ISSN No:-2456-2165

Customer E-mail Categorization and Topic Modeling

Eshita Gangwar[1], Rutvi Sutaria[2]

Abstract:- Customer service centers are most crucial part of any company. They represent the company and communicate with the customers on its behalf. These centers also impart valuable information through customer feedback. As they form the bridge between the company and its customers, it is important they convey information timely and effectively. Emails are the most popular means of business communication. In order to efficiently and effectively utilize time, the customer service centers need to extract the relevant information from these emails. Then they need organizing it and promptly respond to the customers accordingly. In this paper, we propose categorizing emails based on the customer reviews. The polarity (positive or negative) of the customer emails is determined along with its probability and also determine the topic of the email. After preliminary data pre-processing, we use the Naive-Bayes algorithm based approach to classify the emails and then the topic modeling is performed. This way, the project helps to classify emails and determine their subjects to save valuable time for the employees.

Keywords:- *Emails, Customer Care, Categorization, Topic Modeling.*

I. INTRODUCTION

Emails are convenient and the most popular means of communication for customers. They help in avoiding long waiting hours for telephonic conversation as well as in preserving the record of the communication that is happening between the customer service center and the customer. It is crucial that these emails be properly organized at the customer service so as to spontaneously and appropriately reply. Also, if properly utilized, these emails can yield invaluable data for the companies to access the needs and demands of the customers.As technology advances, the customers have more interactive and dynamic relation with the company. The amount of the emails received by customers at the customer service center is increasing rapidly. Handling this enormous amount of email manually can be a very time-consuming and complex task. The problem can be solved if emails are automatically classified on the basis of their content.

Sentiment Analysis, also known as opinion mining sometimes, is a method to assess written or spoken language to determine whether the expression is a positive, negative,or neutral, and to what extent. It is also the most common classification tool for textual analysis.Sentiment analysis is important because it helps businesses understand the likes and dislikes of customers about the company. It helps companies evaluate the social sentiment of their brand.Sentiment Analysis is crucial and should be treatedseriously as customer feedback contains a plethora of useful information which if properly harnesses can help companies achieve next to impossible goals. But, just knowing what customers are talking about it is not enough. Companies must also understand how they feel. Sentiment analysis is one way to discover these feelings. The customer service center employees can thus determine the sentiment of the email and prioritize or respond appropriately.

Massive amounts of data are collected on daily basis in so many companies. The huge amount of information makes it difficult to understand the data or to find what we are searching. We have to employ different methods to organize the data systematically and extract relevant information from it.Topic modeling is one such powerful technique which allows us to not only organize but also summarize large amount of textual data. It is helpful for determining different topical patterns present in the dataset which otherwise remain invisible.Topic modelling is a method for finding subjects (i.e. topics) from a data that best describe the content present in the document. There are many methods which can be used to obtain topic models. For our project, we will be using Bag of Words technique to find the topics of the emails at the customer service centers. This will help the employees to summarize the content of the emails.

Today, the algorithm-based sentiment analysis tools can handle huge volumes of customer feedback consistently and accurately. Paired with topic modeling, sentiment analysis reveals the customers opinion about topics regarding different products and services. In our research, we provide with various features for extraction for Naive Bayes.

Algorithm based classifier for the project and the subject of the email through topic modeling. This classification helps the customer service employees to prioritize emails. It also assists them to assess the customer responses to particular products and formulate market strategies in accordance. Furthermore, the customers inputs are useful as they inadvertently provide the businesses with insightful information regarding the current trends and requires of customers.

II. PREVIOUS WORK

A. Polarity Categorization on Product Reviews

The focus of the research in this paper was to tackle the problem of sentiment polarity categorization. The dataset is collected from amazon.com. The dataset contains 376 instances of reviews of Nokia mobile in the form of a text file. Two classification algorithms namely Nave Bayes and Support Vector Machine Algorithms are taken to classify the reviews as positive, negative or neutral.

