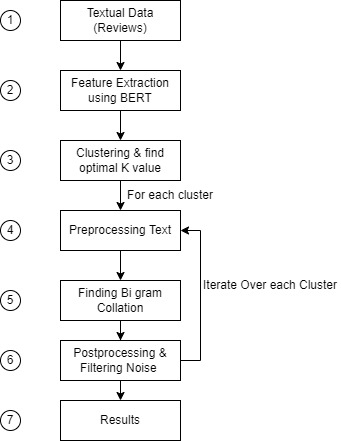
|  |  |
| --- | --- |
| Problem Modeling | Auto Labeling Engine (Bi Gram) |
| Short Description | This models extract the bi gram labels from the data using machine learning technique. |
| Technology & Imp. Lib | Python, Tensor flow, Sentence transformer, Scitkit Learn, Nltk, Numpy, Matplot lib etc. |
| Technique | Machine Learning, Deep Learning (Transfer Learning), Clustering, Bigram Collation |
| Resource Used | Ryzen 6 Core, 16 GB RAM, GPU 6GB 1060 |
| Experiment Duration | 18 hours |
| Experiment Assumption | 1. Data provided contain only English as a language which can be understood by the common (non-domain) person. 2. Most of the Data contains sequential information. No Ad hoc word provided. 3. Data can contains semantics words which are not present in English dictionary. 4. Solution is scalable enough provide decent result to other domain data. 5. Ratio of Known dictionary words are greater than unknown words. |

Auto Labeling (Bi gram) Problem

Solution:



|  |  |  |
| --- | --- | --- |
| 1 | Textual Data | Amazon Reviews were used as a part of the experiment provided in the assignment |
| 2 | Feature Extraction | Since the data (Amazon Reviews) was a natural data with a proper human sequencing and containing generic domain. I had used BERT pre trained network to extract features in 512 dimension vectors, since these model are state of the art and capable of finding the right attention on each words. I did not try other models due to time constraint. |
| 3 | Clustering and finding Optimal Cluster | Once the features were extracted, they were sent to clustering model with the range of 1 to 10 of K value. Once the clustering were done their inertia were used for plotting and getting the optimal K. |
| 4 | Preprocessing Text | For each cluster processing were done i.e. cleaning the text (removing punctuation, stop words, making text into lower case.) |
| 5 | Finding Bi Gram Collation | Once the preprocessing was done each cluster cleaned data was sent a Bigram collation model to generate the bi gram which was representing the bi gram label for a cluster. |
| 6 | Post Processing & Noise Filtering | Once the bi gram is found. Textual parsers are used to cleaned the bi gram labels |
| 7 | Results | Results are in the string format. |