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

SYMBIOSIS
STATISTICAL INSTITUTE

॥वसुधैव कुटुम्बकम्॥

ACADEMIC YEAR 2022 - 23

Under the guidance of

Name of Project Guide / Mentor

Mr. Aakash Gupta

Designation: Business Research Analyst – II

amazon | Development Centre India

INTERNSHIP LETTER

Amirta V
No. 18 Kannagi street, Srinivasa nagar, Near perumal koil, New perungalathur
Chennai – 600063
TN
IN

Dear Amirta,

On behalf of **Amazon Development Centre (India) Private Limited**, a company incorporated under the laws of India, having its registered office at # 26/1, Brigade Gateway, World Trade Centre, 10th Floor, Dr. Rajkumar Road, Malleshwaram (W) Bangalore - 560 055. Karnataka India (hereinafter the "Company" or "Amazon India"), we are very pleased to issue this Internship Letter for the position of an **Intern** at **Bangalore**, India.

Your internship with the Company will be subject to your acceptance of this Internship Letter and the terms and conditions set forth hereinbelow on or before 10 business days in the manner provided for by the Company.

Upon your acceptance of this Internship Letter, the same shall form a valid and binding agreement between Amazon India and you, and you shall be bound by the terms and conditions stipulated herein below.

1. Date of Commencement

Your internship with Amazon India will commence on **23-Jan-2023** and shall end as per the provisions contained in Section 12 herein below. The said duration of internship shall hereinafter be referred to as the "Term".

2. Duties

2.1 You will be engaged in the position of **Business Research Analyst**. Your manager will advise you about your duties and responsibilities after your joining with us. You will be expected to perform your duties to the best of your ability at all times as per the responsibilities advised by your manager at the time of joining or as amended from time to time, as well as such other tasks as may be required by Amazon India.

2.2 You will be required to comply with Amazon India's rules, regulations and policies from time to time in force, including, without limitation, those policies set out in Amazon India's Policies and Procedures (as may be applicable to you), as communicated to you. Amazon India reserves the right to change Amazon India's Policies and Procedures from time to time at its sole discretion and you shall be bound by the same, so far as may be applicable to you.

2.3 You acknowledge that during the Term, as the business of Amazon India changes, it may be necessary to rotate you in other departments / units. Amazon India therefore reserves the right to change your role and responsibilities from time to time at its sole discretion and without assigning any reason, it being understood that you will not be assigned responsibilities which you cannot reasonably perform.

2.4 Unless specified in writing, you shall not be authorised to enter into any contractual obligations on behalf of Amazon India or its affiliates including creating a lien (statutory or other), security interest, mortgage, pledge, assignment, encumbrance, chattel or conditional sale or other title retention agreement or any other financial obligations or otherwise on behalf of Amazon India or its affiliates.

3. Hours of Work

The normal business hours of the office, at which you work, will apply to you and these will be advised on commencement of internship and when there is a change. You may be required to work in shifts for different workhours or workdays during the week depending on the business or team that you may be working for. You will be advised by your manager or department about such requirements at the time of joining and from time to time during the course of your internship, as appropriate. Certain business teams also operate on 24x7 basis and hence, may have rotational shifts or related requirements for their respective team members. Please refer to Amazon India's Policies and Procedures for further details.

4. Place of Work

Your initial place of work will be at Amazon India's facility in Bangalore. However, you should be aware that the Company and/or its affiliates have offices throughout the world and because of the nature of your duties, the Company has the right to transfer you from one place to another or from one section to another or from one unit to any other unit of the Company, its parent company or to any of its sister concerns, which are either existing or may be set up in future. The decision of the Company in this regard shall be final and binding on you. As you are joining during the period of the Covid-19 Pandemic, you may be permitted to work from a location of your choice in India with the prior approval of your manager under the condition that you are willing to get back to the location mentioned above as and when required by Amazon.

5. Remuneration

5.1 Your internship stipend will be Rs.**70,000** per month made payable in arrears and subject to all lawful deductions of tax.

5.2 Amazon India has the right to deduct from your stipend any sums which you may owe Amazon India, including without limitation, any over-payments or loans made to you by Amazon India or any demand raised by any judicial or quasi-judicial authority for your acts or omissions and / or losses suffered by Amazon India as a result of your negligence or breach of the terms contained in this Internship Letter/Amazon India’s Policies (as may be applicable to you), or your failure to return Amazon India’s property.

5.3 You will be reimbursed for any reasonable expenses incurred by you in the course of the performance of your internship on behalf of Amazon India, subject to your compliance with the Expenses Policy contained in Amazon India’s Policies and Procedures (as may be applicable to you).

6. Leave and Benefits

You will not be entitled to any leaves or such other employee benefits during the term of your internship with Amazon India.

