1. **Identifying the problem**

***Descriptive Analysis***

Amazon reviews are often the most publicly visible reviews of consumer products. We know that Amazon Product Reviews Matter to Merchants [1] because those reviews have a tremendous impact on how we make purchase decisions. Studies have shown that there is a strong relationship between reviews and sale conversation rates. Reviews and ratings also help increase discoverability as potential buyers tend to filter their options based on ratings. Hence, customer feedback in form of reviews and ratings has become an important source for businesses to improve upon.

The problem today is that for many businesses, it can be tough to analyze a large corpus of customer reviews and quantify how good or bad a product is and with that understand what features make a product favorable or unfavorable. Additionally, relevant information about products and its features can sometimes be hard to identify or extract from a large volume of reviews. In Amazon specifically, there are also numerous cases where products have an overall average rating of 3 which is considered neutral. Such a situation can primary arise in 2 scenarios - one where majority of the customers generally rate a product 3, and the second case being where there’s a nearly equal number of customers giving a rating to a product that falls on the either extreme side of the rating scale. This happens when certain features specifically appeal to certain people. In such cases, it become pertinent for merchants to evaluate reviews on a user level rather than looking at the overall average rating. This can be very hard when the reviews are numerous. And so, there needs to be an easier way for merchants to be able to gauge user sentiment - be it positive or negative and also understand on a broader level, the common reasons for users projecting that sentiment.

One of the major factors contributing to this issue is the subjectivity involved on the user’s part in rating products. For instance, for the same product, two users can give exactly opposite rating based on the same reason - a user would probably give 5.0 to a dessert because it is sweet but another user would probably give 1.0 also because it is too sweet.

There exists some data on the web to understand and address such issues such as the ‘Fine Food Reviews’ dataset sourced from Amazon that we have leveraged for this analysis. The data contains reviews of various food items reviewed by users on Amazon. Each product reviewed has a review and rating associated with it given by a single user. There are other attributes provided as well. The dataset has been elaborated on in the ‘Data’ section.

***Normative Analysis***

*-*

***Stakeholders***

Amazon as a business heavily relies upon the relationship formed between its users and vendors through the use of their robust feedback system based upon product reviews and ratings. Hence, the stakeholders in this case is Amazon itself, the vendors selling products on Amazon that depend upon customer feedback for product intel, and customers or Amazon users who depend upon the feedback given by other customers to decide which products they should buy.

***Impact***

In this analysis, we have applied techniques such as TFIDF, Word2Vec and LDA for feature extraction that are eventually used in the binary classification of sentiment of reviews. Topic modeling of reviews using LDA can help cluster reviews of a similar kind. For eg – a product could have a positive sentiment associated with it for various reasons and user reviews can be grouped into categories depending on the reasons associated with that sentiment. Such structure can help merchants better understand their user feedback in context with the product features. It’s also equally helpful to other users who depend upon reading these reviews before they make a judgement about buying a product.

Apart from providing valuable market intelligence to businesses by helping them better understand their products and customers, the impact of this project can also be extended across other areas as this project can serve as a layer for transfer learning onto other sectors where a problem of such nature persists. For example - there can be websites where there is no provision for ratings or websites where there’s only user comments on articles and blog posts and there is need to quantify or understand customer perception.

1. **Objectives and Metrics**

***Objective***

By implementing sentiment analysis on the ‘Amazon Fine Food Reviews’ dataset, we intend to address the abovementioned problems faced by many businesses and merchants by -

* Building a model that can predict the sentiment (positive or negative), given a review
* Understanding the features that make a product good or bad by studying the user reviews

In transitioning to a sentiment classification model, we will be able to develop a method or a process that will have the ability to process text and classify it as a negative or positive sentiment. In the case where there is a mismatch between rating and a review - a review that projects a positive sentiment but is otherwise rated 3, our process might be able to correct that and classify it as a positive sentiment relative to other reviews and give the probability associated with that classification. Leveraging topic modelling can also help optimize the reviews but giving a more concise summary or a headline of a review.

