EXCEL JOURNAL OF ENGINEERING TECHNOLOGY AND MANAGEMENT SCIENCE
(An Peer Reviewed International Multidisciplinary Journal) *ISSN 2249-9032 (Print)*
Vol. I No.25 - February 2024 *ISSN 2277-3339 (Online)*
Impact Factor 5.136 (IIFS)

Sentiment Prediction and Summarization of Customer Opinion using Python

**** Dr. Kavita T. Rangari***

Abstract

Customers can submit reviews for numerous products on websites like Amazon and Flipkart. As e-commerce grows in popularity, so does the quantity of consumer reviews that a product receives. A single product may have hundreds of thousands of reviews, each of which may be lengthy and repetitious. As a result, computerized review summarization offers a lot of potential for assisting buyers in making quick selections about certain items. Text summarization is way of reducing the text content of a document without losing any information. Automatic text summarization is one of the area of natural language processing.

The process of creating a summary from review sentences is known as review summarizing.In this paper, given a product review, a shorter version of the review is created. Web scraping is used to collect reviews from popular e-commerce websites. Sentiment analysis is about determining the text given by the user whether it is Positive, Negative or Neutral. This project is about text summarization which includes sentiment analysis.

In this paper, the researcher present a summarization model for monitoring the generated opinions from online reviews or comment of the customer. The effectiveness of different developed settings of our model was evaluated through several experiments carried out

* Sinhgad Collage of Commerce, Pune, Maharashtra, India

on the data sets and opinions collected from various social networking sites. The obtained results show that our model can generate opinion summaries. Also, sum up the bigger review which can reducethe processing time and improve their performance.

Keyword: Sentiment analysis, e-commerce, summarization, natural language, opinion

Introduction

It is important to understand what a summary is before delving into the specifics of summarizing a document or a large comment. A summary is a text made up of one or more texts that, while condensed, capture key points from the original text. Presenting a condensed, meaningful version of the original text is the aim of automatic text summarization. The primary advantage of utilizing a summary is that it helps consumers save time. Extractive summarization and abstract summarization are the two types of text summarization techniques. The extractive summary approach condenses the sentence from the original document and applies it to the summary. A summary of the findings based on the trained dataset is what the abstract summary aims to accomplish.[1]

Text summarization is a way of reducing the length of the information without losing its meaning. Automated text summarization is called auto-combining. Online reviews are very important for any online business that wants to maintain their online reputation by aggregating several documents. Customer reviews are beneficial for customers and allow a business to advertise to them in new ways, while also attracting more serious customers. Reviews help producers attract serious customers and create strong customer relationships that will help the business in the future. [2]

Online reviews create the fundamental experience sharing forum between customers andproviders. Customers who submit evaluations frequently return to see how their remarks werereceived. Customers can build a link with an online business through this type of social involvement[2]. They also assist service companies in establishing a stronger web presence. Itaids in the improvement of their website's ranking. That is incredibly crucial to search engineslike Google, Yahoo, and Bing.

Related Works

Yang, L. [3] in "Sentiment analysis usingproduct review data," Xing Fang & Justin Zhan primarily used a feed-forward neural network equipped with an attention-based

encoder to overcome the abstractive summarization difficulty. evaluated the product reviews using sentiment analysis and rated each review with a star. Star ratings are given based on how positive or terrible the review is judged to be. Mason, R. et al. in "Micro summarization of Online Reviews: An Experimental Study," Sentiment analysis, summarization, and entity recognition are the three broad architectures that were combined to provide users with important content. The task at hand explores both supervised and unsupervised approaches.[4]

Shiva Kumar, K., Priyanka, M., Rishitha, M., Teja, D. D., &Madhuri, N. (2021)[5]. In the article "Text summarization with sentimental analysis" In order to save time and preserve the overall significance of the content, the researcher aims to create summaries for the vast amounts of text data. It makes it easier to understand the text's contents clearly. In order to solve the summary limit, our software additionally gives the customer the option to select the ratio and word count. The GUI implementation in the research is restricted to news items. The summary of the news story and its details, including the title, author, publishing date, and sentiment analysis, are intended to be displayed by the GUI. The solution is superior to other news summarizers since it incorporates sentiment analysis.