7. Confidential Information and Confidentiality Obligations

7.1 “Confidential Information” means and includes any information that relates to the business of the Company that is not generally available to the public. Without limiting the foregoing, Confidential Information includes:

(1) the identity of, contractual terms with, and any information relating to, the Company's business partners, customers, services clients, sellers, agents, employees, contractors, investors, joint venturers, vendors, or suppliers and the terms on which the Company does business with each such entity, or generally;

(2) computer code (including source code and object code) or software developed, modified, or used by the Company;

(3) data of any sort compiled by the Company, including, but not limited to, data relating to products and services, advertising and marketing, and existing or prospective customers, clients, vendors, or business partners;

(4) algorithms, procedures or techniques, or the essential ideas and principles underlying such algorithms, procedures or techniques, developed by, or whose workings are otherwise known to, the Company (but excluding any public domain algorithms, procedures, or techniques), whether or not such algorithms, procedures or techniques are embodied in a computer program, including, but not limited to, techniques for identifying prospective customers, communicating effectively with prospective or current customers, reducing operating costs, or increasing system reliability;

(5) the fact that the Company uses, has used, or has evaluated for potential use any particular database, source of data, algorithm, procedure or technique, or the essential ideas and principles underlying such algorithm, procedure or technique, developed or supplied by a party other than the Company (including any algorithms, procedures or techniques in the public domain), whether or not such algorithms, procedures or techniques are embodied in a computer program;

(6) pricing or marketing strategies developed, investigated, acquired (from a third party or otherwise), evaluated, modified, tested or employed by the Company, or any information related to, or that might reasonably be expected to lead to, the development of such strategies;

(7) information about the Company's future plans, including, but not limited to, plans for expanding into new products, geographical areas, market segments, or services;

(8) any information that would typically be included in the Company's financial statements, including, but not limited to, the amount of the Company's assets, liabilities, net worth, revenues, expenses, or net income;

(9) the following information which shall hereinafter be referred to as the "Disclosure Information":

(a) any and all algorithms, procedures or techniques related to the Company's business activities or to your work with the Company, and the essential ideas and principles underlying such algorithms, procedures or techniques, conceived, originated, adapted, discovered, developed, acquired by the Company (from a third party or otherwise), evaluated, tested, or applied by you during the course of your internship with the Company, whether or not such algorithms, procedures or techniques are embodied in a computer program;

(b) any and all pricing or marketing strategies, the essential ideas and principles on which such strategies are based, and any information that might reasonably be expected to lead to the development of such strategies, conceived, originated, adapted, discovered, developed, acquired by the Company (from a third party or otherwise), evaluated, tested, or applied by you during the course of your internship with the Company;

(c) information relating to any and all products and services, and the essential ideas and principles underlying any and all products and services, conceived, originated, adapted, discovered, developed, acquired by the Company (from a third party or otherwise), evaluated, tested, or applied by you during the course of your internship with the Company, whether or not such products or services are marketed, sold, or provided by the Company; and

(d) any other ideas or information conceived, originated, adapted, discovered, developed, acquired by the Company (from a third party or otherwise), evaluated, tested, or applied by you during the course of your internship with the Company, if the idea or information could reasonably be expected to prove useful or valuable to the Company;

(10) any other information gained in the course of your internship with the Company that could reasonably be expected to prove deleterious to the Company if disclosed to third parties, including without limitation, any information that could reasonably be expected to aid a competitor or potential competitor of the Company in competing more effectively with the Company;

(11) any information received by the Company from third parties, whether or not under obligation of confidentiality;

(12) any information derived from any of the above, including any intellectual property rights attached thereto; and

(13) any copies of the above mentioned information.

7.2 Confidentiality Obligations:

(i) You acknowledge that you have acquired and/or will acquire Confidential Information during the course of, or incident to, your internship with the Company, and that the ability of the Company to continue in business could be seriously jeopardized if such Confidential Information were to be used by you or by other persons or firms to compete with the Company. Accordingly, you agree that you shall not, directly or indirectly, at any time, during the term of your internship with the Company or at any time thereafter, and without regard to when or for what reason, if any, such internship shall terminate, use or cause to be used any Confidential Information in connection with any activity or business except the business of the Company, and shall not disclose or cause to be disclosed any Confidential Information to any individual, partnership, corporation, or other entity unless such disclosure has been specifically authorized in writing by the Company, or except as may be required by any applicable law or by order of a court of competent jurisdiction, or any regulatory or governmental body. Further, you agree that you will give the Company prompt notice of any such order/direction of a court/ regulatory or governmental body so that the Company may seek relief by way of a protective order or other appropriate remedy, and further will provide any assistance which the Company may reasonably require in order to secure such order or such remedy (with your expenses reasonably incurred in providing such assistance to be reimbursed by the Company). In the event such protective order or other remedy is not obtained, you shall furnish only that portion of the Confidential Information which is legally required by the governmental entity or regulatory authority; and will use reasonable efforts to obtain confidential treatment for any Confidential Information so disclosed.