B. Approaches, Tools and Applications for Sentiment Analysis Implementation

In this paper,Andrea et al. (2015) describe the different approaches and tools available for Sentiment Analysis.Their elaborate on different approaches that can be used for analyzing sentiment along with the different types of features and techniques associated with these and also the corresponding advantages and disadvantages associated with these different types of sentiment analyzing approaches.They also give an overview of different tools used for Sentiment Analysis over a period of time with respect to the different methods which can be used. Furthermore, they describe different fields where sentiment analysis can be applied. Some of these areas are politics, finance, public actions and business and many other fields as well.The sentiments can be classified as positive or negative depending upon the content of the document. In their research work, the authors of this paper define three different levels of classification of sentiment. They suggest that the sentiments in a text can be classified as the document level, the sentence level and the aspect level.The paper also explains the different approaches as one of the major task of the project, like Machine learning approach, lexicon based approach and a hybrid of these both. A detailed study is done on these and their advantages and limitations are discussed. Also different features and techniques related to each of the approach is described in this work.It also describes in detail the tools used for analysis, as the other major part of their research work, such as EMOTICONS, LIWC, SentiStrength, SentiWordNet, SenticNet and Happiness Index The paper basically classified Sentiment Analysis depending upon different approaches and tools.

C. Sentiment analysis using product review data

The aim of this paper is to tackle one of the fundamental problems of sentiment analysis.Fang and Zhan (2015) aim to find a solution for one of fundamental problems of sentiment polarity categorization.A process of sentiment polarity categorization is outlined in this paper along with which a detailed descriptions of the different processes involved at different stages is also explained.They use a dataset in this study which contains product reviews available from Amazon.com. This dataset is available online and is provided by the company itself to be used in general for research.For their work, they conduct experiments for both sentence-level and review-level categorization on the collected dataset. At last, they also provide insights into future work to be conducted on sentiment analysis.They initially propose an algorithm. This algorithm is executed later for identifying negation phrases. This is followed by implementation of mathematical approach which helps in computing the sentiment score on the product reviews from the dataset.Later, a method for creating a feature vector is shown for categorization of sentiment polarity.After this initial implementation of algorithm, two experiments are conducted to categorize sentiment polarity. They are based on two levels: sentence level and review level. Performance of all the three different classifiers are compared to one another and evaluated based on their corresponding experimental outputs.The three different classifiers used in this paper for their research are Naive Bayesian classifier, the Random Forest classifier and the Supervised Vector Machine classifier.The experiments are conducted on these classifier and their results are compared with one another to understand the classification better.

D. Emotion Detection in Email Customer Care

The paper shows how to extract salient features and identify emotion in emails at customer service centers. these features show customer anger, dissatisfaction with the business, and warnings like take legal action, report to higher authorities or to leave.

E. Using Text Mining for Automated Customer Inquiry Classification

This paper has illustrated the use of customer inquiry classification and intention analysis on customer inquiries, primarily in the form of unstructured multi-lingual text data. Inquiry classification helps isolate inquiries related to the product. Intention analysis isolates sales/procurement/licensing related queries within these. The paper has demonstrated an automated way of such analysis resulting in significant savings of manual efforts, without compromising on the accuracy of the analysis.

III. APPROACH

In our work, the email can be viewed as a set of sentences. The purpose of our project is to help find the topic of the email and also determine the polarity of the email.For our project, we are using a dataset from customer service centers for four different products. The emails contain customer feedback on an MP3 player, a camera, a mobile phone and a router.The first procedure is of topic determination. In this module, the email dataset isconverted into a form which is easy to understand for finding the probable topic. Thus data pre-processing is performed.Firstly, we will employ the technique known as the tokenization. The technique we are using in our work is called the word tokenizer. All the tokens that have been extracted may not be useful for our classification. Therefore the unnecessary words,known as the stop words, are eliminated using the technique known as stop words removal.The remaining

tokens are subjected to the process of stemming. Stemming attempts to reduce a word to its root form. Therefore, the document is then represented in the form of root words rather than the original words. Other method called Part of speech tagging is then applied. POS tagging associates a word in the test to the corresponding part of speech. The topic is then determined using Bag of Words technique.Remaining tokens after pre-processing can express a negative or positive opinion of the customers. For example, upset represent negative emotion while satisfaction demonstrates positive opinion.This is then fed into the Naive Bayes classifier in the Sentiment Analysis module.Naive Bayes classifier is trained using a feature set which then determines the polarity of emails from the incoming preprocessed datasets as either positive or negative.

