International Journal of Progressive Research in Science and Engineering
Volume-1, Issue-3, June-2020
www.ijprse.com

Bank Customer Complaints Analysis Using Natural Language Processing and Data Mining

Chandana C[1], Neelashree N[1], Nikitha G N[1], Nisargapriya J[1], Vishwesh J[2]

[1]*Student, Department of Computer Science Engineering, GSSSIETW, Mysuru, Karnataka, India.*

[2]*Assistant Professor, Department of Computer Science Engineering, GSSSIETW, Mysuru, Karnataka, India.*

Corresponding Author: chandananayak.08@gmail.com

Abstract The banking sector has undergone a major revolution with the advent of digital transformation. The entry of Fintech and tech giants such as Google, Amazon, and Facebook have introduced convenient banking that is easy to understand and use. In this competitive environment, banks are realizing the importance of customer service and satisfaction and want to pay close attention to the Voice of Customer to improve the customer experience. By analyzing and getting insights from customer feedback, banks will have better information to make strategic decisions. In their quest to better understand their customers, banks are seeking artificial intelligence (AI) solutions in the form the of sentiment analysis. What is sentiment analysis? In simple words, sentiment analysis is the process of detecting a customer's reaction to a product, brand, situation or event through texts, posts, reviews, and other digital content. Using sentiment analysis, business leaders can gain deep insight into how their customers think and feel. The analysis can help in tracking customer opinions over a period of time, determine customer segmentation, plan product improvements, prioritize customer service issues, and many more business use cases.

Key Words:- ***t-SNE(t-Distributed Stochastic Neighbour Embedding), LDA(Latent Dirichlet Allocation), NLP(Natural language processing).***

I. INTRODUCTION

The rapid increase in the quantity of customer data has promoted the necessity to analyze these data. Recent progress in text mining has enabled analysis of unstructured text data such as customer suggestions, customer complaints and customer feedback. Much research has been attempted to use insights gained from text mining to identify customer needs to guide development of market-oriented products. However, the previous research has a drawback that identifies limited customer needs based on product features. To overcome the limitation, this paper presents application of text mining analysis of customer complaints to identify customers' true needs by using the Outcome-Driven Innovation (ODI) method. This paper provides a method to analyze customer complaints by using the concept of job. The ODI-based analysis contributes to identification of customer latent needs during the pre-execution and post-execution steps of product use bycustomers that previous methods cannot discover. To explain how the proposed method can identify customer requirements, we present a case study of stand-type air conditioners. The analysis identified two needs that experts had not identified but regarded as important. This research helps to identify requirements of all the points at which customers want to obtain help from the product.

For example, a sentence like "The customer service of XYZ bank is frustrating" – the system identifies "customer service" as a feature, "XYZ bank" as the object, and "frustrating" as a negative opinion. The algorithm arrives at a relationship between the opinion and object to extract relevant information. Today, several banks study and track customer behaviour through websites, transactions, voice notes, social media, and other digital channels. The aim being, to map and monitor a customer's journey with a bank and how those paths affect the quality of service or the sale of financial products and services. Financial institutions are collecting data through polls or interviews to capture customers' opinions towards specific product or service. Analysing the unstructured data through semantic processing offers a comprehensive View of customer satisfaction; classifying it under negative, neutral and positive feedback. Using the insights, banks can deliver better customer service.

II. METHODOLOGY

This project is implemented using java programming language. Both servlet and JSP technologies are used to create a web application. Servlet are java programs are precompiled which can create dynamic web contents. There are many interfaces and class in the servlet API such as http servlet, servlet request, servlet response etc. JSP is used to create a

web application just as servlet.it can be thought of as an extension to servlet because it provides more functionality than servlet. MySQL server is used as a backend.

A. t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-Distributed Stochastic Neighbour Embedding (t-SNE) is a non-linear technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. It is extensively applied in image processing, NLP, genomic data and speech processing. To keep things simple, here's a brief overview of working of t-SNE:

The algorithms start by calculating the probability of similarity of points in high-dimensional space and calculating the probability of similarity of points in the corresponding low-dimensional space. The similarity of points is calculated as the conditional probability that a point A would choose point B as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian (normal distribution) centered at A.