(i) During the course of your internship with the Company and at the date of termination thereof (hereinafter the "Date of Termination"), you shall promptly disclose and deliver over to the Company, without additional compensation, in writing, or in such form and manner as the Company may reasonably require, the Disclosure Information defined in Section 7.1(9) hereinabove, to the extent that such disclosure could reasonably be expected to be of interest to the Company.

(i) Nothing in this Internship Letter shall be deemed to dilute or waive any rights related to the protection of trade secrets that the Company may have under common law or any applicable statutes.

8. Intellectual Property Rights

8.1 All patents, copyrights, trade secrets, trade/commercial names, proprietary rights, logos, slogans and all other intellectual property rights developed by or for the Company by any person, including but not limited to intellectual property rights relating to any and/or all of the Confidential Information, ("Intellectual Property Rights") shall be owned by the Company.

8.2 For good and valuable consideration, the receipt and sufficiency of which is hereby acknowledged, you hereby agree to irrevocably, perpetually and unconditionally sell, assign, transfer and convey to the Company and its successors your entire right, title and interest in the Confidential Information and/or Intellectual Property Rights and any improvements thereto throughout the world, including, without limitation:

(i) all patents, copyrights, trade secrets, trade/commercial names, logos, other proprietary rights and all other intellectual property rights in the Confidential Information and all rights to secure registrations, renewals and extensions of the same;

(i) all rights to make, have made, use, practice, import, export and otherwise fully exploit the Confidential Information and any and all improvements that you or Company may hereafter make or develop;

(i) all rights to file and prosecute applications for patent, copyright and all other intellectual property protection covering the Confidential Information and improvements thereon, and the processes and designs embodied therein, in India, the United States and in every other country and jurisdiction throughout the world;

(i) all rights under any patent, copyright and all other intellectual property which may be issued on the Confidential Information or the improvements thereon, and any processes and designs therein, and all rights to enjoy the same; and

(i) all documents, notes, notebooks, drawings, schematics, prototypes, magnetically encoded media, electronically stored information, or other materials related to the Confidential Information.

8.3 During the period of your internship with the Company and as may be reasonably necessary subsequent to your internship, you agree to cooperate with the Company as may be necessary to obtain patent, copyright and all other intellectual property protection for the Intellectual Property Rights and improvements thereto throughout the world and agree to do such further acts and execute and deliver to the Company such instruments as may be required to perfect, register or enforce the Company's ownership of the rights assigned, transferred or conveyed. If such cooperation is required after the Date of Termination, the Company shall compensate you at a reasonable rate for the time and related expenses actually spent by you at the Company's request. If you fail or refuse to execute any such instruments, you hereby appoint the Company as your attorney-in-fact to act on your behalf and to execute such instruments. This appointment shall be irrevocable and deemed to be a power coupled with an interest.

8.4 For the purposes of the assignment, transfer or conveyance referred to hereinabove, you acknowledge and covenant that your internship with the Company and the benefits received thereunder shall be treated as good and valuable consideration and that you are not entitled to any further consideration in any form or manner whatsoever in relation thereto.

8.5 Notwithstanding any other provision hereof to the contrary, this Internship Letter does not obligate you to assign or offer to assign to the Company any of your rights in an invention for which no equipment, supplies, facilities, Intellectual Property Rights, Confidential Information or trade secret information of the Company was used and which was developed entirely on your own time, unless (a) the invention relates (i) directly to the business of the Company, or (ii) to the Company's actual or demonstrably anticipated research or development, or (b) the invention results from or is related to, any work performed by you for the Company.

8.6 No Grant of Rights.

You agree that all rights, title and interest in the Intellectual Property Rights and Confidential Information shall be owned exclusively by the Company. Nothing herein contained shall be construed as a grant by implication, estoppel or otherwise, of a license of any kind by either you to the Company, or by the Company to you, for example, to make, have made, use or sell any product using the Intellectual Property Rights, Confidential Information, or as a license under any patent, patent application, utility model, copyright, mask work right, or any other intellectual property right.

9. Data Protection

9.1 You authorise Amazon India to collect, process and transfer all your personal information obtained by Amazon India for the purpose of proactively managing the relationship.

9.2 You further authorise the transfer to, and storage of, your personal information in the worldwide database currently located in Seattle, Washington, U.S.A. (or such other location as Amazon India determines from time to time). Human Resources and selected management throughout the Amazon group worldwide will be authorised to access this database.

