International Journal of Current Engineering and Technology
©2021 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Aspect based sentiment classification using machine learning for online Reviews

Miss. P. B. Nehe and Prof. A. N. Nawathe

Department of Computer Engineering AVCOE, Sangamner Savitribai Phule Pune University Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

The tourism and travel sector is improving services using a large amount of data collected from different sources. The easy access to comments, evaluations and experiences of different tourists has made the planning of tourism rich and complex. Therefore, a big challenge faced by tourism sector is to use the gathered data for detecting tourist preferences. Unfortunately, some user's comments are irrelevant and complex for understanding these becomes hard for recommendation. Aspect based sentiment classification methods have shown promise in overcome the noise. In existing not much work on aspect based sentiment with classification. This paper presents a framework of aspect based sentiment classification recommendation system that will not only identify the aspects very efficiently but can perform classification task with high accuracy using machine learning na¨ive Bayes and Decision Tree algorithms. The framework helps tourists find the best place, hotel and restaurant in a city, and performance has been evaluated by conducting experiments on Yelp and foursquare real-time datasets.

Keywords: *Machine Learning, Consumer Reviews, Aspect Based Sentiment Analysis,Text mining.*

Introduction

The travel industry is a powerfully developing industry and significant for some regions and nations as key industry. Millions of visitors visit traveler puts each year and offer their feelings on different sites, for example, TripAdvisor and Opinion Table. These sentiments give a general perspective on an opinion holder in regards to the tourist place. In any case, there are countless suppositions which are accessible on a specific spot and it is hard for a typical user to audit/read all these accessible assessments and settle on whether to visit a spot or not. Various sentiment mining strategies have been proposed to manage the huge number of suppositions and these techniques help to characterize the conclusions into positive and negative. In any case, these recently proposed techniques don't manage different perspectives present in a feeling. Rather, these strategies simply call attention to the general sentiments of every opinion. Subsequently, new aspects based opinion mining techniques were proposed. These strategies enable clients to extricate various viewpoints from feelings and group every perspective in the assessments into positive and negative. For example in a given sentence, ”Nourishment is delightful yet administration is slow. “The words ”nourishment” and ”administration” allude to viewpoints and ”flavorful” is the positive assessment of the nourishment perspective and ”moderate” is the negative opinion of service aspect.

In terms of aspects extraction, firstly identifying the implicit aspects is a problem. Implicit aspects do not directly appear in any opinion but it indicates to an important aspect. For instance, in a given sentence “yesterday my sister and I visited sayaji hotel, the taste was superb,” the user did not mention any important aspect in this sentence. But the indication of this sentence is implying a “food” aspect. Secondly, identifying the coreferential aspects is a difficulty. It is common that people use different words and expressions to describe the same aspect. For instance, in a restaurant opinion, atmosphere and ambience refer to the same aspect and these are coreferential to each other. Thirdly, identifying the infrequent aspects is also very cumbersome. Due to large amount of explicit aspects available aspect extraction methods discarded the infrequent aspects. However, some infrequent aspects may be coreferential of frequent aspect or may be important for a tourist place; for instance, Air conditioner and Bed are less frequent aspects but these aspects are important for hotels.

This paper shows a powerful framework of aspect based estimation order by presenting advanced machine learning algorithms. The structure comprises of two fundamental components choice tree-based viewpoint recognizable proof strategy, which allows readers to identify explicit, implicit and infrequent aspects, and groups coreferential aspects from tourist sentiments aspect-based sentiment classification using machine learning algorithms that has three stages In the main stage, Stanford Basic Dependency strategy is utilized to channel sentence parts between slant words and aspects in a given opinion sentence. In the second stage, filtered phrases are used to build features like n-grams and Part-Of-Speech tags. Lastly, machine learning algorithms are applied to identify features to classify the opinions about aspects into positive or negative.

