International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

Sentiment Analysis For Customer service

E.kodhai, B.nivetha, K.sriakila, G.suvalakshmi

Abstract: — NLP can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation. By utilizing Natural Language Processing the customer experience will improve . E-mail is still the most commonly used digital customer service channel 54% of customers have used E-mail customer service channel. The proposal of this paper is to develop an algorithm where customer E-mails are scanned, analyze the sentiment from the body of the message and automate customer e-mail categorization and prioritization for the banking sector. The main goals are to collect bank query related E-mail data, ranging from general information, escalations and request, Develop a machine learning algorithm that can perform text mining and sentimental analysis, Provide priority based categorized information for management to prioritize and improve customer service.

Keywords – e-mail categorization, priority, sentiment analysis.

I. INTRODUCTION

Customer service is the provision of service to customers before, during and after a purchase. According to Turban et al[1] , "Customer service is a series of activities designed to enhance the level of customer satisfaction – that is, the feeling that a product or service has met the customer expectation." Customer satisfaction is an ambiguous and abstract concept and the actual manifestation of the state of satisfaction will vary from person to person and product/service to product/service. The state of satisfaction depends on a number of both psychological and physical variables which correlate with satisfaction behaviours such as return and recommend rate. The level of satisfaction can also vary depending on other options the customer may have and other products against which the customer can compare the organization's services.

For any successful company, customer service is a core pillar and it plays a pivotal role in how a company is perceived by its customers and in turn the long term revenue. How a customer request is handled speaks volumes about how committed a company is in pursuing customer delight. In a nutshell, the role of a customer service department in an organization is to interact with the customers, answer questions, resolve support issues, establish credibility, and nurture relationships. The most important objective for the customer service is to be timely and helpful[2]. With E-mail channel constituting more than half customer service requests, we could utilize AI and machine learning to further enhance how companies and its customer service departments respond to E-mail requests. Let's take an example of existing use case of AI powering customer service processes. Digital Genius[6] is an AI-powered customer service tool that uses machine learning and natural language processing to completely automate the consumer support process. Using a proprietary technology referred to as Conversational Process Automation, the AI platform is capable of understanding conversations, automating repetitive tasks and personalizing user interactions for the purpose of assuring quality service. Our primary goal is to improve customer experience by utilizing AI and machine learning to automate customer e-mail categorization and prioritization for the banking sector. We are trying to collect bank related enquiries from consumers and analyze them to process at a faster rate and to avoid human intervention in categorizing those emails and forwarding to the concerned departments for faster action. Customer service plays an integral part in maintaining the consumers and especially where there is money involved in the process. It is very crucial and critical to have the consumers satisfied as early as possible whenever a complaint or request is raised from them.

The section two follows the related works that are referred and analayzed for our proposal paper .The existing paper's of the merits and demerits were found and the demerits are overcome in our proposal paper. The section three follows proposed system and it follows by the third section modules and conclusion.

II. RELATED WORK

Anju Radhakrishnan and Vaidhei[9] proposed in a paper that spam messages plays a major problem in email classification. So in their paper they overcome the spam messages by email classification with the help of machine learning algorithm. In machine learning , they used Naïve Bayes and J48 decision tree are done to test their efficiency of email classfication as spam. Their experiment mostly focused on with two combinations with preprocessing techniques and text categorization.

Revised Manuscript Received on April, 02 2020.
* Correspondence Author

E.kodhai*, computer science and engineering, sri manakula vinayagar engineering college, pudhucherry, india . Email:kodhaiej@gmail.com

B.nivetha, computer science and engineering, sri manakula vinayagar engineering college, pudhucherry, india . Email:nivetha.birundha@gmail.com

K.sriakila, computer science and engineering, sri manakula vinayagar engineering college, pudhucherry, india . Email:srriakila1998@gmail.com

G.suvalakshmi, computer science and engineering, sri manakula vinayagar engineering college, pudhucherry, india . Email:suvi9252@gmail.com

They also used Enron Corpus as a data set. The TD-IDF value is taken as a score value,then the value is taken as minimum size of email classification time. But the have classified only the spam e-mails.

Slhami Sel et al [15] proposed that on ignoring spam e-mails, there are remaining lots of e-mails received every day.In order to know the importance of received e-mails,the subject or body of each e-mail must be checked . In this study they proposed an unsupervised system to classify received emails.Received e-mails' texts are determined by a method of NLP called as Word2Vec algorithm. According to the similarities, processed data are grouped by kmeans algorithm with an unsupervised learning model. The subject or content of each email must be checked.

