MACHINE LEARNING

(DS7A-709 and DS7A707)

A MINOR PROJECT REPORT ON

Classification of Customer Reviews

Session 2022-23

In partial fulfilment of the award of the degree

Of

MASTER OF TECHNOLOGY

In

DATA SCIENCE

Submitted by

HIMANI SARATHE

DS7A-2208

Submitted to

RESPECTED DR.DINESH BHATI

SCHOOL OF DATA SCIENCE AND FORECASTING DEVI AHILYA VISHWAVIDYALAYA
INDORE (M.P)

STATEMENT OF ORIGINALITY

Following the requirements for the Degree of Master of Technology in DATA SCIENCE in the SCHOOL OF DATA SCIENCE AND FORECASTING, I present this report entitled – MINOR PROJECT ON **Classification of Customer Reviews**. This report is completed under the Supervision of:

DR DINESH BHATI

I declare that the work presented in the report is my work except as acknowledged in the text and footnotes, and that to my knowledge this material has not been submitted either in whole or in part, for any other degree at this University or any other such Institution.

Name of the Student:
HIMANI SARATHE
Date:30/11/2022

SCHOOL OF DATA SCIENCE AND FORECASTING DEVI AHILYA VISHWAVIDYALAYA
INDORE (M.P)

CERTIFICATE

This is to certify that the dissertation entitled "----MINOR PROJECT ON **Classification of Customer Reviews** -----" submitted by –HIMANI SARATHE--- is approved for the award of Master of Technology in DATA SCIENCE.

ACKNOWLEDGEMENT

We would like to express our gratitude to the DINESH BHATI Faculty of the SCHOOL OF DATA SCIENCE AND FORECASTING Department for guidance and support throughout this work. He has been a constant source of inspiration to us throughout this work. We consider ourselves extremely fortunate for having had the opportunity to learn and work under her guidance over the entire period. I also express my sincere thanks to all the teachers of the SCHOOL OF DATA SCIENCE AND FORECASTING, who gave me the golden opportunity to do this wonderful MINOR PROJECT on the topic "**Classification of Customer Reviews**", which also helped me in doing a lot of research and I came to know about so many new things I am thankful to them. Last but not least I would like to thank all my friends and family members who were involved directly or indirectly in my work.

HIMANI SARATHE

Classification of Customer Reviews Using Machine Learning Algorithms and web scraping

Introduction

Customer reviews have been commonly recognized as valuable sources for marketing intelligence and sentiment analysis (Dickinger and Mazanec 2015). Sentiment analysis seeks to build a system for analyzing and evaluating customer reviews reflected on websites, blogs, Twitter, or Instagram. In recent years, with the expansion of online systems, customer reviews have a powerful impact on business development and attracting potential customers. Therefore, review categorization becomes the key technology to organize textual data. Review categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns (Uğuz 2011). In fact, customer sentiment is very important for the hospitality industry and plays an important role in providing better quality services (e.g., more adaptation to customer requirements and customization of services),

Related Works

Web Scraping

Web scraping is **the process of collecting and parsing raw data from the Web**, and the Python community has come up with some pretty powerful web scraping tools. The Internet hosts perhaps the greatest source of information on the planet.

Sentiment Analytics

Online user-generated content in various social media and websites, such as consumer experiences, user feedback, and product reviews, has increa consumers' and businesses' primary information source and businesses (Duan et al. 2016). Basically, customer reviews demonstrate customer experi- ence

in relation to the organization, which is very important in understanding customer thoughts. These reviews have a major impact on other customers' decisions and are the basis for business improvement. The number of reviews has increased over the past few years, and attention to hidden features in thesell increases the performance of the hotels. In other words, while customers use these reviews in their decision companies, compa- nies use this information to grow products.

Classification

Classification is one of the most commonly used methods in machine learning. It is a process of finding a set of models that allows data classes to be identified and classification massification is to determine the class of future data objects by using past information. In classification, a training set is usually used to learn the model, and the learned information is then tested on the test set. Many classification algorithms have been developed in the literature so far since there is no perfect algorithm for all data sets (Gulsoy and Kulluk 2019).

Proposed Sentiment Classification Fra challenging to study the predictive accuracy of sentiment analytics for the nalytics analyticsl industry. a Such task is more methodological (e.g., clothe using design

factors) than technical (e.g., improving a new classification algorithm) (Fu et al. 2018). Hence, we need an integrative effort to examine how different methods in sentiment analytics influence predictive accuracy, and how to the nsure predictive accuracy of semantic analytics via a systematic approach. Specifically, this article attempts to address the following research questions:

(1) What are the key steps of sentiment analytics for hotel industry?
(2) What are the key design factors of feature engineering?
(3) How do these design factors influence the predictive accuracsentimentment analytics for hotel industry?
(4) How can machine learning methods be systematically incorporated to improve the predictive accuracy of sentiment analytics?

The parts of the proposed system structure are shown in Figure 1. These parts are explained in the following subsections:

Data Collection

The review data were collected from the TripAdvisor.com. TripAdvisor.com is one of the most famous and largest trawebsitessite. A corpus or data collection can be defined as a set of text documents that can be classified under many subsets (Hu and Chen 2016). The corpus contains 400 documents of different lengths. In this data collection, each document was saved in a separate database.

