IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10545

Restaurant and Cuisine Review Sentiment Analysis using SVM

Vishal Lokam[1], Vaishnavi Shinde[2], Abhishek Raikar[3], Vivek Kate[4], Gaurav Jumnake[5]

Student, Computer Engineering, Zeal College of Engineering and Research, Pune, India[1,2,3,4]

Professor, Computer Engineering, Zeal College of Engineering and Research, Pune, India[5]

Abstract: Nowadays, due to users critically reviewing the restaurant and related services, there is an abundance of the reviews that are available on the internet. There is an opportunity to get more detailed information from the customer's review using sentiment analysis. Certain steps are taken to conduct this sentiment analysis. Implementing a machine learning based approach for sentiment analysis increases the performance and accuracy of the sentiment analysis task, as compared to the lexicon based approach. The accuracy achieved for the chosen dataset was between 75% to 80% and it can be increased furthermore with some changes to the process. The model is trained using domain specific data, so it performs better than a lexicon based approach.

Keywords: Sentiment analysis, Data mining, ML, NLP, SVM

I. INTRODUCTION

Sentiment analysis has rooted its way by making people understand the importance of reviews and opinions. It is the study of opinion where one's emotional tone, attitude and sentiment expressed at a particular time is inspected through online mode. A survey says that 61% of the population is dependent on the online reviews and ratings for a particular product in the market, movies, restaurants and cuisines and makes choices accordingly. Classification of reviews and understanding sentiment polarity of the review helps the reader which product is best and worst for them. Through this, restaurant owners will understand their strengths, what differentiates them from others, what issues need to be addressed before they impact their reputation and much. The major disadvantage of lexicon- based approach is that it doesn't contain all the domain specific terms. It is versatile in the sense that it can be used for general purpose usage but lacks accuracy and performance when dealing with a specific domain. Sentiment score is mostly influenced by domain and certain words can be positive in one domain but considered negative in another. So here we are implementing Machine learning based approach which predicts the sentiment polarity of the review using the Support Vector Machine(SVM) algorithm. Supervised classifiers in machine learning are provided with inputs that are previously labelled information for training. SVM provides high accuracy and speed for large data sets.

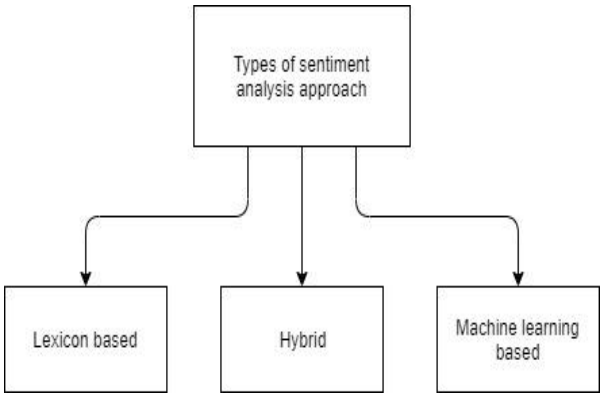

Fig. 1.1 Types of sentiment analysis approach

Generally there are three approaches for sentiment analysis:-

1] Lexicon based approach.
2] Machine learning based approach.
3] Hybrid approach.

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10545

A Lexicon based approach uses dictionaries of polarity words to determine the overall polarity of the sentence. In Machine learning based approach, classifiers are trained and then these classifiers are used for labeling the input. Hybrid approach is the combination of both the lexicon based and machine learning based approach.

II. LITERATURE REVIEW

S. Tiwari, A. Verma, P. Garg and D. Bansal, who worked on "Social Media Sentiment Analysis On Twitter Datasets". They used a twitter API to get the tweets required. For improving the performance of analysis, preprocessing on the collected data was done. Preprocessing includes tokenization, removing unnecessary things (URLs, hashtags, usernames, additional white space etc.), stop words (using NLTK stop word corpus), converting all uppercase to lowercase. After this comes the feature extraction process, which was implemented using a bag of words feature selection. The SVM (Support vector machine) was used as a classifier and trained on training samples. Then this trained model was implemented on a test sample for emotion detection.

M. Wongkar and A. Angdresey, who worked on "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter" followed a similar approach as above. Data was collected using crawler data from twitter.This data was preprocessed like earlier. After that the data was parsed and tokenized to select some important words. They implemented a Naive Bayes classifier for analysis of the data . And showed that it performs better than other classifiers like SVM and KNN for the domain that they were using for.

A. John, A. John and R. Sheik, who worked on "Context Deployed Sentiment Analysis Using Hybrid Lexicon". They used a labelled dataset from sentiment140.com. After that came pre processing and noise removal of the data. They implemented a hybrid lexicon approach in which first the classification is done using SentiWordNet. SentiWordNet is a general purpose lexicon used for assigning sentiment scores. Domain specific classification was also done. The SentiWordNet score was preferred as a final score.