Muhammad Ali Hassan and Nhamo Mtetwa [16] proposed that the use of different feature extraction methods coupled with two different supervised machine learning classifiers evaluated using four performance metrics on two publicly available spam email datasets for spam filtering. They highlight the importance of the correct coupling of feature extraction and classifier, and the merits of using two independent datasets. In their paper, they use four performance metrics: accuracy, precision, recall and f1- score and they used the datasets for the combination of Ling-spam and Enron Corpus.

III. PROPOSED SYSTEM

The proposal of the paper is to address common customer frustration points with respect to customer service in banks. So, there is a need to improve Customer Service department's responsiveness to e-mail queries/escalations by better data categorization and prioritization.

Here the sentimental analysis is used to analyze customer E-mails, assign priority, judge sentiment in the communication and divert the E-mail to relevant departments for further processing, email classification in general, particularly utilized natural language processing and prioritization by sentiment analysis. By extracting data from e-mails and perform sentiment analysis . In sentimental analysis, the emails can be classified into cards , loans, account , pins and it can be prioritize by scoring matrix algorithm.

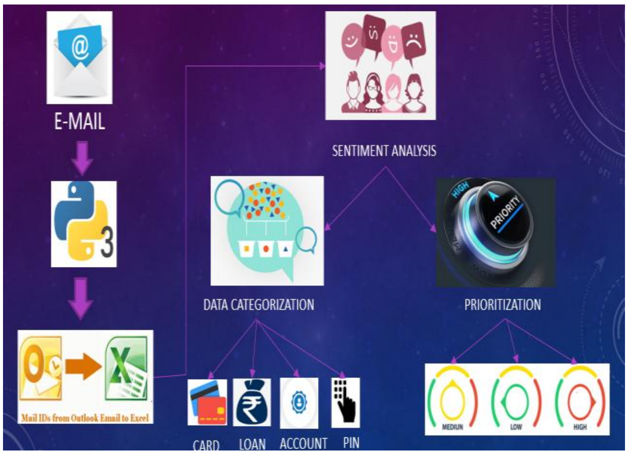

Fig 1 Architecture of Proposed System

Collect bank query related Email data ranging from general information ,escalation and request. Fig .1 shows as Data extractions is solution to get the content generated automatically by importing it directly from software .after importing of data into access ,we then exported it as an excel worksheet. the data were sentimentally analyzed and move on to the next step which is data categorization and prioritization .as we have taken banking as our domain so data categorization were on fields such as credit, loan, account, pin etc,. prioritization is done according to the urgency and importance and we can give three levels of priority highly critical, critical, low critical.

The modules of the proposed paper are as follows:

1.Sentimental analysis
2.Text mining
3.Corpus
4.Frequency of words
5.Polarity classification
6.Time series analysis

3.1 SENTIMENT ANALYSIS

Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression[5]. Polarity: if the speaker expresses a positive or negative opinion

Subject: the thing that is being talked about
Opinion holder: the person, or entity that expresses the opinion

Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs and social media. With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service.

Text information can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about something. Opinions are usually subjective expressions that describe people's sentiments, appraisals, and feelings toward a subject or topic.Sentiment analysis, just as many other NLP problems, can be modeled as a classification problem where two sub-problems must be resolved.

- Classifying a sentence as *subjective* or *objective*, known as **subjectivity classification**.
- Classifying a sentence as expressing a *positive*, *negative* or *neutral* opinion, known as **polarity classification**.

In an opinion, the entity the text talks about can be an object, its components, its aspects, its attributes, or its features. It could also be a product, a service, an individual, an organization, an event, or a topic. To perform the sentiment analysis, the data collected has been made to go through the below steps to obtain efficient results from the data.

- Text Mining
- Corpus Building
- Sentiment Analysis and Polarity classification

3.2 TEXT MINING: BAG OF WORDS

Text mining is the process of deriving high quality information from text. High quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data and finally evaluate and interpret the output[17]. Text mining was done on the dataset to preprocess the data and have it prepared for sentiment extraction and analysis. There are several preprocessing functions to preprocess the text data. Text stored in the Body column of the email data is cleaned up for further analysis. Two libraries were used majorly to clean up and preprocess the data. They are gsub and tm. Gsub library: gsub() function replaces all matches of a string, if the parameter is a string vector, returns a string vector of the same length and with the same attributes (after possible coercion to character). Elements of string vectors which are not substituted will be returned unchanged (including any declared encoding). Using the gsub library, the punctuations, digits, hyperlinks and other additional unwanted strings have been removed from the dataset.

TM library:

Tm library has several functions that will preprocess and clean up the data. Below are the tm_map library functions that are used in the dataset to preprocess the email body content.

Removing punctuations – using the removePunctuation function, the body content of the email has been stripped off punctuations and the data frame is now without any punctuations.