Review of Literature

This paper [1] presents another broadly useful estimation dictionary Sentiment Lexicon and contrasts it and five existing vocabularies: Hu and Liu Opinion Lexicon, Multi-viewpoint Question Answering (MPQA) Subjectivity Lexicon, General Inquirer, National Research Council Canada (NRC) Word-Sentiment Association Lexicon and Semantic Orientation Calculator vocabulary. The adequacy of the opinion dictionaries for estimation order at the report level and sentence level was assessed utilizing an Amazon item survey informational collection and a news features informational index.

This paper [2] proposed Aspect based feeling grouping techniques are accessible in the writing however extremely restricted research work has focused on the programmed viewpoint recognizable proof and extraction of verifiable, rare, and coreferential angles. Viewpoint based characterization experiences the nearness of insignificant sentences in a regular client survey. Such sentences make the information uproarious and corrupt the characterization exactness of the AI calculations. This paper exhibits a fluffy angle based conclusion characterization framework which effectively removes viewpoints from client sentiments and perform close to precise grouping.

In this paper [3] Assessment mining or feeling examination is the computational investigation of individuals' sentiments, evaluations, frames of mind, and feelings toward substances, for example, items, administrations, associations, people, occasions, and their various angles. It has been a functioning exploration zone in regular language preparing and Web mining as of late. Scientists have contemplated assessment mining at the report, sentence and angle levels. Viewpoint level (called angle based assessment mining) is regularly wanted in functional applications as it gives the definite suppositions or estimations about various parts of elements and substances themselves, which are normally required for activity. Viewpoint extraction and substance extraction are in this way two center undertakings of angle based supposition mining.

This paper [4] presents a music recommendation framework dependent on an assessment force metric, named improved Sentiment Metric (eSM) that is the relationship of a vocabulary based estimation metric with a remedy factor dependent on the client's profile. This remedy factor is found by methods for abstract tests, led in a research center condition. In light of the test results, the remedy factor is defined and used to alter the last supposition force. The clients' assumptions are separated from sentences posted on interpersonal organizations and the music proposal framework is performed through a system of low multifaceted nature for cell phones, which recommends melodies dependent on the present client's slant force. Likewise, the structure was manufactured thinking about ergonomic criteria of ease of use.

This paper [5] presents an empirical comparison between SVM and ANN regarding document-level sentiment analysis. We discuss requirements, resulting models and contexts in which both approaches achieve better levels of classification accuracy. We adopt a standard evaluation context with popular supervised methods for feature selection and weighting in a traditional bag-of-words model. Except for some unbalanced data contexts, our experiments indicated that ANN produce superior or at least comparable results to SVM's. Especially on the benchmark dataset of Movies reviews, ANN outperformed SVM by a statistically significant difference, even on the context of unbalanced data.

In this study [6], they conduct a comparative assessment of the performance of three popular ensemble methods (Bagging, Boosting, and Random Subspace) based on five base learners (Naive Bayes, Maximum Entropy, Decision Tree, K Nearest Neighbor, and Support Vector Machine) for sentiment classification. Moreover, ten public sentiment analysis datasets were investigated to verify the effectiveness of ensemble learning for sentiment analysis. Based on a total of 1200 comparative group experiments, empirical results reveal that ensemble methods substantially improve the performance of individual base learners for sentiment classification.

This work [7] proposes an expansion of Bing Liu's viewpoint based feeling mining approach so as to apply it to the travel industry space. The expansion worries with the way that clients allude distinctively to various types of items when composing surveys on the Web. Since Liu's methodology is centered around physical item audits, it couldn't be straightforwardly applied to the travel industry area, which presents includes that are not considered by the model. Through an itemized investigation of on-line the travel industry item surveys, also found these highlights and afterward model them in our expansion, proposing the utilization of new and progressively complex NLP-based standards for the undertakings of abstract and supposition arrangement at the viewpoint level. additionally involve the undertaking of feeling perception and rundown and propose new strategies to assist clients with processing the tremendous accessibility of sentiments in a simple way.