10. Exclusivity

During your internship, you will be required to devote your full time, attention and abilities to your assignment, and to act in the best interests of Amazon India at all times. You shall not, without the written consent of Amazon India, be in any way directly or indirectly engaged or concerned in any other business or undertaking or undertake any internship therein.

11. Relationship of parties

This internship opportunity neither creates the relationship of employer and employee between the Company and you, nor does it assure or guarantee future employment with the Company.

12. Termination of Internship

12.1 Your internship will automatically end on **16-Jun-2023**, unless terminated earlier as per the provisions of this Section.

12.2 This Internship Letter may be terminated either by the Company or by you at any point of time during the Term, without providing any reasons for such termination. Such termination shall be valid and effective only if communicated to the other party in writing at least one day prior to the date of termination.

12.3 On the expiry or sooner termination of your internship for any reason whatsoever, you will return to Amazon India, without delay, all assets belonging to Amazon India, correspondence, records, specifications, models, notes, formulations, lists, papers, reports and other documents and all copies thereof and other property belonging to Amazon India or relating to its business affairs or dealing which are in your possession or under your control. At Amazon India’s option, you agree to provide a written certification of your compliance with this Section. Further, you agree to sign a termination certificate in accordance with Amazon India’s Policies and Procedures, which will reaffirm your compliance of your post-termination obligations, including return of Amazon India’s property/properties and releasing Amazon India from all claims, liabilities and obligations. Where Amazon has made any excess payment to you as part of your relieving formalities, whether or not such excess payment is termed “Full and Final Settlement”, you shall be obligated and liable to repay such excess amount forthwith upon being notified by Amazon.

13. Background Investigation

13.1 It is Amazon India’s policy to investigate all its new interns. Your internship is conditional upon the information contained in your application form and/or curriculum vitae being true and accurate, including (but not limited to) your educational and professional qualifications, the documents furnished by you being genuine, and upon reference checks to be conducted by Amazon India being successfully completed.

13.2 You authorise Amazon India to conduct such searches with government or enforcement authorities as are necessary to enable it to verify that you do not hold any criminal convictions.

14. Foreign Nationals

14.1 In case you are not an Indian national and, under any law, are required to obtain applicable visa / work permit / authorisation or permission from appropriate government authorities to work in India, you are required to ensure all such permissions are obtained before commencement of internship with Amazon India.

14.2 You are also required to ensure all future correspondence and permissions for continued stay and internship in the country as per the governing law are complied with at all times. If required, Amazon shall be at liberty to demand copies / originals of such permission.

14.3 It is made clear that possessing valid work permit / authorisation at all times of your internship is an inherent requirement of your internship with Amazon India. Any time after the execution of this Internship Letter, if it is found that you do not have required work permit / visa, Amazon India shall terminate your internship, without notice, with immediate effect, without any liability towards you.

15. Representations and Warranties

You hereby represent and warrant to the Company that:

15.1 you shall not, during the course of your internship with the Company, use or disclose any document/s that in any way constitutes confidential, proprietary of trade secret information of a third party, except pursuant to written authorization by such third party to do so;

15.2 you are not in unauthorized possession or control of any document/s that in any way constitutes confidential, proprietary of trade secret information of a third party;

15.3 You confirm that there are no other agreements executed by you with third parties that conflict with the terms and conditions of your internship with Amazon India or that restrict your ability to execute this Internship Letter.

15.4 You hereby represent and warrant that the information furnished by you for the purpose of your internship with the Company is true and correct to the best of your information, knowledge and belief.

16. Notices

All notices issued by you to the Company or by the Company to you shall be sent either by registered post, courier through a recognised courier service provider or email transmission which shall be deemed to have been received the next working day provided the notice is also sent by registered post the next working day after email transmission.

17. Waiver

Failure of the Company to insist upon strict adherence of any term of this Internship Letter on any occasion/s shall not be considered a waiver thereof or deprive the Company of the right thereafter to insist upon strict adherence to that term or any other term of this Internship Letter.

18. Severability

The holding of any provision of this Internship Letter to be illegal, invalid, or unenforceable by a court of competent jurisdiction shall not affect any other provision hereof, which shall remain in full force and effect.

19. Liability for Breach

You acknowledge and accept that your breach of any of the terms contained in this Internship Letter and/or Amazon India’s Policies and Procedures (as may be applicable to you) may cause the Company irreparable harm for which there is no adequate remedy at law, and therefore, the Company shall be entitled to the issuance by a court of competent jurisdiction of an order of injunction, restraining order, or other equitable relief in favor of itself, without the necessity of posting a bond, restraining you from committing or continuing to commit any such violation. Exercise or waiver by the Company of its rights to obtain an injunction, restraining order, or other equitable relief hereunder shall not be deemed a waiver of any right to assert any other remedy the Company may have at law or in equity. In any legal action or other proceeding by the Company against you in connection with this Internship Letter (e.g., for recovery of damages or other relief), the Company will be entitled to recover its reasonable attorneys' fees and other costs incurred.