In the research paper titled "Aspect based opinion mining on Restaurant review" by I.K.C.U. Perera and H.A. Caldera,.The data was collected from a website named Zomato which is a restaurant review website. The content can be incomplete or malformed so preprocessing is done on the data. This process will remove noise, white spaces, stop words etc. to increase the overall accuracy of the outcome. In addition to this the words are transformed into simple letters because SentiWordNet considers words with capital and small letters different even though they are the same word but with different cases. It also removes slang which are common while posting a review online. The next step is to find aspects among the reviews i.e finding important features mentioned in the reviews. For this POS tagger is used for tokenization. It tags each word with grammatical types like adjectives, adverbs, nouns, pronouns etc..In most cases nouns and noun phrases are considered aspects. After tokenizing the data using POS tagger, nouns(singular, plurals) and proper nouns(singular/plural) are extracted. After this the frequency of these words is calculated using a frequent word finding algorithm. The words are arranged in descending order according to their frequency and the top five words are selected as important aspects. In the next phase aspect related opinion words should be identified. For this purpose a dependency parser is used. Representation of grammatical relations between words in a sentence is represented by using Stanford dependencies. The output of the dependency parser is provided to the SentiWordNet. The opinion words are extracted and given a positive or negative score. The aspect is considered negative if summation is negative and the aspect is considered positive if the summation is positive.

III. METHODOLOGY

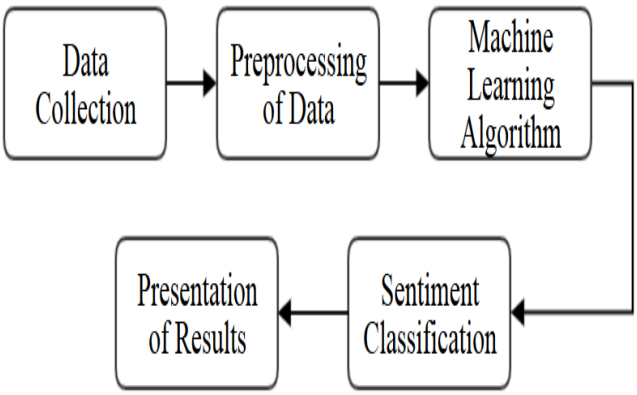

Fig 1.2 Architecture

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10545

This flowchart shows the flow of operations for the task of developing a machine learning model for sentiment classification of restaurant reviews work.

Training dataset:- The dataset required for training the machine learning model is acquired from www.kaggle.com. This dataset contains a large number of reviews which are labelled positive or negative. Generally for the purpose of developing a model, the dataset is divided into a 7:3 ratio. Where 70% data is used for training and 30% is used for testing purposes.

Preprocessing:- Reviews contain words which are not useful for training purposes. These words don't convey the sentiment of the user. So, before training begins, data needs to be cleaned. Stop words, pronouns, noise, white spaces, slang etc. should be removed. These words hold no value in the context of sentiment analysis. Removing these words reduces system resources and improves the performance of the trained model.

Features extraction:- It is the step in which we determine important words which have some meaning. These words determine the overall polarity of the sentence.

Most basic form of feature extraction is counting the occurrence of a particular word in a document also known as a bag of words method. For the purpose of this system, TFIDF(term frequency-inverse document frequency) is used. It is more versatile than the bag of words approach, but not too complicated to understand and implement.

Classification:- For the purpose of classification of reviews in this system, SVM(support vector machine) algorithm is implemented. In the SVM data points are plotted in n-dimensional space. The relative position of these points according to the hyper-plane decides the polarity of the sentence. The model trained using the SVM algorithm is tested by providing a test dataset to it for classification and comparing the result to the expected result.

IV. PERFORMANCE EVALUATION

Using a machine learning algorithm(in this case SVM), improves the reliability of the prediction. Sentiment classification using machine learning approach brings robustness to the system. Prediction is not dependent on very large dictionaries for determining whether the sentiment is positive or negative. A large labelled dataset is required in our process flow for training the classification model. The dataset used for training the model was split into 2 parts. One part contained 70% of the total dataset. It was used for training the model. Remaining 30% part is used for testing the performance of the trained model. Accuracy of the model is 79%, precision score is 0.8 and recall score is 0.73. This model can further be extracted and hosted on the internet to be used for sentiment analysis.

V. CONCLUSION

In this paper, the machine learning model for predicting the sentiment of the review was implemented and results were shown in a user friendly way. Machine learning models are trained for a specific domain, so the same model cannot be used for sentiment analysis of reviews from different domains like movie reviews, resort reviews etc. Using a model developed for one domain for another domain gives sub-optimal results and may affect performance. Accuracy for the model can be further increased by some changes to the methodology.Although machine learning provides better accuracy than lexical based approach, still better results can be achieved by using Deep learning techniques. So, the future work would be to implement sentiment analysis using deep learning techniques.

REFERENCES

[1]. S. Tiwari, A. Verma, P. Garg and D. Bansal, "Social Media Sentiment Analysis On Twitter Datasets," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 925-927.
[2]. M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1-5.
[3]. A. John, A. John and R. Sheik, "Context Deployed Sentiment Analysis Using Hybrid Lexicon," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5.
[4]. I. K. C. U. Perera and H. A. Caldera, "Aspect based opinion mining on restaurant reviews," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 542-546
[5]. https://www.kaggle.com/apekshakom/sentiment-analysis-of-restaurant-reviews
[6]. Kusrini and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 365-368