**I. PROPOSED FRAMEWORK**

We used the below framework for our research:

**A. Dataset and its Features**

We used 5-core Amazon review dataset [1] provided by Jianmo Ni. The chosen dataset contains product reviews of Cell phones and Accessories purchased from amazon.com. It includes 1,128,437 rows and 11 features as explained below. Each row corresponds to a customer review, and includes the feature variables:

* reviewerID - ID of the reviewer
* asin - ID of the product
* reviewerName - name of the reviewer
* vote - helpful votes of the review
* style - a dictionary of the product metadata
* reviewText – customer review text
* overall - rating of the product
* summary - summary of the review
* unixReviewTime - time of the review (unix time)
* reviewTime - time of the review (raw)
* image - images that users post after they have received the product

**B. Preliminary Feature Selection**

Since our research is focused towards studying the sentiments from customer reviews and how it corroborates to the ratings; relevant features are selected for the analysis. Features – “reviewText”, “overall” and “summary” are considered.

**C. Data Pre-processing**

We used Bag of Words approach to analyse the reviews. For this approach, data is pre-processed using the following techniques:

* Convert the text to lower case
* Expand the contractions so we do not miss out any relevant sentiments such as haven’t etc.
* Remove the punctuations, digits and special characters
* Tokenize the text, filter out the adjectives used in the review and create a new column in data frame
* Use Negation handling to append preceding and successive negation clauses, before removing the stop words. For example – convert “not worth” to “not\_worth”
* Remove the stop words

**D. Analysis**

In order to study the sentiments and nature of words used in the review texts, the reviews are divided into 3 subgroups – positive, neutral and negative based on their corresponding ratings. Reviews with overall rating greater than 3, less than 3 and equal to 3 are labelled as positive, negative and neutral respectively. Word Cloud is used to visualise the most frequent words used in positive, negative and neutral reviews. The words are paired in order to conserve the ultimate contextual meaning of the sentence. For this motive, first unigram approach is used. But it did not work well for the negative reviews as it could not hold the negation clauses such as – “not recommend”. Bigram and trigram approaches which hold 2 words and 3 words together respectively are used so that the meaning of the text is conserved.

**E. Classification and Prediction Models**

[1] <https://nijianmo.github.io/amazon/index.html>