It then tries to minimize the difference between these conditional probabilities (or similarities) in higher-dimensional and lower-dimensional space for a perfect representation of data points in lower-dimensional space.

To measure the minimization of the sum of difference of conditional probability t-SNE minimizes the sum of Kullback-Leibler divergence of overall data points using a gradient descent method.

Production Function:

1. Collected datasets of bank customer's complaints from Kaggle
2. Pre-processing of obtained datasets
3. Sentence Segmentation
4. Convert Sentence segmentation into Tokenization
5. The selected datasets are catharized using nltk library
6. T-nse model applied for deploying into model The obtained result is showed in the graph

General Constraints:

The results generated have to be entered in to the system and any error or any value entered out of the boundary will not be understood by the system. In any case if the database crashes, the whole information collected and the results generated will be of no use.

B. LDA Algorithm

LDA is a form of unsupervised learning that views documents as bags of words (that is order does not matter). LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words. Now you may be asking "ok so how does it find topics?" Well the answer is simple: it reverses engineers this process. To do this it does the following for each document *m*:

1. Assume there are *k* topics across all of the documents
2. Distribute these *k* topics across document *m* (this distribution is known as α and can be symmetric or asymmetric, more on this later) by assigning each word a topic.
3. For each word *w* in document *m*, assume its topic is wrong but every other word is assigned the correct topic.
4. Probabilistically assign word *w* a topic based on two things:
   - what topics are in document *m*
   - how many times word *w* has been assigned a particular topic across all of the documents (this distribution is called *β*, more on this later)
5. Repeat this process a number of times for each document and you're done!

C. Natural Language Processing (NLP)

NLP is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

D. Analysis

Stage1:

The datasets are collected form the Kaggle.com which consists of 50 Bank datasets and user review system for each bank.

Stage2:

Data Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

IJPRSE
Progressive Research
International Journal of Progressive Research in Science and Engineering
Volume-1, Issue-3, June-2020
www.ijprse.com

Missing Data: This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

1. Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
2. Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value

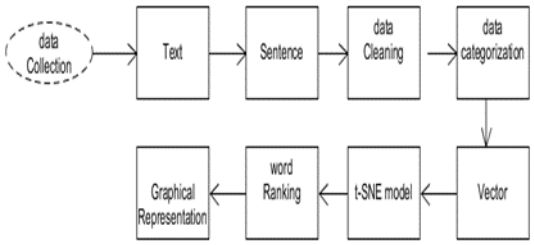

Fig.1. Flow Diagram

Stage 3:

1. Remove punctuation

Punctuation can provide grammatical context to a sentence which supports our understanding. But for our vectorizer which counts the number of words and not the context, it does not add value, so we remove all special characters. eg: How are you?->How are you

2. Tokenization

Tokenizing separates text into units such as sentences or words. It gives structure to previously unstructured text. eg: Plata o Plomo-> 'Plata','o','Plomo'.

3. Remove stop words

Stop words are common words that will likely appear in any text. They don't tell us much about our data so we remove them. eg: silver or lead is fine for me-> silver, lead, fine.

4. Stemming

Stemming helps reduce a word to its stem form. It often makes sense to treat related words in the same way. It removes suffices, like "ing", "ly", "s", etc. by a simple rule-based approach. It reduces the corpus of words but often the actual words get neglected. eg: Entitling, Entitled->Entit

5. Lemmatizing

Lemmatizing derives the canonical form ('lemma') of a word. i.e the root form. It is better than stemming as it uses a dictionary-based approach i.e a morphological analysis to the root word.eg: Entitling, Entitled->Entitle

6. Vectorising Data

Vectorising is the process of encoding text as integers i.e. numeric form to create feature vectors so that machine learning algorithms can understand our data.

Stage 4:

The obtained data is also trained using deep learning like LSTM.

Stage 5:

The trained data are classified and visualized.

III. RESULTS AND DISCUSSION

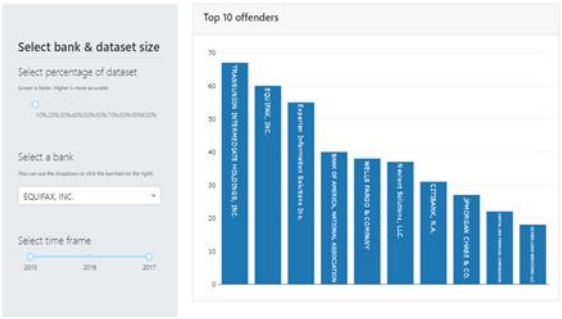

Fig.2. Snapshot showing the top 10 fraud banks and user options.