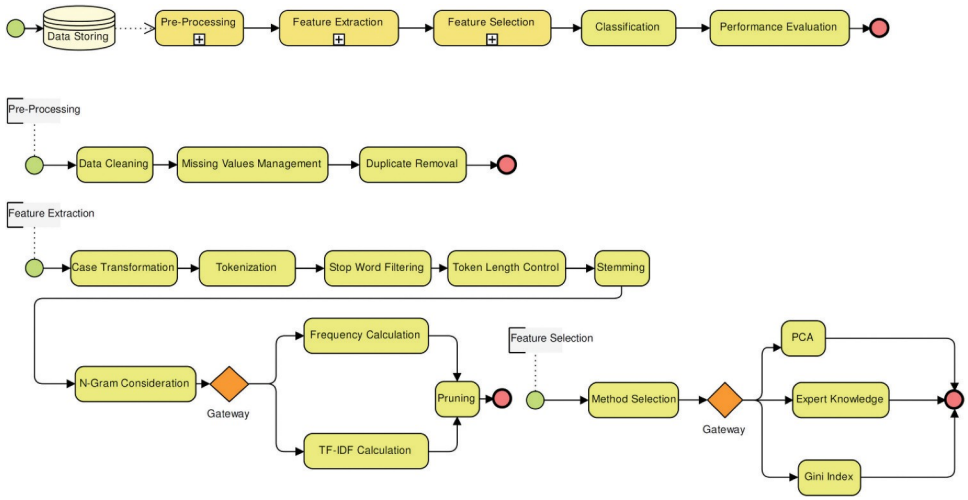


Figure 1. System structure.

Natural language processing

Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data

Re Library

A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression (or if a given regular expression matches a particular string, which comes down to the same thing).

Pre-processing and Feature Extraction

Text pre-processing is an important step in the text classification process. This step ceduce the errors and enhance the accuracy of classification (Bahassine et al. 2018; Uysal and Gunal 2014). The main objective of this endeavor is to get rid of noisy and nonmeaningful words (Bahassine et al. 2018).

Each review was subject to the following procedure:

Data Cleaning

- duplicate removal, delete digits, punctuation marks, and numbers.

- delete stop-words and non-useful words lke: pronouns, artic,les, andpropositions. (Bahassine et al. 2018)

Transformation and Tokenization

Tokenization is the process of splitting reviews into pieces called tokens.

Removing of Stop-words

Words such as conjunctions and pronouns that are not related to the concept of the text are called stop-words. This process involves removing certain common words such as ‘a’, ‘an’, the’, etc., that occur commonly in documents’s. It is important

removing these high-frequency words because they may misclassify the documents (Uğuz 2011).

Stemming

Stemming is a process of reducing inflected words into one form (stem or root) by removing prefixes, suff,xes, and infixes (Bahassine et al. 2018). The stemming process leaves out the root forms of the words. Thereby, terms sharing the same root that seem like different words due to their affixes can be determined. For example, “computer”, “computing”, “computation”, and “computes” all have the same comput root (Uğuz 2011).

Term Weighting

After the words are transformed into terms, the presentation form of the document, which means the expression thereof, terms have to be determined. This process is called term weighting. Thereby, each document could be written in a vector form depending on the terms they contained. To obtain the weight vector,frequencyy-inverse document frequency (TF-IDF) feature weighting algorithm is used as its weight scheme. N-gram refers to a sequence of n tokens based on words.

Pruning of the Words

The pruning procely filters less frequent features in a document collection. The term vector is very high-dimensional and sparse. Also, it is

seen that a number of elements in the term vectare is “0”. Therefore, we prune the words that appear less than two times in the documents. This process decreases the term vector dimension further (Uğuz 2011).

Feature Selection

Feature selection is a process that selects a subset from the original feature set according to some criteria of feature importance (Uğuz 2011). A major pro- blem of sentiment categorization is the high dimensionality of the feature space due to a large number of terms. This problem may cause the tcomputationalnal complexity of machine learning methods used for sentimecategorygor- ization to be increased and may bring about inefficiency and results of low accuracy due to irrelevant terms in the feature space. For a solution to this problem, two techniques are used in this study: feature ranking and feature selection (Uğuz 2011).

Sentiment mining has beco a heated research in recent years. One of the important means of sentiment mining is sentiment categorization. For many problems of sentiment categorization, a good feature selection method can not only reduce the computational complexity but also increase the categorization performance. Feature selection is a process that selects a set of new features from the original features and forms a distinct feature space. Apart from this, feature selection is also perceived as a prerequisite for text categorization, so its significance and importance can be imagined (Wang et al. 2015).

On the other hand, feature extraction produces a large feature set and creates a high-dimensional vector space, which will ultimately lower the efficiency and effectiveness of sentiment classification. As a result, it is critical to select features with significant sentiment distinguishing ability and reduce the dimension of vector space (Wang et al. 2013). Features selection is effective in the reduction of large data in text classification. It can enhance the classification process.