News-Focused Opinion Summarization Model

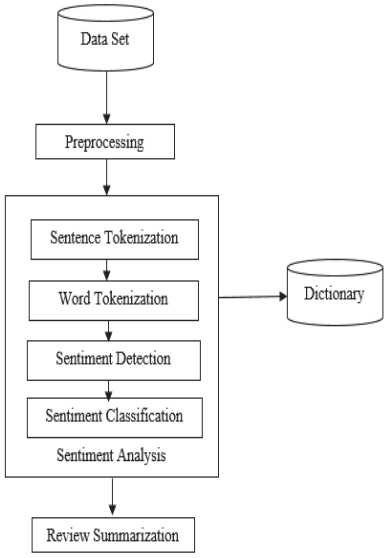

Fig 1: Proposed Methodology of the System

Fig 1 shows the diagram explaining the flow of the system. Social network data collection is a stage where data is collected from various sources. Data is collected in a variety of formats, including structured and unstructured data. Collected data is uploaded to HDFS for further processing. One of the most important stages of data processing is where data processing is done with data cleaning to confirm the accuracy of the data to ensure the efficiency and ease of the analysis process. Using pre-processing techniques, noisy and inconsistent data are reduced and converted to CSV format. The CSV file is uploaded to the HIVE. Considering the nature of the data, the researcher used multiple data sets one by one in the initial phase to draw conclusions.

Classification Analysis using Sentiment Analysis

Due to wireless technology, the Web is becoming a convenient place to learn digitally, share ideas and opinions on products and services. Online product and service ratings will reach millions, making it difficult to monitor and appreciate customer feedback. Sentiment analysis is a modern field of research that refers to natural language processing, computational linguistics and text analytics and classifies the polarity of the opinion stated[11]. Sentiment analysis is referred to as the process of extracting data, opinions, criticism. To predict sentence sentiment via the natural language process (NLP). In sentiment analysis, the text is divided into three stages: positive, negative, or neutral. The researcher[12] analyzes the data and describes “good” and “bad” feelings as positive and negative.

Proposed Classification Algorithm

Input: Dataset Input file

Output: Data set with polarity Positive, Negative and Neutral

Notations: P: dictionary word Polarity, R: Review, D: Dataset

Begin

Step 1: Transfer data to HDFS using Hadoop put command.

Step 2:Load data from HDFS to Hive Data warehouse

Step 3: Perform Sentence Tokenization

Step 4: Perform Word Tokenization

Step 5: While R in D do

while words in Review do

if word = = word in dictionary then
word_polarity = P;
end
end
Step6:Sum_polarity=sum(word_polarity_of_Review)
ifSum_polarity> 0.0 then
Given review is positive review
else if Sum_polarity< 0.0 then
Given review is a negative review
else
Given review is a neutral review
Step 7: End

Information handling could be a valuable asset that traces potential arrangements for the millions of individuals who work with different information applications and are continuously on the move and request speedy reactions. We are using a Hadoop HIVE for data taking care of purposes. The studies examiner stacked the data into the HIVE for handling from customers. The customer review and word reference tables need to be made utilizing the Hive arrange some time recently any data is stacked into them. The researcher carried out the tokenization of the sentence. One table is made, sentence_tokenize. Customer reviews are tokenized in this sentence, be that as it may they are furthermore cheapened. To tokenize words is just a strategy for tokenizing sentences freely of words. To perform tokenization, a word tokenize table has been made by the researcher. The query's words are isolated utilizing HIVE explodework . This inquiry is utilized to begin the MapReduce work. This inquiry joins tables to decide the extremity of each word word_tokenize and word reference. Sometime recently being relegated extremity amid the execution of the inquiry, each survey word is looked up in a word reference. After allotting extremity, the researcher asks whether customerreviewsmust to be categorized as positive, negative, or neutral. The common limit of each and each word within the review is enlisted to choose the conclusion of each overview. In case a reviews add up to extremity is more prominent than zero, it is considered positive.

ProposedAlgorithm for opinion Summarization

The process of predicting the solution of the review is accomplished through steps as follows.