Removing Stopwords –Stopwords considered as noise in the text. Text may contain stop words such as is, am, are, this, a, an, the, etc which is stored in the default dictionary as English. The common words can be removed using this function remove Words and any other words that needs to be removed can be added to the stop words dictionary.

Perform Stemming - Stemming is a process of linguistic normalization, which reduces words to their word root word or chops off the derivational affixes. Stemming helps us increase accuracy in our mined text by removing suffixes and reducing words to their basic forms For example, connection, connected, connecting word reduce to a common word "connect".

3.3 CORPUS

A corpus represents a collection of texts, typically labeled with text annotations. A Corpus contains two types of metadata. Corpus metadata contains corpus specific metadata in form of tag-value pairs as shown Fig 2. Document level metadata contains document specific metadata but is stored in the corpus as a data frame. Document level metadata is typically used for semantic reasons or for perform reasons[18].

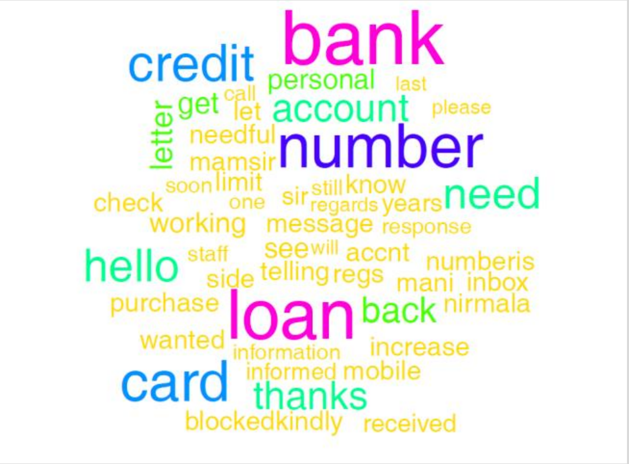

Fig. 2 Corpus

3.4 FREQUENCY OF WORDS

Now that the corpus is built and the word cloud of the data is visualized, the frequency of words occurring across all documents have to be captured and there are three important terminologies to store the corpus with respect to terms.

- Document Term Matrix
- Term Document Matrix
- Term Frequency – Inverse Document Frequency

A Document Term Matrix (DTM) or a Term Document Matrix (TDM) is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a DTM, rows correspond to documents in the collection and columns correspond to terms whereas in a TDM, rows correspond to terms and the columns correspond to the documents in the collection. TF-IDF is a matrix that diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. It can be quantified as an inverse function of the number of documents in which it occurs. For the model, we created a Document Term Matrix to find out the frequency of the more common words and for calculating the polarity using the frequency.

3.5 POLARITY CLASSIFICATION

The main objective of the sentiment analysis was to develop a reliable and scalable mechanism where customer e-mails are scanned for indicators of priority and sentiment i.e. if the text is indicating a positive or negative customer experience. Based on this mechanism, priority can be assigned, and sentiment for each customer e-mail[19]. For the team to build such a model, multiple steps were to be considered, like coming up with a sentiment scoring matrix, an algorithm which would assign criticality to each of the E-mail based on the scoring matrix.

With respect to the sentiment score values scores were categorized into the following categories:

- Positive – Greater than 1
- Negative – lesser than .75
- Neutral - .75 to 1

Once the sentiment values are obtained, then the important step to be performed before assigning the polarity to the body of the emails is to remove the Sparse terms. Sparse terms are terms which have a low frequency in the DTM so that the generalization of a model is improved. This is done using the remove Sparse terms function in the tm_map library and the level at which the frequency is to be filtered has to be specified. In the model, we have used 0.995 as the sparse value.

Based on the above sentiment scores, the criticality factor has to be assigned based on the five categories given below and written into the csv file which would help for final visualization and also to identify how the sentiment of the emails are distributed across the categories.

- Highly Critical – Less than .5
- Critical - .5 to 1
- Priority 1 – Equal to 1
- Priority 2 – 1 to 2
- Priority 3 – Greater than 2

A Lexicon is the collection of information about the words of a language about the lexical categories to which they belong. Lexicon is the total stock of words and word elements that have a meaning which helps us in deciding the subjective classification of the Sentiment Analysis.

