PROJECT TITLE

Defect classification using customer Q & A

Amazon

Bagmane Constellation B, Bangalore, India 560037

SUBMITTED BY

Ms. Amirta V

M.Sc. (Applied Statistics)

PRN:21060641004

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
STATISTICAL INSTITUTE

ACADEMIC YEAR 2022 - 23

Under the guidance of

Name of Project Guide / Mentor

Mr. Aakash Gupta

Designation: Business Research Analyst – II

Email Id: aakasgu@amazon.com

Contents

1. Executive Summary

The return of products in the retail business can pose various challenges for the retailers. While returns are an integral part of the customer experience, they can have significant implications for the profitability and customer satisfaction of retail business. This project focuses on reducing the returns by building a text classification model to classify the defects using customer Q&A. Two separate classification models are built, one for defect classification and the other one for sub-defect classification.

2. Study Background

Online shopping platforms like operate through a sophisticated system that enables customers to browse and purchase products conveniently. Customers are provided with a user-friendly website interface where customers can search for products based on categories, keywords, or specific criteria. The customers navigate to the desired product's detail page, where they find detailed product listings that include images, descriptions, specifications, pricing, and customer reviews. It is crucial for sellers and manufacturers to provide comprehensive and accurate information on the product detail page about the product. These information's help customers to make informed purchase decisions. After deciding on a product, customers can add it to their virtual shopping cart, once customers are ready to proceed with the purchase, they initiate the checkout process by completing the payment. Once the customers receive the product, if they are not satisfied with their purchase, they may have the option to return the product.

Return of products in the retail business can pose various hazards and challenges for retailers. While returns are an integral part of the customer experience, they can have significant implications for the profitability, operational efficiency, and customer satisfaction of a retail business. Finding and reducing the returns are very important.

Customers may return a product for various reasons, depending on their individual circumstances and expectations. Few common reasons for the returns are product being defective, receiving wrong product, unintentionally ordered the wrong product etc. These reasons are listed by customers when they return the product. But actual problem arises when there is no reason mentioned by the customer while returning the product. In such situation other data sources can be used for reducing the returns. One such scope is questions asked by customers.

Customers often struggle to make purchase decisions, because of incomplete information, difficulty in finding the relevant information in the detail page or inaccurate information in the detail page. These gap in the detail page leads customers to the Q&A section. As a result, the time taken by customer to make buying decision increases and it leads to increase in negative customer experience.

It is essential to minimize the need for customers to rely solely on the Q&A section for clarification and ensure a smoother online shopping experience. This is done by identifying the issues affecting the customer's buying decision and issues driving customers to return the product.

3. Literature Review

An integrated review technique is followed in this project work. Using text classification businesses can make the most out of unstructured data. Several research works are being carried on to classify the text data. Authors of paper 1 (refer to bibliography) show how Multinomial Naïve Bayes (MNB) can be used for unbalanced text classification. Furthermore, the authors of paper 2 (refer to bibliography) had applied multiple machine learning models such as Random Forest and Gaussian Naïve Bayes to different labelled data to measure and compare the accuracy of the classifiers. Authors Hanumahu et al. of paper 3 (refer to biboliography) applied XGBoost Classification to do a fake news classification for Indonesian news. These papers were used as references to proceed further in the project.

4. Aims & Objectives

To explore and process the customer Q&A data and developing a machine learning model to identify and classifying the defects from the question asked by customers.

5. Methodology

The data required to achieve the above-mentioned objectives were retrieved using SQL query from Amazon's internal database. The duration of the data used for the study is from 1 September 2022 to 1 February 2023 (5 months). There were 12 M records in the dataset.

As the objective of the study is to classifying defects from the questions, the main focus here will be the questions

Below are the steps followed in the project:

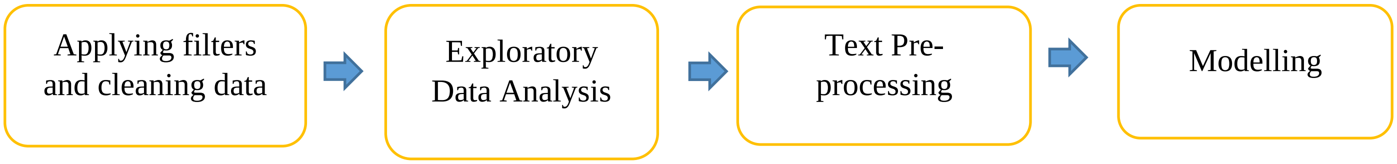

Fig 1: Methodology of the project

i. Understanding the problem statement and the data

It is the crucial part of the process. Sating the business problem in SMART way helped to understand the need and benefits of solving the problem for the users. The data used for solving the problem is text data. Reading through a lot of data helped in understanding the context of the text.

ii. Cleaning data and applying filters

a) Removing duplicate entries

The duplicate entries were removed.

b) Questions with no answers

Some of the questions doesn’t have answer which means that the respective questions have not been answered yet. Even though the answers are not present, still the questions would be considered for the study.

c) Filtering based on Language

For now, the Questions asked in English language are only considered for this study.

d) Selecting distinct questions

A question can have multiple answers which is the cause of duplicate questions. For this study distinct questions are considered for the modelling part

iii. Exploratory data analysis

EDA was performed in the data to understand the data in a holistically. The insights are discussed in the result section.

iv. Text pre- processing and modelling

Before text processing the major there were 2 major steps which were as follows:

Creating taxonomy:
As stated in the aims and objective the aim is to build a classification model for which a taxonomy has to be created which will have two levels to it first is overall defect class and the second level is called as sub- defect which gives the defect information in more granular level. The taxonomy was created as per the business requirement.

