x)$ = sigmoid (Z)
- sigmoid(t) = $1/(1+e^{-t})$
- **Cost function**

$$Cost\big(h_\theta(x), Y(actual)\big) = \begin{cases} -\log\big(h_\theta(x)\big) & if\ y = 1 \\ -\log\big(1 - h_\theta(x)\big) & if\ y = 0 \end{cases}$$

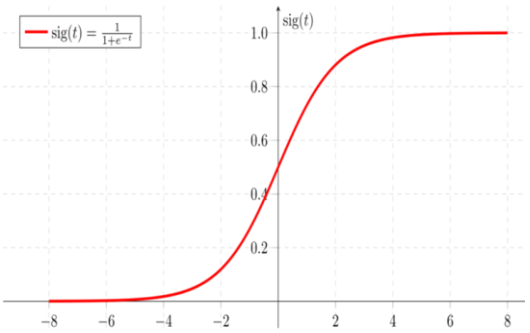

Figure 1: Sigmoid function

C) Naive Bayes Classifier

A classification procedure based on Bayes' theorem by making an assumption that there exists independence among the features. To make it as simple as possible to understand, in general terms we can say that "a classifier which makes the use of Naïve Bayes algorithm presuppose that the occurrence of a precise feature in a class is not related to the occurrence of any other feature". For illustration, a fruit which is black in colour, oval in shape and having a diagonal length about 3cm may be considered to be an grape.

Naive Bayes model is easy & simple to build and is highly useful for classifying text in very large data sets. Along with its simplicity, it is also known to perform even better than highly advanced & sophisticated classification algorithms. Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Where $P(c|x)$ = Posterior Probability

$P(c)$ = Class Prior Priority

$P(x|c)$ = Likelihood

$P(x)$ = Predictor Prior Probability

$P(c|x) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

There are 3 types of models are available for Naïve Bayes classifier in scikit-learn library:

Gaussian: It is used as a classifier by assuming that the features will follow a normal distribution.

Multinomial: It is used in case of disjoint counts. Now consider Bernoulli trials which is a step ahead and count how frequently letter occurs in the statement instead of letter occurring in the statement i.e., you can assume it as taking the number of times outcome number xp is observed over the q trials.

Table 1: Sample Reviews

Sentiment	Movie Reviews
Positive	The Puppet King is an amazing movie, don't take me wrong. The people who are very close to me know "To what extent I love the movie Zorokin.
Negative	It's completely waste of time and money both to watch Zorokin. Had an interesting discussion with one of my colleagues at work about how the movie "James and his Destiny" sucks.

Bernoulli: The binomial model is useful in those cases when your specification points are binary (i.e. consist of zeros and ones). One application would be to classify the text with 'bag of paragraph' model where the 0s & 1s are "word is a substring of the sentence" and "word is not a substring of the sentence" respectively.

Multinomial model is used under Naïve Bayes classifier.

4. RESULTS AND OBSERVATIONS

Naive Bayes, Logistic regression and k-Nearest Neighbours are three types of machine learning classifiers used in this research and the analysis is form at a sentence level. The operations of splitting the data set into two parts for were tested for a data set which is obtained from Unigram and Bigrams. Table 1 shows the accuracy shown by different classifiers and Table 2 shows the accuracy and confusion matrix.

Table 2: Accuracy & Confusion matrix obtained

Classifier	Accuracy	Confusion Matrix
K Nearest Neighbours	98.6994	589 9 9 777
Logistic Regression	99.3497	593 5 4 782
Naïve Bayes	98.9161	586 12 3 783

5. CONCLUSION

Information related to the layman reviews given on a particular product are mostly available in a number of heterogeneous information domains. In this paper, data set is executed from heterogeneous sources and derived relevant tokens by applying train test split and countvectorization techniques. There Machine Learning algorithms (K Nearest Neighbours, Logistic Regression, Naïve Bayes) are implemented in order to train the data set, to analyse the reviews and predicted the sentiment of the reviews either positive or negative, and predicted with high accuracy.

An effective and efficient system has been automated to collect the movie reviews thus enabling the anatomization to be carried out rapidly in such a manner so that the outcome of the evaluation can be used proficiently hitherto its usefulness is superannuated. Naïve Bayes can be used ultra-efficiently and appropriately in implementing sentiment analysis on movie reviews / brand reviews and many more so as to understand the emotions and behavioural preferences of the consumer so as to provide better customer experience.

The future scope of work is to verify whether a hybrid technique can be used by applying the permutations and combinations of the above mentioned classifiers in order to achieve better accuracy.

6. REFERENCES

[1] B. Liu, “Sentiment analysis and operation mining”, Synthesis Lectures on Human Language Technologies, pp. 152-153, 2016.

[2] P. Nakov Tiedemann, "Combining word-level and character-level models for machine translation between closely-related languages", Meeting of the Association for Computational Linguistics: Short Papers, pp. 301-305, 2012.

[3] B. Wen, T. T. He, L. Luo, L. Song, Q. Wang, "Text Sentiment Classification Research Based on Semantic Comprehension", Computer Science, pp. 261-264, 2010.

[4] Shengyi Jiang, Guansong Pang, Meiling Wu, Limin Kuang. School of Informatics, Guangdong University of Foreign Studies, 510420 Guangzhou, China,vol 39,pp. 1503-1509,2012

[5] N. Aston, J. Liddle, W. Hu, “Twitter sentiments in data stream with perceptron of [J]”, Journal of Computer and Communications, pp. 11-16, 2014.

[6] Hodeghatta Umesh Rao, “Sentiment Analysis of Hollywood Movies on Twitter”,IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),pp. 1401-1404, 2013.

[7] Umesh Rao Hodeghatta,”Sentiment Analysis of Hollywood Movies on Twitter”, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 25-29, 2013

[8] Ali hasan, Sana Moin, Ahmad Karim, Shahaboddin Shamshirband “Machine Learning-Based Sentiment Analysis for Twitter Accounts”, Mathematical and Computational Applications, 2018

[9] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani.,” Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python”, International Journal of Computer Applications,vol.165, 2017

[10] Mitali Desai , Mayuri A. Mehta ”Techniques for sentiment analysis of Twitter data: A comprehensive survey”, 2016 International Conference on Computing Communication and Automation (ICCCA), pp. 149-154,2016