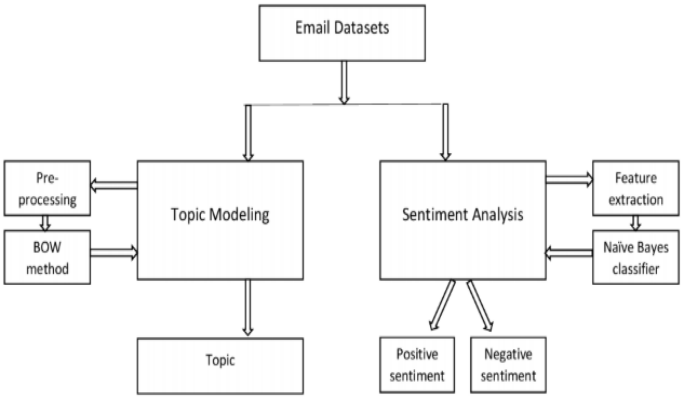

Fig 1:- Overview of the process

IV. ALGORITHM

A. Naive Bayes

Naive Bayes algorithm is one of the Bayesian theorems and a supervised machine learning technique. It can be used for binary as well as multiclass classification problems. It is named Naive because it simplifies the probabilities for all hypothesis to make their corresponding calculation easy. The algorithm works by finding out the probability of the different attributes of data being associated with the certain class. Naive Bayes classifier works under the assumption that the presence of one feature in a class is independent of the presence of other features. The classier selects the most likely classification for a given set of the attribute values.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

- P(c—x) is the posterior probability of class(target) given predictor(attribute).
- P(c) is the prior probability of class.
- P(x—c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

B. Bag of Word

Topic Modeling is a Machine Learning technique used in NLP Natural Language Processing(NLP) where the ”topic” or the subject of a set of documents is determined. BOW is a type of topic model that uses frequency to show why certain parts of data occur repeatedly. In BOW each email document is assumed to be composed of a set of different words. The topics are determined from the content of the emails in the corpus. These words are used as features which are then used for training. These topics then produce the subjects. BOW identifies topics based on occurrence of different terms which are automatically determined. It basically gives different topics that would represent the document.

V. DATASET

The dataset contains around 200 emails. These emails are comprised of four different datasets.The first dataset contains emails which contain customers review for an MP3 player.The second dataset belongs to a camera manufacturing company. The dataset contains reviews regarding a particular camera. The third dataset consists of emails collecte from the customer service center of a mobile company for a particular mobile phone.The final dataset contains reviews for a router.

VI. EXPERIMENT

A. Sentiment Module

This module helps to determine the polarity of the incoming customer emails with the help of Naive Bayes classifier.Firstly, we need to find a method to train our Naive-Bayes classifier to classify our mails as positive or negative. For this purpose, we use the words in the positive and negative reviews as their features. The classifier, after training, has thus learned to associate the words from positive reviews as features for positive sentiment and similarly those from negative reviews as features for negative sentiment.The feature set, which contains the features extracted after training the classifier, is then loaded onto the module.The Naive Bayes algorithm based classifier is then used to determine the sentiment of the emails based on the content from the customers.The function sentiment() is then defined where the new text is finally classified.

B. Topic Module

This module contains two distinct functions; clean() and topic().In the clean function, the data is pre-processed so that only relevant words are retained for topic determination. Pre-processing involves processes such as tokenization stop-words removal, stemming and POS tagging. After the cleaning of the dataset, it is now ready for determination of topics and is thus sent to the topic function The topic function, which uses the Bag of Words approach, is then used to choose the most appropriate topic for the corresponding emails.

VII. RESULT

In our project we analyzed on four different datasets (1. MP3 Player, 2. Camera, 3. Mobile phone, 4. Router). The analysis module 1 displayed the polarity of the emails and the corresponding suggested topics by the Bag of Words (BOW) technique. It also showed individual pie-charts for each of the datasets. In most charts, the amount of positive customer review is lesser than the negative customer reviews. The topics show a trend with certain topics occurring repeatedly as compared to others.The topics show a trend with certain topics occurring repeatedly as compared to others. The similar topics are mostly associated with the same sentiment. (For example,in dataset 3, the topics battery and charger occur repeatedly and all are associated with negative sentiment.)