***Metric***

#### Some of the standard metrics that we will look at to evaluate model performance is misclassification rate, F1-score/cross entropy, precision, recall, ROC area, and confusion matrix amongst others.

1. **Understanding the State-of-the-Art**

Sentiment analysis is a actually a pretty mature field where many algorithms already exist and one in which modules and libraries have been developed for automatic detection of sentiment. However, the accuracy goes drastically low when the rating is twisted with individually unrelated, unique and random features. Traditional approaches on sentiment analysis from scratch use word count or frequencies in the text which are assigned a sentiment value by experts and they often disregard the order of words[3]. To accurately rate a review that can be classified into 1 of the 5 categories still remains a very big challenge. For example - a review that contains the phrase “this candy is not that bad” and originally rated 4 can be misclassified as rated 1 or 2 because it contains the word “bad”. Such nuances in the way people characterize their sentiments towards a product make it a difficult task but using methods such as n-grams can be addressed. Recent work has been focused on other complicated RNN models such as recursive neural tensor network [1]. The data set and the question were actually drawn from a paper coming from Stanford.[4] Though, in their paper the main challenge was not sentiment analysis, their highest test accuracy was about 40% in their studies of users tastes and preferences changing and evolving over time. This low accuracy also showed that this was a challenging data set to analyze on.

1. **Hypothesis and Approach**

#### **Hypotheses**

* A review containing words like “excellent” or “horrible” which signify extreme sentiments will likely be classified towards the extreme end of the binary classification

(look at words)

* A review that is rated a score that falls on either extreme end of the classification spectrum is likely to get a higher review helpfulness score

(relation between rating and score)

* A review with a higher helpfulness score is more likely to be verbose specific to each product type

(relation between review length and help score)

* Data mining method X will produce a better ROC curve than method Y when applied to data Z.
* Adding predictor A to existing model M will improve the ROC performance of M.

**Data**

This dataset sourced from Stanford SNAP[4] consists of reviews of fine food products sold on Amazon.

The data spans over a period of more than 10 years, including all ~500,000 reviews up to October 2012 in one sqlite database. Reviews include product and user information, ratings, and a plain text review. Below is more information on the dataset:

Number of reviews: 568,454

Number of users: 256,059

Number of products: 74,258

Timespan: Oct 1999  -  Oct 2012

Number of Attributes/Columns in data: 10

The column or features in the dataset:

1. Id
2. ProductId -  unique identifier for the product
3. UserId - unique identifier for the user
4. Profile Name
5. Helpfulness Numerator  -  number of users who found the review helpful
6. Helpfulness Denominator  - number of users who indicated whether they found the review helpful or not
7. Score  - Rating between 1 and 5
8. Time  - Timestamp for the review
9. Summary  - Brief summary of the review

Text  - Text of the review

#### The ratings and reviews are both given by users whose economic, demographic, and other important data such as personal taste and preference is unknown. In such a situation the reviews and ratings will definitely be very biased, for eg. one user might feel a product is rightly priced while the other user might feel that the same listing is overpriced. Another bias that exists is that the data is highly skewed towards the higher ratings of 4 and 5.

**Approach**

Our data pipeline involved pre-processing data, performing exploratory data analysis on it and then building classification models and evaluating them. Below are the methods and steps followed -

* Data pre-processing:
  + Re-naming columns
  + Checking datatype and appropriate datatype conversion of variables (timestamp conversion)
  + Handling missing values and dealing with duplicated data – dropped 0.05% of the data
  + Feature Engineering - creation of variables – upvotes and total responses using variable ‘helpfulness\_score’ , creation of variable ‘extended\_review’ which is a concatenation of ‘summary’ and ‘text’ , and creation of variable word count for each review
  + Removal of neutral reviews - with rating 3
  + Classification of reviews rated 4 and 5 to 1 (positive sentiment) and 1 and 2 to 0 (negative sentiment)
* Exploratory Data Analysis:

(paste plots and mention takeaways)

* Modeling: An essential part of this process was Text Processing.

(write about text process techniques, math formula, models used,metric comparison,hypothesis results)