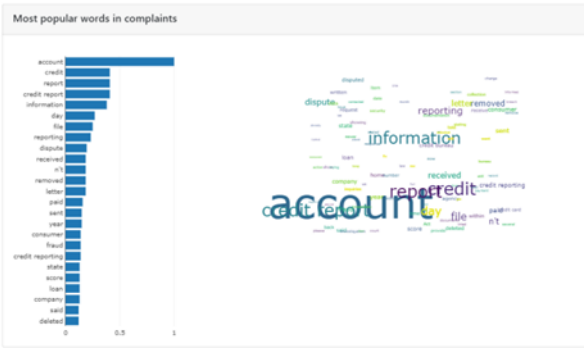

Fig.3. Snapshot showing the range of most popularly used words with Word cloud representation.

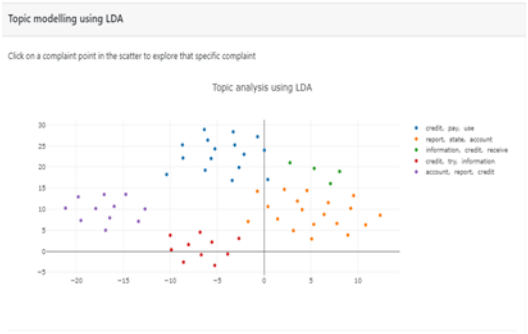

Fig.4. Snapshot showing LDA analysis of the complaints

IV. CONCLUSION

In This Project we focus on exploring and analyzing Consumer Finance Complaints data, to find how many similar complaints are there in relation to the same bank or service or product. These datasets fall under the complaints of Credit reporting, Mortgage, Debt Collection, Consumer Loan and Banking Accounting. By using Natural language processing, we analysis as well as TSNE modelling is applied to obtain valuable information about complaints in fifty 50. The banks that are receiving customer complaints filed against them will analyze the complaint data to provide results on where the most complaints are being filed, what products/ services are producing the most complaints and other useful data. Our model will assist banks in identifying the types of errors for resolution, leading to increased customer satisfaction to drive revenue and profitability and also for the customer to understand which banks have the more complaint analysis on the bank product.

REFERENCES

[1]. Lengnick-Hall, C.A. Customer contributions to quality: A different view of the customer-oriented firm. Acad. Manag. Rev. 1996, 21, 791–824.

[2]. Nambisan, S. Designing virtual customer environment for new product development: Toward a theory. Acad. Manag. Rev. 2002, 27, 392–413.

[3]. Rigby, D.; Zook, C. Open-market innovation. Harv. Bus. Rev. 2002, 80, 80–81.

[4]. Atuahene-Gima, K. An Exploratory Analysis of the impact of market orientation on new product performance: A contingency approach. J. Prod. Innov.Manag.1995, 12, 19.

[5]. Zhan, J.; Loh, H.T.; Liu, Y. Gather customer concerns from online product reviews—A text summarization approach. Expert Syst. Appl. 2009, 36, 2107–2115.

[6]. Decker, R.; Trusov, M. Estimating aggregate consumer preferences from online product reviews. Int. J. Res. Mark. 2010, 27, 293–307.

[7]. Park, Y.; Lee, S. How to design and utilize online customer centre to support new product concept generation. Expert Syst. Appl. 2011, 38, 10638–10647.

[8]. Aguwa, C.C.; Monplaisir, L.; Turgut, O. Voice of the customer: Customer satisfaction ratio based analysis. Expert Syst. Appl. 2012, 39, 10112–10119.

[9]. Wang, Y.; Tseng, M.M. A Naïve Bayes approach to map customer requirements to product variants. J. Intell. Manuf. 2015, 26, 501–509.

[10]. Aguwa, C.; Olya, M.H.; Monplaisir, L. Modeling of fuzzy-based voice of customer for business decision analytics. Knowl.-Based Syst. 2017, 125,136–145.