20. Governing Law and Jurisdiction

Your internship, and any disputes which may arise under, out of, or in connection with your internship, shall be governed by and construed in accordance with the laws of India; and the Courts having territorial jurisdiction over the registered office of the Company shall alone have exclusive jurisdiction to try and entertain such disputes to the exclusion of any other Courts situated elsewhere.

21. Agreement/Modifications

The terms described in this Internship Letter and in Amazon's Policies and Procedures (as may be applicable to you), will cumulatively constitute the terms of your internship, and shall supersede any previous discussions, offers, or agreements relating to your internship, or the subject matter hereof. Any additions to, deletions of, or modifications of these terms are valid and effective only if the same are carried out in writing and signed by you and an officer of Amazon India.

22. Headings

The Section headings appearing in this Internship Letter are used for convenience of reference only and shall not be considered a part of this Internship Letter or in any way modify, amend or affect the meaning of any of its provisions.

23. Survival

Your obligations under Sections 7, 12, 17, 18, 19, 20 and this Section 23 hereof shall survive the termination of this Internship Letter and of your internship with the Company.

You undertake to be bound by any rules and regulations enforced by Amazon India from time to time in relation to the conduct, discipline, medical leave and holidays or on any matters relating to service conditions which will be deemed as rules, regulations and order as a part of these terms of internship.

For and on behalf of Amazon Development Centre (India) Private Limited

AUTHORIZATION

By

Signed by:ZUBAIR CHISHTI
Date: 2022.12.06 12:26:32 +05:30
Location: India

ACCEPTANCE

I acknowledge receipt of this Internship Letter and, after reading and understanding the same, I accept the same on the terms set out herein.

amazon | Development Centre India

17th Jul 2023

TO WHOMSOEVER IT MAY CONCERN

Sub: Internship of Amirta V at Amazon India ,Bengaluru,India

This is to certify that **Amirta V** a student of **Symbiosis Statistical Institute** has completed her internship with **Amazon Development Centre India Pvt. Ltd.** from **23-Jan-2023** to **16-Jun-2023**

The Project was titled **DEFECT CLASSIFICATION USING CUSTOMER Q&A FOR US- HARDLINES** and assigned by **RBS Defect Reduction Variable** team under the guidance of **Nilanjan Dutta.** She has successfully completed all the requirements of the project.

We wish her the very best in her future endeavors.

Signature Not Verified

Digitally signed by DS AMAZON DEVELOPMENT CENTRE (INDIA)
PRIVATE LIMITED 6
Date: 2023.07.17 11:09:52 IST
Location: Bengaluru

Regards,

For Amazon Development Centre India Pvt. Ltd.

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

SYMBIOSIS
STATISTICAL INSTITUTE

॥वसुधैव कुटुम्बकम्॥

ACADEMIC YEAR 2022 - 23

Under the guidance of

Name of Project Guide / Mentor

Mr. Aakash Gupta

Designation: Business Research Analyst – II

Contents

1. **Executive Summary** ……………………………………………………………...03

2. **Study Background** ……………………………………………………………..04

3. **Literature Review (If any)** …………………………………………………….05

4. **Aims & Objectives** ……………………………………………………………..05

5. **Methodology** ………………………………………………………………………06

6. **Results** ……………………………………………………………………………..11

7. **Discussion and Conclusion** …………………………………………………….12

8. **Recommendation** ………………………………………………………………….13

9. **Acknowledgement** ……………………………………………………………...13

10. **References (including Bibliography and web references)** …………………………...14

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

In this project, an integrated review technique was adopted to explore various aspects of text classification and its application in leveraging unstructured data for business purposes. The literature review focused on multiple research works conducted in the field of text classification.

One of the reviewed papers demonstrated that the predictive performance of Multinomial Naive Bayes (MNB) can be enhanced by applying appropriate data transformations. Specifically, after vectorization, the authors normalized the length of the feature vectors. This paper served as a valuable resource for understanding how MNB can be effectively used for multiclass text classification.

In another paper, the authors explored the application of multiple machines learning models, including logistic regression, support vector machine, Gaussian Naive Bayes, and Random Forest, for text classification tasks on different types of labelled documents. They measured the accuracy of these classifiers and concluded that the performance of the classifiers is influenced by the training corpus to some extent. They emphasized the importance of high-quality training data corpora in deriving classifiers with good performance. However, it was noted that the paper did not provide visualizations of accuracy measures, limiting the ability to comprehend the differences between the classifiers.