1. Collect data from websites and convert it into CSV format.
2. Store data into HDFS using Hadoop put Command.
3. Loading the customer dataset from HDFS as input for the algorithm using HIVE.
4. Process Data in MapReduce Environment using HIVE.
5. Classify review dataset as Positive, Negative and Neutral using HIVE and store back into HDFS.
6. Consider negative reviews for further processing
7. Extract key Feature attributes from the review to build a predictive model.
8. Identify Topics and Sub Topics from the dataset and consider it as the training set.
9. Consider the negative review dataset as the testing set.
10. Map customer reviews to sum up the longer review.
11. Map customer reviews for predictive solution.

Review Data Processing

Every day retailers get hundreds of reviews through the company website and social media. The researcher performed the classification of reviews as positive, negative and neutral. Thechallenge to analyze this data and identifying the area to be the focus on the improvement. Using built-in features of the Python NLTK package we can identify topics using a classification-based approach. By taking into consideration of the identified topic, We can summarize the review based on mapped review data. To do so the researcher builds the dictionary by studying the datasets used for analysis. As this dictionary is developed by the researcher by its won but the organization can involve the people who are involved in handling feedback and making decision for preparing the same kind of dictionary. The researcher prepared a dictionary and found the key features shown in Figure 2 and Figure 3.

Data Structure for Review Mapping

In order to uniquely identify each review, the categorization features of the review include a region of the review that contains the customer review, a customer Id, and an arbitrary identifier. Positive(1), negative(-1) and neutral(0) sentiments are represented in the sentiment.

Extracting key features from Review

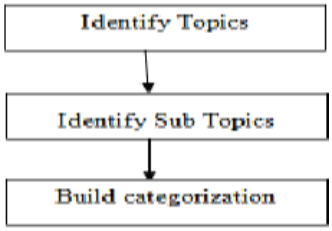

Fig 2: Key Feature

Finding the various subjects from the reviews is the first step. The issues in the reviews can be defined using a variety of fundamental techniques, including Word Frequency, Inverse Text Frequency, and more widely used ways like LDA. However, after examining the reviews, the researcher determined the subject, its subtopics, and its essential characteristics. Topic modeling techniques categorize subjects according to the keywords that are present in the text. Table 1 displays the subjects that were identified.

In order to determine the subtopics, the recognized topics listed in Table 1 will be taken into consideration. The main characteristics of the subtopics will solely depend on the subject.

Based on the topics, Build a categorization. Categorization can be considered as a network of topics, subtopics and keywords.

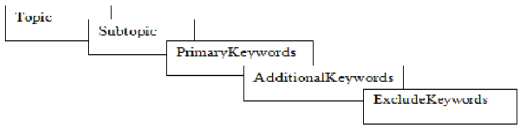

Fig 3 : Topic Hierarchy of categorization

A CSV file format is used for the categorization. Each subtopic has three layers of keywords: PrimaryKeywords, AdditionalKeywords, and ExcludeKeywords. The classification file is manually updated with the keywords for each topic. The TfIDf, Bigram frequencies, and LDA techniques can help you identify the right set of keywords. Although there's no easier technique to create keywords, some of the recommended strategies are available. The keywords that are mostly related to the subject are called primary keywords. Extra keywords are distinct for the subtopic. Although it is not necessary for these keywords to be exclusive of one another across themes, it is advised that they stay exclusive within the same sub-topic. Keywords that are used less frequently than the other two are known as exclude keywords. Snapshot of sample categorization:

Topic	Subtopic	PrimaryKeywords	AdditionalKeywords	ExcludeKeywords
staff	rude staff	impatient*	disappointed*	
staff	bad behavior	staff*	arrogant*,no politeness*	pathetic*,service*
staff	staff argue	not ready*,listen*,problem*,seriously*,packed*,no ramp*	push*,trolley*	floors*
staff	horrible staff behavior	rude*	bad*	staff*
staff	behavior behavior	not good*	space*	
staff	worst behavior	parking*	service*	
staff	worst behavior	people*	security*	
staff	bad behavior	very bad*,rude*,people*,management*,don't help*,customers*	no manners*	
staff	worst behavior	behavior*,no price*	tag*	product*
service	pathetic service	security*,guard*	billing*	behave*
service	space problem	inside*	space*	outside*
service	vegetable	fixed*	quantity*	
service	bad service	service*,stock*,price*	cheap*	quality*
service	customer service	service*,criminal*,security*,staff*,management*,useless*	canvas*	bag*
service	not good service	not good*,wants*,buy*,product*,wont*	go*	again*
service	worst service	experience*,mistake*	material*	pathetic*
service	space problem	parking*,worst*	customer*	service*
service	bad service	too much*,crowd*,week*,end*	poor*	management*
service	worst service	parking*	space*	
service	terrible service	experience*,employees*,products*	not found*	store*
service	parking problem	horrible*,crowd*,parking*,mess*	crowd*	management*
service	parking problem	not inform*,people*,cheat*	worst*	management*
service	worst service	stores*,gift*	packing*	policies*
service	disappointed service	security*,system*,parking*,system*,no space*	people*	park*

Fig 4:Key feature categorization

Mapping Customer Review for the Proposed Algorithm

The customer review analysis includes mapping one or more sub-topics for each customer. Some reviews may not have any corresponding feedback, which requires manual checking to ensure no taxonomy topics were skipped and they can be revised. For the proposed model, a researcher has used python programming to map customer reviews. The program utilizes a function that maps the reviews with categorized topics. The output of this function is a summarized version of the review, which is useful for them who need longer reviews to be condensed. The proposed model system is based on key features determined by the result.

Selection of Key feature Data Sets for Opinion Summarization

For the automatic review summarization from the mapped review, researcher have selected key feature attributes in terms of various words. Researcher have developed the key feature dictionary as shown in Figure 4. This dictionary is developed by focusing on the topic hierarchy as shown in Figure 2.

The frequent words of researcher develop dictionary are analyzed for attribute selection of customer review. The training data is separated based on the topic. The topics considered are shown in Table 1. The researcher has considered four key featured attributes and their related mapped key sets. Set 1 is formed using the staff key topic. Set 2 is formed using the service key topic. Set 3 is formed using a billing key topic and Set 4 is formed using a product key topic. These key topic words are frequently used words for the reviews.

Sr. No	Topic
1	staff
2	service
3	billing
4	Product

Table 1: Key Topic

These key feature attribute data sets are used for the proposed model of review summarization for the review. The proposed model is implemented using the Python programming language.

Topic	Word Set
staff	rude,impatient,disappointed,bad staff,arrogant, no politeness,pathetic,service,argue,not ready,listen,problem,seriously,trolley,horrible,rude,behavior, not good, space, worst, parking, service,worst,people,security,bad,very bad,rude,people,management,don't help,customers,no manners,worst,behavior,no price,tag,product
	pathetic,security,guard,billing,behave,space,inside space,outside space,vegetable, fixed, quantity, bad service,stock,price,cheep,quality,customer,service, staff,management,useless,bag,not good,wants to buy,product,wont,go,again,worst, experience,mistake,material,pathetic,space,parking worst,customer service,bad,too

service	much,crowd,week end,poor,management, worst,parking,space,terrible, experience,employees,products,not found,store parking,horrible crowd, parking,mess, crowd,management,not inform,people,cheat,worst management, worst,stores,gift, packing,policies,disappointed,security system,parking system,no space,people,park, less,counter,no space,stand crowd,too bad,bad,very rude,less space,parking worst, customer care,no space
bill	overbill,refund long,waiting period,counters close,public more wait,bad experience, disappointing,items charge higher,less counter,carry bag charging,behaviour,swipe twice,swiping wrong,items,not present cart,purchase,cashier wrongly,staff

Table 2: Extracted Key Feature Words

Summary Construction

The conception of the proposed model is based on the review summarization and topic-based text summarization approach, where the relevance scoring of sentences not only requires processing the information content to be summarized, but also requires to carry out an alignment process with external or contextual information of interest. An overview of the proposed model is shown in Figure 1.