In this paper [8], author propose a novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus (i.e., the given review corpus) and one domain-independent corpus (i.e., the contrasting corpus). We capture this disparity via a measure called domain relevance (DR), which characterizes the relevance of a term to a text collection. We first extract a list of candidate opinion features from the domain review corpus by defining a set of syntactic dependence rules. For each extracted candidate feature, we then estimate its intrinsic-domain relevance (IDR) and extrinsic-domain relevance (EDR) scores on the domaindependent and domain-independent corpora, respectively. Candidate features that are less generic (EDR score less than a threshold) and more domain-specific (IDR score greater than another threshold) are then confirmed as opinion features.

In this paper [9], author propose a sentiment classification method for the categorization of tourist reviews according to the sentiment expressed. We also give the results of the application of our sentiment analysis method on a real data set extracted from the Am Fost Acolo tourist review Web site. In our analysis we were focused on investigating the relation between the opinion holder and the accuracy of the review sentiment with the review score.

In this paper [10], author solve the problem in a different setting where the user provides some seed words for a few aspect categories and the model extracts and clusters aspect terms into categories simultaneously. This setting is important because categorizing aspects is a subjective task for different application purposes, different categorizations may be needed. Some form of user guidance is desired. In this paper, author propose two statistical models to solve this seeded problem, which aim to discover exactly what the user wants.

Proposed Methodology

Architecture presents an overview of the proposed framework for aspect identification and classification. In Step 1 (Data Collection), tourist reviews about tourist places like hotels and restaurants are collected from multiple social media platforms and websites. Step 2 (Data Preprocessing), suppresses noise and redundancy, and cleaned reviews are transformed into sentences. Step 3 (Aspect Identification) finds aspects from preprocessed datasets using a hybrid aspect identification method. Step 4 (Classification) uses machine learning to classify the identified aspects into positive or negative sentiment.

A. Project Modules

1) Review Data Collection - In data collection, reviews are collected from popular social media websites using crawler and APIs. The datasets have different numbers of reviews in each domain. In the restaurant domain, there are 2000 reviews with 1000 positive and 1000 negative. In the hotel domain, there are 4000 reviews with 2000 positive and 2000 negative. London is chosen as a city of interest in the case study.

2) Data Preprocessing - Data preprocessing removes redundancy and ambiguity inherit in the data and transforms the reviews into sentences to facilitate sentence-level aspect-based classification. First, sentences are extracted by identifying the delimiters (e.g. dot, exclamation or question mark). Next, redundant information, e.g. duplicate sentences, is removed. Finally, ambiguous, vague or misspelled terms are corrected.

1. Stop word Removal-This technique removes stop words like is, are,they,but etc.

Initialize i,j

for i=1 to no of words in documents for j=1 no of words in stopword list

if

Words(i)==Stopwords(j) then eliminate words(i) end if end for

2. Tokenization-This technique removes Special character and images.

Initialize feature vector bg feature =[0,0..0] for token in text.tokenize() do if token in dict then token idx=getindex(dict,token) bg feature[token idx]++ else continue end if end for

3. Stemming– Removes suffix and prefix and Find Original words for e.g.- 1. played – play 2.Clustering -

cluster

The word w

Input = Normalize(input) if normalizeValidate(input) then return input; for each rule in rules do if input match with rule then

Stem = ExtractStem(input,rules) if not TestStemLength(Rule) then end for return input

3) Aspect Identification - The objective of aspect identification method is to identify aspects that are important and relevant to a tourist place. This paper proposes a hybrid aspect identification method that can identify both explicit and implicit aspects from reviews about tourist places based on categorize.

4) Data classification - A machine learning algorithm classifies each aspect in a consumer review into positive or negative by considering all aspects and their linkages to sentiment words. For example, in a restaurant review, the tourist likes the food but dislikes the service. The class of this review depends on the sentiment words and phrases linked to aspects. When multiple aspects are considered, the situation becomes more complex; machine learning algorithms are very efficient and helpful.