Furthermore, a paper focused on the classification of Indonesian news as hoaxes or valid news using XGBoost classification. This paper was referenced to understand the application of XGBoost in classifying text data in the project.

Overall, these reviewed papers contributed to the project by providing insights into the application of different algorithms, such as Multinomial Naive Bayes, logistic regression, support vector machine, Gaussian Naive Bayes, Random Forest, and XGBoost, for text classification tasks. The literature review emphasized the significance of appropriate data transformations, the influence of training corpora on classifier performance, and the usage of XGBoost for text classification. However, it also highlighted the need for visualization of accuracy measures in order to gain a better understanding of the classifier performance.

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:

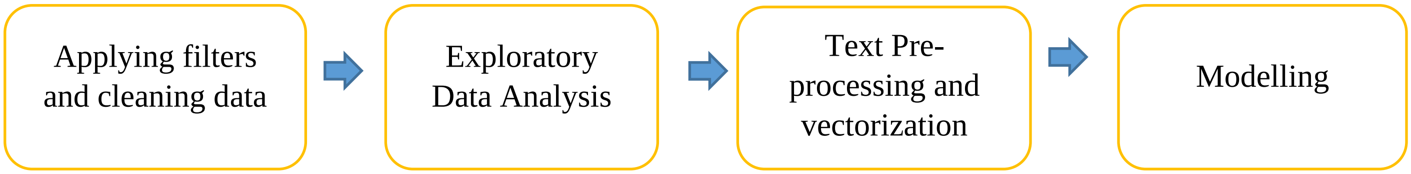
Applying filters and cleaning data
Exploratory Data Analysis
Text Pre-processing and vectorization
Modelling

Fig 1: Methodology of the project

i. Understanding the problem statement and the data

It is the crucial part of the process. Stating 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

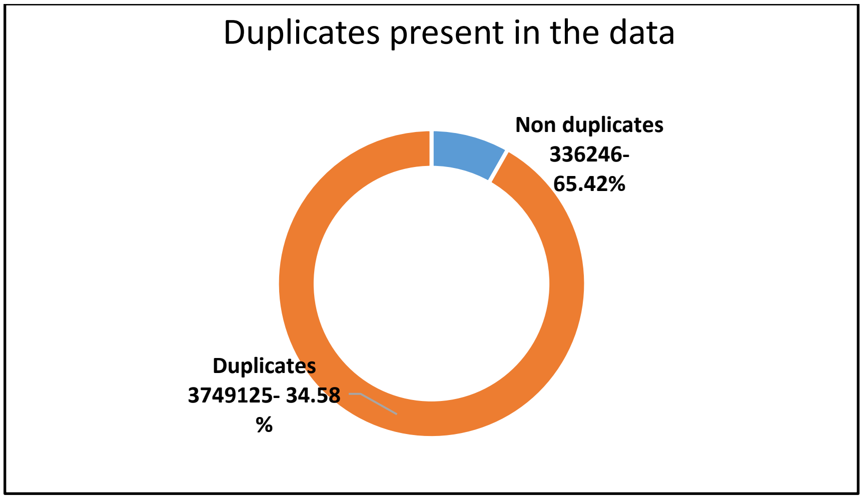

Fig 2: Duplicates in the data

The duplicate entries were removed.

b) Removing deleted Q&A from DP

The Q&A deleted from the DP cannot be included in the study, as it has already been already been deleted. Analyzing such records will not be significantly helpful in achieving the objective.

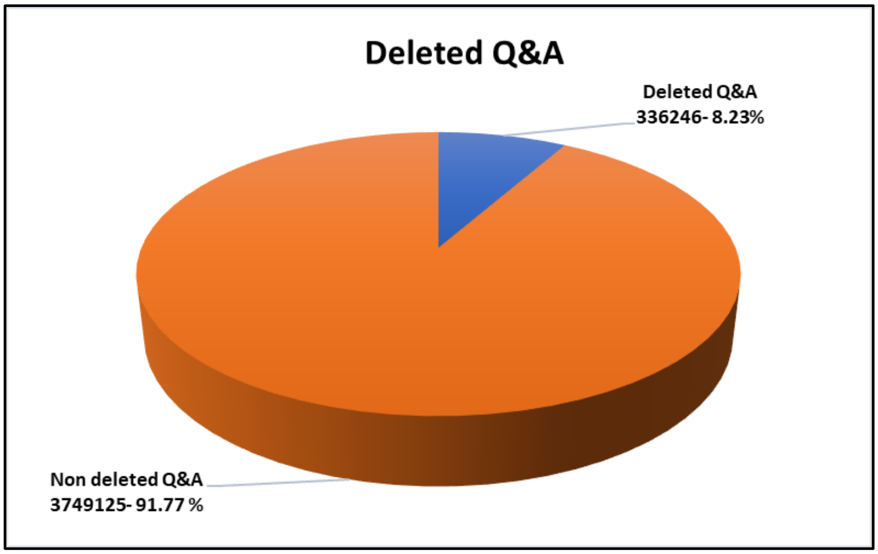

Fig 3: Deleted Q&A

c) 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.