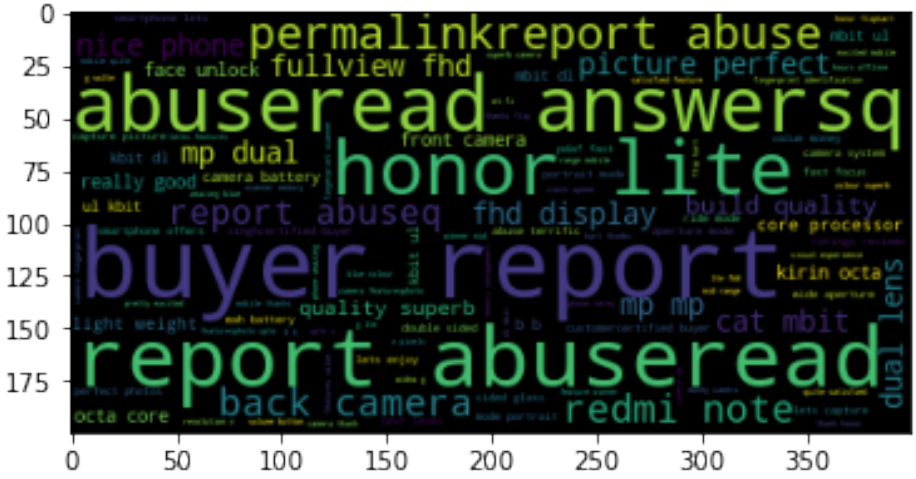
Feature selection deletes irrelevant and noisy data and chooses a representative subset of all data to minimize the complexity of the classification process (Dadgar, Araghi, and Farahani 2016).

Numerous tfeaturesfeatureres selection can be detected in the literature suchas: as Chi-square (Bahassine et al. 2018) the nd Gini index (Manek et al. 2017). The present research tried to introduce a modified version the of Gini feature selection method which will be presented hereafter (Bahassine et al. 2018). A Gini Index-based feature selection method solves the problem mentioned above. The experiments showed that the weight by Gini Index method has better classification performance (Manek et al. 2017). At the end of the feature selection step, terms of high importance in documents are acquired through the Gini method.

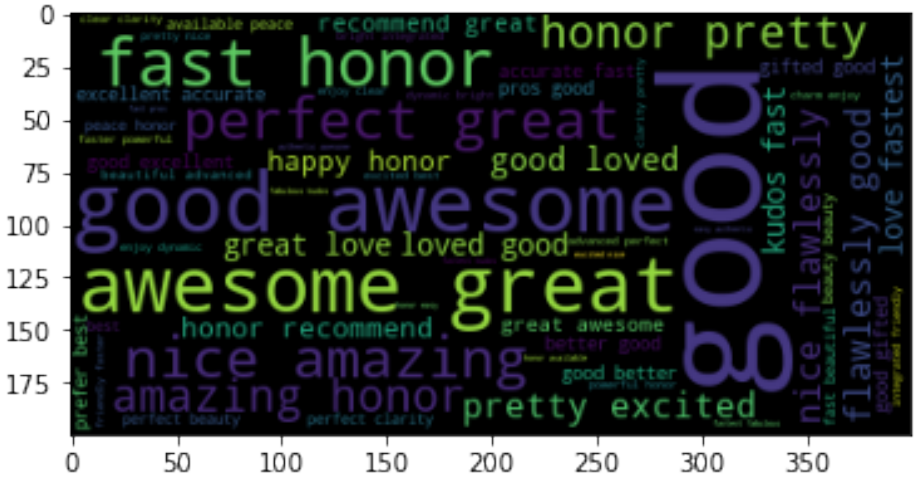
In the current study, feature selection and feature sentiment extraction are used to manage the high dimensionality of a feature space composed of a large

Results:

Wordcloud for Whole Reviews



Wordcloud For Positive Word



Wordcloud For Negative Words

Conclusion

Exploring and analyzing the invaluable hidden data of customer reviews has become a major prerequisite for effective and successful marketing analytics. The large volume of data and Big Data has become a fundamental feature of today's society, yet its ability to analyze, solidify and learn from them, has become a useful element for competition as well as supporting the growth of productivity and innovation. Although big data can certainly be considered as good for decision, large data does not lead to better marketing because they are related to some of the key challenges and issues (i.e. the lack of feature engineering, the need for useful analysis, and so on). Machine learning techniques play an important role in the field of social media analysis (Ducange, Pecori, and Mezzina 2018).

Sentiment classification, which makes up a large part of the voice of the customer in the hotel sector, has become one of the most important problems in the hospitality industry. This is because estimating whether the customers will be unsatisfied is a rather complex problem for hotels. Studies in the literature are predominantly related to the process of sentiment classification. However, there are very few studies on the process of feature identification, feature engineering, and performance comparison of the machine learning algorithm. For this reason, to fill the gap in the literature, feature identification and engineering for this sector have been studied. Then real-life data was collected from an international five-star Iranian hotel chain. Moreover, the data were classified by using six different machine learning algorithms. According to the results, sentiment predictions can be done with 98.9% accuracy.