The researcher deals with designing and implementing review summarization using HIVEand Python for summarization using reviews available on websites. To achieve this goal researcher has proposed a methodology using Hadoop and Python along. The review processing is done using the HIVE tool on the Hadoop platform with sentence, word tokenization which resulted in a positive, negative and neutral review. Then proposed models are constructed on Python and results generated in pre-processing are applied as a training set to test the series of data. The constructed model summarized the longer reviews based on Key Feature and Topic Categorization. The model generate the CSV file as a result of review summarization.

Conclusions

In this paper, we have proposed the algorithm for sentiment analysis for customer review. Based on the results of the evaluation of the opinions Naive Bayes 88.88% and SVM 87.18% of accuracy[6]. In analyzing sentiment, the biggest challenge was in the process

of pre-processing sentiment, especially for sentiment sentences that are not in accordance. For further research, the sentiment data to be processed should be in the form of sentences, not just one or two words. Therefore, it is necessary to carry out a post tagging process to determine at least a subject and predicate in each sentence.
The main objective of our project is to create summaries for the huge customer review text data which results in time saving and preserving its overall meaning. It helps in concise understanding of the textual data. The proposed algorithm of our project is confined only to customer reviews. This study allows organization to know their customers opinion also shorten the customer review by extracting their online review.

References

1. Shiva Kumar, K., Priyanka, M., Rishitha, M., Teja, D. D., &Madhuri, N. (2021). Text summarization with sentimental analysis. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.
2. SM, A. H. (2021, December). Summarization of Customer Reviews in Web Services using Natural Language Processing. In Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7-8 2021, Chennai, India.
3. Yang, L. (2016). Abstractive summarization for amazon reviews.
4. Wang, Q., Liu, P., Zhu, Z., Yin, H., Zhang, Q., & Zhang, L. (2019). A text abstraction summary model based on BERT word embedding and reinforcement learning. Applied Sciences, 9(21),4701.
5. Shiva Kumar, K., Priyanka, M., Rishitha, M., Teja, D. D., &Madhuri, N. (2021). Text summarization with sentimental analysis. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.
6. Rangari, K. T. (2021). Predictive Analysis for Retail Industry Using Big Data Technology (Doctoral dissertation).
7. RANGARI, K. T., & KAIWADE, D. A. (2018). From Deconstruction to Big Data: How Technology is Reshaping the Retail Industry. JournalNX, 85-88.
8. RANGARI, K. T., & KAIWADE, D. A. (2020). Literature Review on Sentiment Analysis in Retail Industry, International conference on Innovations in IT and Management (ICI2TM – 2020)

9. RANGARI, K. T., & KAIWADE, D. A. (2020). Sentiment Analysis of Customer Feedback using HIVE, National conference on Current Trends in Management-Change & Challenges
10. Madhuri, D. K. (2019). A machine learning based framework for sentiment classification: Indian railways case study. Int. J. Innov. Technol. Explor. Eng., 8(4), 441-445.
11. Pak, A., &Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
12. Liang, P. W., & Dai, B. R. (2013). Opinion Mining on Social Media Data. Mobile Data Management (MDM). In 2013 IEEE 14th International Conference on (Vol. 2).
13. Hsieh, H. T., &Hristova, D. (2022). Transformer-based Summarization and Sentiment Analysis of SEC 10-K Annual Reports for Company Performance Prediction.
14. Abuka, G. (2023). Text Summarization and Sentiment Analysis of Drug Reviews: A Transfer Learning Approach (Doctoral dissertation, Middle Tennessee State University).
15. Darmawiguna, I. G. M., Pradnyana, G. A., &Jyotisananda, I. B. (2021, March). Indonesian sentiment summarization for lecturer learning evaluation by using textrank algorithm. In Journal of Physics: Conference Series (Vol. 1810, No. 1, p. 012024). IOP Publishing.
16. SM, A. H. (2021, December). Summarization of Customer Reviews in Web Services using Natural Language Processing. In Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7-8 2021, Chennai, India.
17. Gawde, B., Motwani, S., Jadhav, A., &Makwana, B. Opsum: Topic Based Opinion Summarization And Sentiment Analysis.
18. Mabrouk, A., Redondo, R. P. D., &Kayed, M. (2021). Seopinion: summarization and exploration of opinion from e-commerce websites. Sensors,21(2),636.