d) Filtering based on Language

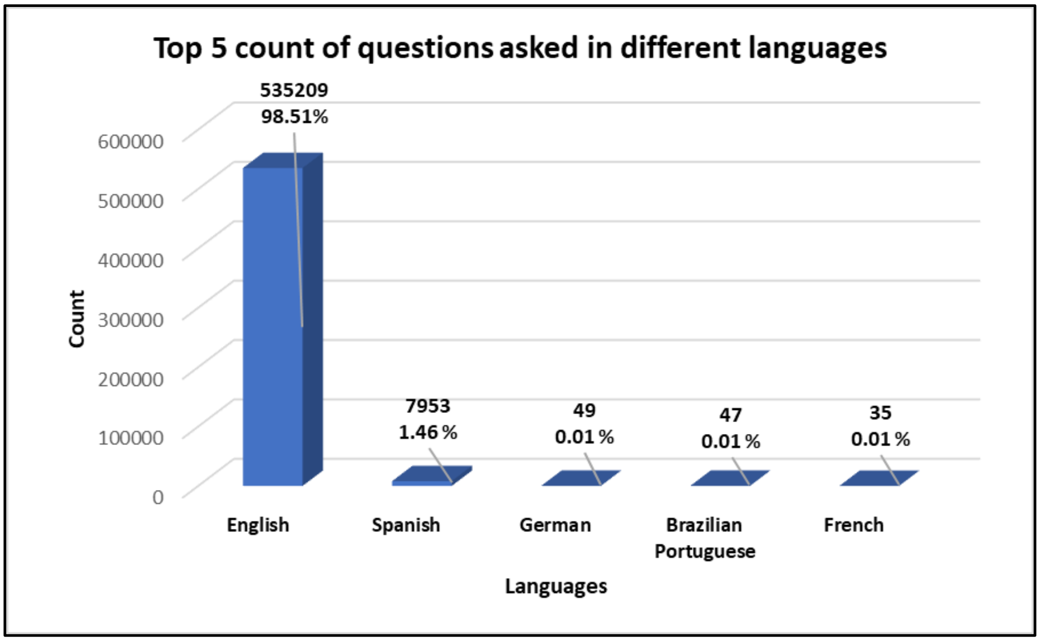

Fig 4: Count of questions asked in different languages

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

e) Selecting distinct questions

A question can have multiple answers which is the cause of duplicate questions. For this study distinct 2.1 M questions are considered for the modelling part.

iii. Exploratory data analysis

EDA was performed in the data to understand the data 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.

Text Vectorization:

For the project TFIDF vectorization method was used for converting the text data into vectors. Other methods such as Word2Vec and Fastext.

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.

vi. Accuracy Measure

F1 score is used for measuring the accuracy of the models. F1 score is the harmonic mean of precision and recall. As the data has high imbalance the recall has to be considered along with the precision. These requirements are satisfied by F1 score.

6. Results

The results are presented below:

- ➢ 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 300-325 observations and each sub- class had 35- 50 observation.

- ➢ The F1 score of all the 4 models mentioned above is represented in the below graph:

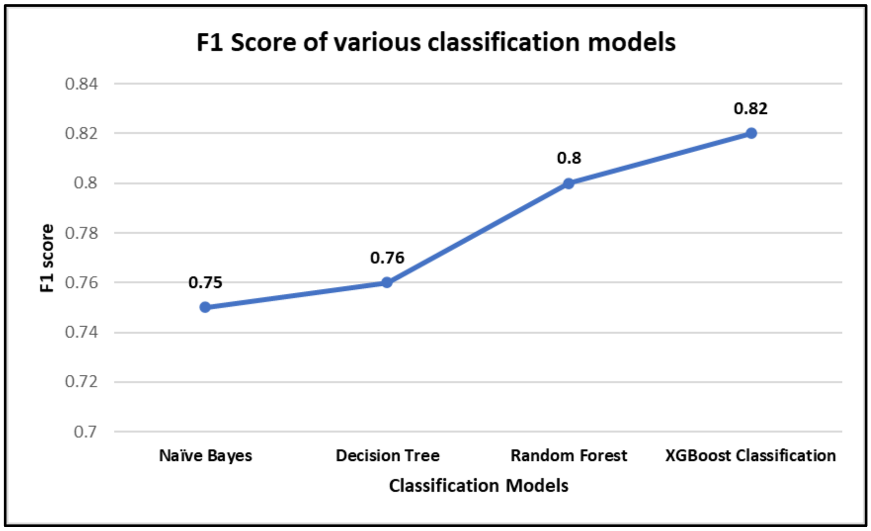

Fig 5: F1 score for various models

7. Discussion and Conclusion

In conclusion, when dealing with unlabeled data in a text classification problem, annotation plays a crucial role. The process of annotating the data provides the labeled examples needed to train a classification model effectively. The quality and accuracy of the annotations are paramount as they directly impact the performance of the classification model.