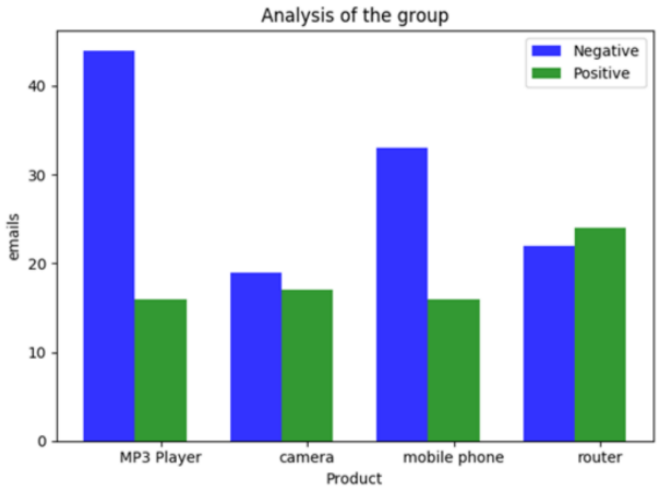

Fig 2:- Comparison graph of all Dataset

Polarity	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Positive(in percentage)	26.7	47.2	32.7	52.2
Negative(in percentage)	73.3	52.8	67.3	47.8

Table 1:- Polarity for each Dataset

VIII. CONCLUSIONS

The research paper focuses on quick and efficient email classification at customer service centers. The emails are classified as positive or negative with the use of Naive-Bayes algorithm based classifier. Also, the subject of the email is determined through topic modeling using LDA technique. Ultimately, the outcome will display the polarity of the email with the probability and suggested topics with their probability count. The topics show a trend with certain topics occurring repeatedly as compared to others. The similar topics are mostly associated with the same sentiment. (For example, in dataset 1, the topics volume and sound occur repeatedly and all are associated with negative sentiment.) Our work will help the customer service employees to get an overview of the content of the incoming customer emails and organize them according to their needs and/or interests. It also helps to reduce their corresponding response time and to analyse the customer feedback on a product or its parts.

FUTURE WORKS

The work can be extended by adding some functionality for the user, such as an interface, so the complaints can be arranged according to the priority. The paper can also help in further categorization depending on the departments within the companies. The emails at the customer service centers can be handled automatically. For this, templates can be created which will help to respond to distinctive types of emails appropriately. Also, in future more data analysis can be performed on the data derived from customer service centers as they can serve as data sources to analyze customer needs and future trends.

REFERENCES

[1]. Alghamdi, R. and Alfalqi, K. (2015). A survey of topic modeling in text mining. International Journal of Advanced Computer Science and Applications, 6, 147153.
[2]. Andrea, A., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125.
[3]. Fang, X. and Zhan, J. (2015). Approaches, tools and applications for sentiment analysis implementation. Journal of Big Data.
[4]. Gupta, N., Gilbert, M., Fabbrizio, G. D., Wu, Z., and Serker, N. H. M. K. (2009). Emotion detection in email customer care. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 1016.
[5]. Jetley, R. P., Gugaliya, J. K., and Javed, S. (2015). Using text mining for automated customer inquiry classification. Journal of Vibration and Acoustics, 122, 4651.
[6]. Joty, S., Carenini, G., Murray, G., and Ng, R. (2009). Finding topics in emails: Is lda enough?. NIPS-2009 workshop on applications for topic models: text and beyond.
[7]. Mixymol, V. (2017). Polarity categorization on product reviews. SInternational Journal on Recent and Innovation Trends in Computing and Communication, 5, 2831.
[8]. Subramanian, S. K. and Ramaraj, N. (2007). Automated classification of customer emails via association rule mining. Information Technology Journal, 6, 567572.
[9]. Xie, P. and P.Xing, E. (2013). Integrating document clustering and topic modeling. Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, 29.