Creating training dataset:
The training data for the model will have the questions asked by customer and the defect class under which the questions are classified. Here the human intervention is needed to make the training data. The context of questions asked by two customers can be same but

the choice of words depends on individual’s vocabulary. This makes it difficult to use NLP technique such as ‘Regex’ to annotate the data or ‘Bag of Words’ model for classification.

To create the training data, a random sample was sampled and manual annotation was carried out. The questions were carefully studied to understand its context in order to label them under appropriate class and sub- class. The problem of imbalance in data was taken care while creating the training data.

Text Pre-processing:

Following are text pre- processing done on the data:

- Converting into lower cases: All the text data were converted into lower cases. Model is case sensitive. So, all text were converted into lower case.
- Expanding Contractions: Contractions such as ‘ Shouldn’t ’ has to be expanded before removing the punctuations. If punctuations are removed these words will never make sense.
- Removing URL: The links which are provided in the questions doesn’t add any information to the model. It has to be removed.
- Removing Numbers, punctuations and emoji’s : The numbers, punctuations and emoji’s present in the text do not significantly contribute for the meaning of the text. Removing them simplifies the data and reduces the corpus size.
- Removing stop words: Stop words are common words that often do not carry much meaning in the context of a specific task, such as "and," "the," or "is." Removing these words helps reduce noise and improve computational efficiency by focusing on more informative words.
- Lemmatization: Lemmatization converts words to their dictionary or base form. In this stemming was not performed only lemmatization was carried out. These techniques help in standardizing words and reducing vocabulary size. Two tools Spacy and NLTK were used for lemmatization. Among these two Spacy performed better in lemmatizing the data.
- Pos- tagging: For the given data pos- tagging is not performed. The reason is discussed in detail in discussion and conclusion section.

v. Modelling

Classification models are built for the text data in solving the business problem. Two separate models are built. One is for defect classification and the other for sub- defect classification. Below are the classification models proposed for using and reasons for choosing them.

Model	**Approach**	**Reason**
Naïve Bayes	Machine Learning	The input data is a text data and the output is a multiclass classification. The training data entries for each class are less in number. Using Naïve Bayes in this situation will provide a good result.
Decision Tree	Machine Learning	The input text data is a single feature and is complex. To classify such data with high accuracy the decision tree is fitted. The tree is grown until pure node is reached ensuring high accuracy.
Random Forest	Machine Learning	Only one decision tree is grown. The data is huge, growing decision tree can result in overfitting. Whereas random forest is an ensemble technique, which combines results from multiple built trees will precisely classify the questions and also will ensure that the model does not over fit.
XGBoost Classification	Machine Learning	Few classes have less data entries leading to imbalance of data which affects the model fitting. XGBoost classification model boosts such classes. The model is scalable and time effective as data size increases.

6. Results

The project aims to

- EDA vitally helps in initial understanding of data. The insights from the EDA are presented below:

- There were 65.4 % that is 77 M duplicate data entries were present and were removed.
- 98.51 %of questions are asked in English language.
- 2.1 M distinct questions were asked.
- Most of the questions are asked before purchasing the product which are called as pre-purchase questions.
- There were only few questions which are asked after purchasing the product which are called as post- purchase question.
- The top 10 type of product arranged based on the number of distinct questions asked were found.
- In this the major focus will only be on question. Questions will be used in the modelling part. The questions will have useful information for finding the issues related to customer's pre- purchase experience.

- In the taxonomy created the defect class, has 10 class and the second level sub- class has been 56 sub- class.

- While creating training data each class had 150-180 observations and each sub- class had 35- 50 observation. The total size of training data for defect classification model is 1

8. Acknowledgement

I express my sincere gratitude to **Amazon** for providing me this chance to get the industry exposure, my manager **Mr. Nilanjan Dutt**, my mentor **Mr. Aakash** and my onboarding buddy **Mr. Mohammed Muzwar** for their time and kind support throughout the project. I express my sincere thanks to my institute **Symbiosis Statistical Institute, Pune** for their encouragement and support to explore summer internship opportunities.

9. References

1. Frank, E., & Bouckaert, R. R. (2006). Naive bayes for text classification with unbalanced classes. In *Knowledge Discovery in Databases: PKDD 2006: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases Berlin, Germany, September 18-22, 2006 Proceedings 10* (pp. 503-510). Springer Berlin Heidelberg.
2. Kumar, R. R., Reddy, M. B., & Praveen, P. (2019). Text classification performance analysis on machine learning. *International Journal of Advanced Science and Technology*, *28*(20), 691-697.
3. Haumahu, J. P., Permana, S. D. H., & Yaddarabullah, Y. (2021, March). Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost). In *IOP Conference Series: Materials Science and Engineering* (Vol. 1098, No. 5, p. 052081). IOP Publishing.