The annotated data serves a dual purpose. Firstly, it is used to build the initial classification model, providing the necessary training examples for the model to learn from. Secondly, the annotated data can be further utilized to increase the data size, which can help enhance the model's performance and generalization capabilities.

Based on the results presented, it is evident that among the four models evaluated, the XGBoost classification model outperformed the others for the given data. The XGBoost algorithm's ability to handle imbalanced data played a significant role in its superior performance. By addressing the imbalance present in the data, the XGBoost model effectively learned patterns and made accurate predictions.

Overall, the success of a text classification model heavily relies on the quality of annotations, which enable the development of an accurate and robust model. Additionally, selecting appropriate algorithms and techniques that can handle specific challenges in the data, such as imbalances, is crucial in achieving optimal performance.

8. Recommendation

In addition to the points mentioned earlier, it is important to consider potential limitations or challenges that may arise when using the XGBoost classification model. One such limitation is the possibility of overestimation in the model's performance, resulting in a higher F1 score.

This overestimation can occur if the training data provided to the model does not cover all possible scenarios or situations that may be encountered in real-world predictions. In such cases, when the model encounters an unknown scenario during prediction, it may struggle to accurately predict the class or provide reliable results.

To mitigate this issue, it is crucial to focus not only on increasing the quantity of data but also on improving the quality of the data. Ensuring that the training data is diverse, representative, and covers a wide range of possible scenarios can help improve the model's ability to generalize to unseen data.

Furthermore, considering the use of deep learning models such as BERT (Bidirectional Encoder Representations from Transformers), RNN (Recurrent Neural Network), lightGBM, or other advanced techniques can be beneficial for addressing complex text classification problems. These models often exhibit strong performance in various NLP tasks and can capture intricate patterns and relationships in the data.

In conclusion, while increasing the quantity of data is important, it is equally important to ensure the quality of the data and consider alternative models, such as deep learning models, to tackle challenges in text classification effectively. By taking these steps, we can enhance the accuracy and reliability of the classification model for unknown scenarios.

9. 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.

10. 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.

Turnitin Originality Report

Processed on: 21-Jun-2023 11:03 IST
ID: 2120149206
Word Count: 2225
Submitted: 1

checking By Amirta V

Similarity Index	Similarity by Source	
3%	Internet Sources:	2%
	Publications:	2%
	Student Papers:	0%

1% match (Internet from 30-May-2023)
https://www.irjmets.com/uploadedfiles/paper/issue_5_may_2023/40297/final/fin_irjmets1685298602.pd

1% match (Internet from 15-Jan-2023)
https://www.medrxiv.org/highwire/citation/156800/endnote-tagged

1% match (Internet from 31-Jan-2023)
https://www.researchgate.net/publication/332213958_Non-Contrast-Enhancing_Tumor_A_New_Frontier_in_Glioblastoma_Research

< 1% match (simran Garg, Devang Chaturvedi, Tanya Jain, Anju Mishra, Anjali Kapoor. "Sentiment Analysis of Twitter Data using Machine Learning: A Case Study of SVM Algorithm", Research Square Platform LLC, 2023)
simran Garg, Devang Chaturvedi, Tanya Jain, Anju Mishra, Anjali Kapoor. "Sentiment Analysis of Twitter Data using Machine Learning: A Case Study of SVM Algorithm", Research Square Platform LLC, 2023

< 1% match (Internet from 22-Sep-2020)
https://sersc.org/journals/index.php/IJAST/article/view/2900

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 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 ……………………………………………………………..03 2. Study Background ………………………………………………………………..04 3. Literature Review (If any) ……………………………………………………05 4. Aims & Objectives ……………………………………………………………….05 5. Methodology …………………………………………………………………...05 6. Results ……………………………………………………………………………….09 7. Discussion and Conclusion ……………………………………………………. 8. Recommendation ………………………………………………………………………. ……………………………………………………………………………………………… 9. Acknowledgement …………………………………………………………………... ……………………………………………………………………………………………… 10. References (including Bibliography and web references) ……………………………... ……………………………………………………………………………………………… 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,