3.6 TIME SERIES ANALYSIS

Data collected initially has been explored for possibility to run a time series model and forecast so that the model can be well suitable for a business scenario. Knowing only the sentiment of E-mails received will only help the organization to understand the criticality and the customer satisfaction of the banks. But if people are to be managed properly and to allocate work in a business environment sector knowing the expected number of E-mails is an important number and plays a significant role for day to day planning of an Operations team. Hence from the data collected, the number of e-mails received in the alias account is taken for analysis and a time series model is built. From the E-mails collected, the date column containing the date of the E-mails received and the number of E-mails on a particular day is calculated using the count function. The extracted values of the date and number of E-mails is stored as a new data frame which is then used for time-series analysis. The new data frame contained is explored and the initial tests are done to meet the requirements of time-series analysis. The date column has to be formatted in a proper date format which is suited for R to explore the data on a day to day basis and also the time factor present in the dates is stripped off from the column of values. Once the date column is sorted out properly, the data frame has to be converted to a time series before models could be built for analysis. Using the xts, ts and zoo libraries in R, data frame has been converted to a time series and then the time series is plotted to have an idea of the trend in the E-mails on a day to day basis.

IV. FEATURES

The potential features are:

- Faster and efficient queue processing
- Digitalized workflow for queue
- Cost savings for the bank by reducing the manual bandwidth
- Touch time reduction of a ticket
- Criticality based solutions
- Improved customer service
- Improved response time

V. CONCLUSIONS

By improving customer service first, the satisfaction towards the bank will improve if the tickets are solved at a faster rate. From the data collected, the functional aspects of the bank has degraded which prompts the management to look into critical areas and take necessary action for improvement. , the project could be used by any other customer service domain with slight fine tunings to the models used and would serve as a good business improvement scheme for both the top and middle management. However, while the team managed to collect E-mails from real banking customers, the trend and flow of E-mails to the inbox might not be the most accurate depiction of a live environment. The data collected is to be considered as representative of the live scenario and could vary from bank to bank and its respective services

REFERENCES

1. https://www.helpscout.com/75-customer-service-facts-quotes-statistics/
2. https://www.groovehq.com/support/customer-support-statistics
3. https://experiencematters.blog/2018/04/10/report-state-cx-management-2018/
4. http://www.iibf.org.in/documents/reseach-report/Report-26.pdf
5. https://monkeylearn.com/sentiment-analysis/
6. https://www.forbes.com/sites/julianmitchell/2018/09/05/how-ai-machine-learning-and-other-disruptive-trends-are-defining-the-future-of-customer-service/#24fcd1e54cdf.
7. M.Rajman , R.Besancon " Text Mining : NLP Techniques and Text Mining Applications " Jan2014.
8. Trupti G. Ghumade , R. A.Deshmukh " A Document classification using Natural Language Processing and Recurrent Neural Networks " Aug 2019.
9. Anju Radhakrishnan , Vaidhei " Email Classification Using Machine Learning Algorithms" Oct 2019.
10. Xi An Jiatong "A Comparative study for Content Based Dynamic Spam Classification Using Four Machine Learning algorithms" Feb2008.
11. Tiwo Ayodele , Rinat Khusainov "Email classification of summarization" Aug 2008.
12. EricHorvitz ,Kirkland Andrew "Notification And Interaction with the prioritized messages" Oct 2006.
13. N Mohammed Abu Basim,nishanth Solomon "Autobot for Precision Farming" Feb2007.

14. Rijul Dhir , Anand Raj "Movie Sucess Prediction Using Machine Learning Algorithms and Their Comparsions" Jan2018.
15. Slhami Sel ,Davut Hanbay" E-Mail Classification Using
16. Natural Language Processing"March 2019.
17. https://ieeexplore.ieee.org/document/8703222 2018..
18. https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing.
19. https://www.geeksforgeeks.org/nlp-custom-corpus/
20. https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

AUTHORS PROFILE

Dr.E.Kodhai is currently working as Associate Professor in the Department of Computer Science and Engineering at Sri Manakula Vinayagar Engineering College affiliated to Pondicherry University, Puducherry, India. She has completed her M.C.A from Cauvery College for women, Trichy affiliated to Bharathidasan University, Trichy, M.E. in Computer Science and Engineering from Vinayaka Mission's Kirupananda Variyar Engineering College, Salem and Ph.D in Computer Science and Engineering from Pondicherry Engineering College, Puducherry. She has 19 years of experience in teaching in various engineering colleges. Her Research interests includes Software Clones and Software Engineering. She has published more than 50 papers in international conference and journals.

B.nivetha, Studying B.tech (computer science and engineering)in sri manakula vinayagar engineering college .her area of interest artificial intelligence Her specialized area is networking and database

K.sriakila, Studying B.tech (computer science and engineering)in sri manakula vinayagar engineering college .her area of interest artificial intelligence Her specialized area is cloud computing

G.suvalakshmi, Studying B.tech (computer science and engineering)in sri manakula vinayagar engineering college .her area of interest artificial intelligence Her specialized area is database

Retrieval Number: D7287049420/2020©BEIESP
DOI: 10.35940/ijeat.7287.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.