B. Architecture

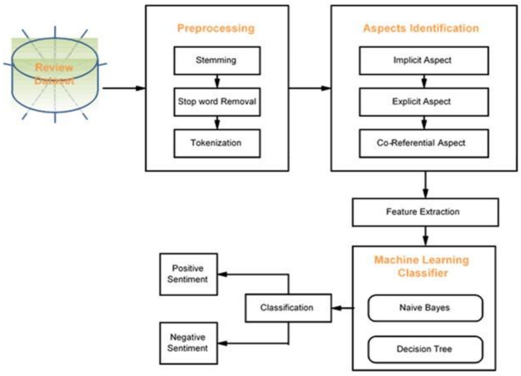

Fig. 1. Proposed System Architecture

C. Algorithm

1. Hybrid Tree Based Aspect Identification Input: Collection of sentences ={S1,S2,S3..Sn}
Output: Aspects assigned to sentences

```
initialize aspects
for all sentences do
    stanford tagger = SPOS (sentences)
    if NN in stanford tagger then
        aspects NN
    end if
end for
initialize aspects groups
for all aspects do
    WordNet sets = WNSS (aspects)
    if TRUE in WordNet sets then
        aspects groups aspects
    end if
end for
frequent aspects = freq _measure (aspects, aspects _groups, 10)
tree = DT (sentences, frequent aspects)
initialize aspect assigned sentences
for all sentences do
    aspect identification = tree (sentences)
    if TRUE in aspect identification then
        aspect assigned sentences aspect identification
    end if
end for
return aspect assigned sentences
```

2. Naive Bayes

Steps:

1. Given training dataset D which consists of documents belonging to different class say Class A and Class B 2. Calculate the prior probability of class A=number of objects of class A/total number of objects Calculate the prior probability of class B=number of objects of class B/total number of objects

3. Find NI, the total no of frequency of each class Na=the total no of frequency of class A Nb=the total no of frequency of class B

4. Find conditional probability of keyword occurrence givena class:
P (value 1/Class A) =count/ni (A)
P (value 1/Class B) =count/ni (B)
P (value 2/Class A) =count/ni (A)
P (value 2/Class B) =count/ni (B)
………………………………….
………………………………….
………………………………….
P (value n/Class B) =count/ni (B)

5. Avoid zero frequency problems by applying uniform distribution

6. Classify Document C based on the probability p(C/W)

a. Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/Class A)……. P(value n /Class A)

b. Find P (B/W) =P (B)*P (value 1/Class B)* P (value 2/Class B)……. P(value n /Class B)

7. Assign document to class that has higher probability.

3. Decision Tree
Input:
Step 1: Upload training dataset
Step 2: Reviews set is the set of input attributes
Step 3: Aspect Classification is the set of output attributes
Step 4: sample is a set of training data
Function Iterative Dichotomiser returns a decision tree

1. Create root node for the tree

2. If (all inputs are positive, return leaf node positive)

If Else (if all inputs are negative, return leaf node negative) Else (Some inputs are positive and some inputs are negative, check condition (Positive¿negative——Positive¡negative), then return result)

3. Calculate the entropy of current state H(S)

4. For each attribute, calculate the entropy with respect to theattribute 'X' denoted by H(S,X)

5. Select the attribute which has maximum value of IG(S,X) 6. Remove the attribute that offers highest value from the set of attributes

7. Repeat until we run out of all attributes or the decision tree has all leaf nodes.

Output:
Dataset value will be retrieved.

Results and Discussion

Experimental evaluation is done to compare the naive bayes and decision tree for evaluating the performance. The simulation platform used is built using Java framework on Windows platform. The system does not require any specific hardware to run; any standard machine is capable of running the application.

