**justification of why the research question is important**

* This section includes: why are we doing this, how can this help for the social good, and possible ways to do this

Let us explain a few things about this research project before anything starts. This part of the writing will contain what this project is and why this is needed for the social good. We, as a group, are the people who are interested in social good and having to bring this beauty to the society by using the powers we gained over the educational journey of Minerva. Here, we have the ability to do some analysis on the dataset, make predictions, and most importantly the power to transform ideas into reality. As people who believe that there is lots of information, nowadays, that is used to manipulate or change human behavior for different purposes, we wanted to make an algorithm that is going to help people make a safer decision making in real life.

This is the reason why we landed on Amazon data as Amazon is a place where a lot of people buy and order from, especially after the COVID-19 pandemic where there has been a huge rise in the need of these contactless orderings and deliveries. However, it also has some drawbacks as the products we get may not be the things we have expected from the descriptions that are provided to us as customers. Things like clothes or anything that has to do with the design such as colors or performance of the things we buy like computers may not accurately deliver their flaws. Providers just make the descriptions convincing to buy rather than accurately delivering the descriptions of the product.

Furthermore, look at most of us, just reading the descriptions and briefly checking the reviews seldomly rather than thoroughly going through all the reviews before making a decision on buying this product. We just read the reviews that are at the top most which can be affected by the upvotes that are quite easy to manipulate. Thus, what we are going to do is important as our goal is to make an algorithm that checks the reviews (if it exists) to check whether the product a user is about to buy is accurately explained by the provider, or whether it is likely to regret buying this product that has misleading information.

It can help a lot of people, and save a lot of money from the users especially during these days when goods become expensive to buy, as well as having to pivot the society into the basin of attraction where the providers will have an easier time selling their product if they become honest. So, that by itself makes this project to be working for the social good.

**A clear explanation of the strategies we’ve used to collect data**

* This section includes: where we got the data, why we chose that database, and other things that we had from a class where to collect the data (have we used any skillsets to collect like HTML queries, or etc).

For this project, we got the data for Amazon reviews on products and the size of the dataset is 30GB. The data was collected by the University of California, San Diego. As it was collected by the institution, it lessened the stress of having skewed data that is purposely skewed to include positive or negative reviews only. And also, as there are other groups that have used this service or are going to use this dataset, it is easy to collaborate with outside authorities to gain and share insights that we get from the dataset as well.

Thus, as we are mainly going to use the dataset that is available already, we did not use any web-scraping techniques including HTML queries or going through Kaggle to find the data. This dataset itself was huge in size and had diverse categories that seemed to allow us to check the validation of the algorithm in diverse settings.

We tried to think of as broad as possible when it comes to data gathering as we want to have non-skewed and non-biased data as possible. We needed a lot of data to train the machine learning algorithm to tell us the likelihood of whether this product that I am about to buy is going to be a misleading information driven choice or not.

We also tried to choose the dataset that is going to be most applicable to our daily life, which is the Amazon reviews dataset where it contains different reviews that tells us whether the customer has negative or positive sentiments to tell the fact (just like how humans would have done when we are reading the reviews before buying the product). We thought of the goal, the reason for the project, and then defined the dataset’s requirements which made us find a dataset that fits the most.