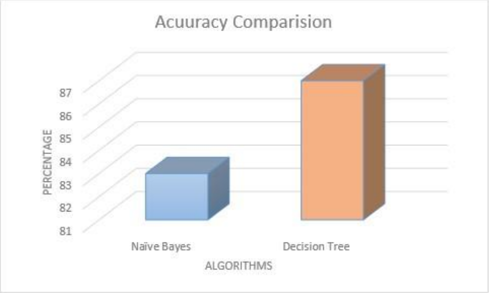

Fig. 2. Graph 1

Table 1:Comparative Result

Sr. No.	Naive Bayes	Decision Tree
1	83%	87%

Conclusion

This proposed system presented an aspect-based sentiment classification framework that classifies reviews about aspects into positive or negative. In this framework, a tree-based aspects extraction method is proposed that extracts both explicit and implicit aspects from tourist opinions. It extracts frequent nouns and noun phrases from reviews text, and then groups similar nouns using WordNet. Decision tree is employed on reviews where review words are used as internal nodes and extracted noun as leaf of a tree. Opinion-less and irrelevant sentences are first removed by employing Stanford Basic Dependency on each sentence. Next, features are extracted from the remaining sentences with N-Grams and POS Tags to train the classifiers. Lastly, machine learning algorithms are applied to the extracted features to train the classifiers.

References

[1] Muhammad Afzaal, Muhammad Usman, Alvis Fong,” Tourism Mobile App with Aspect-Based Sentiment Classification Framework for Tourist Reviews” IEEE Transactions on Consumer Electronics Vol: 65, May 2019.

[2] C. S. Khoo and S. B. Johnkhan, “Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons,” Jour. Inform. Scien., vol. 44, no. 4, pp. 491-511, Aug. 2018, DOI: 10.1177/0165551517703514

[3] M. Afzaal, M. Usman, A. C. M. Fong, S. Fong, and Y. Zhuang, “Fuzzy Aspect Based Opinion Classification System for Mining Tourist Reviews,” Adva. In Fuzzy Sys., vol. 2016, Oct. 2016, DOI:10.1155/2016/6965725

[4] L. Zhang and B. Liu, “Aspect and entity extraction for opinion mining,” in Data Mining Knowledge Discovery For Big Data, Berlin, Germany: Springer, 2014, pp. 1-40, DOI: 10.1007/978-3- 642-40837-3

[5] R. L. Rosa, D. Z. Rodriguez, and G. Bressan, “Music recommendation system based on user’s sentiments extracted from social networks,” IEEE Trans. Consum. Electron., vol. 61, no. 3, pp. 359-367, Aug. 2015, DOI: 10.1109/TCE.2015.7298296

[6] R. Moraes, J. F. Valiati, and W. P. G. Neto, “Document-level sentiment classification: An empirical comparison between SVM and ANN,” Expert Sys. With Appli., vol. 40, no. 2, pp. 621-633, Feb. 2013, DOI: 10.1016/j.eswa.2012.07.059

[7] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, “Sentiment classification: The contribution of ensemble learning,” Decision Supp. Sys., vol. 57, pp. 77-93, Jan. 2014, DOI: 10.1016/j.dss.2013.08.002

[8] E. Marrese-Taylor, J. D. Velasquez, and F. Bravo-Marquez, “A novel de-´ terministic approach for aspect-based opinion mining in tourism products reviews,” Expert Sys. With Appli., vol. 41, no. 17, pp. 7764-7775, Dec. 2014, DOI: 10.1016/j.eswa.2014.05.045

[9] Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, “Identifying features in opinion mining via intrinsic and extrinsic domain relevance,” IEEE Trans. Know. And Data Engi., vol. 26, no. 3, pp. 623-634, Mar. 2014, DOI: 10.1109/TKDE.2013.26

[10] M. Colhon, C. Badic˘ a, and A. S¸endre, ”Relating the opinion holder˘ and the review accuracy in sentiment analysis of tourist reviews,” in Int. Conf. Knowledge Sci., Eng. and Manage., 2014, pp. 246-257, DOI: 10.1007/978-3-319-12096-622

[11] A. Mukherjee and B. Liu, ”Aspect extraction through semisupervised modeling,” in Proc. 50th Annu. Meeting Assoc. for Computational Linguistics, 2012, pp. 339-348.