SENTIMENT ANALYSIS AND PREDICTION OF RATINGS ON *Amazon customer review dataset for video games*

A PREPRINT

Amogh Vikram Sirohi
GCCIS, Rochester Institute of technology
Rochester, NY 14623
as4483@rit.edu

Dhrumil Mehta
GCCIS, Rochester Institute of technology
Rochester, NY 14623
dm9076@rit.edu

Sharjeel Ansari
GCCIS, Rochester Institute of technology
Rochester, NY 14623
sa2676@rit.edu

Vivek Lad
GCCIS, Rochester Institute of technology
Rochester, NY 14623
vl9244@rit.edu

August 10, 2020

1 Unsupervised learning on Word2vec embeddings

Word2Vec is a 2-layered Neural Network used for Natural Language Processing. It processes the text data by converting words to vectors. Thus, the input to the model is text corpus and the output is a set of vectors in numerical form which can be understood by deep neural networks.

1.1 Data Preparation

Since, implementing BERT is computationally expensive and scales as $\mathcal{O}(n(n-1)/2)$, we took first 2500 data-points from the dataset. Plotting the Class label and their frequency, we get the following graph: As it is evident above, the

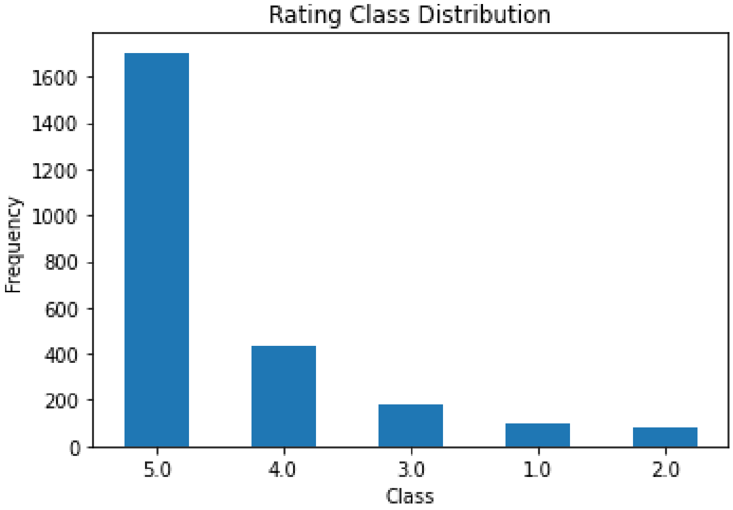


Figure 1: Frequency distribution of class labels, pre-sampling

dataset is biased towards class label 5.0, which is ≈ 68% of the entire dataset.
Therefore, before passing it through the ML algorithms, we oversample the dataset based on class labels, for which we used **SMOTETomek**, which is **S**ynthetic **M**inority **O**versampling **T**echnique Tomek method from *imblearn.combine* library. The resulting distribution is as depicted in Figure 2.

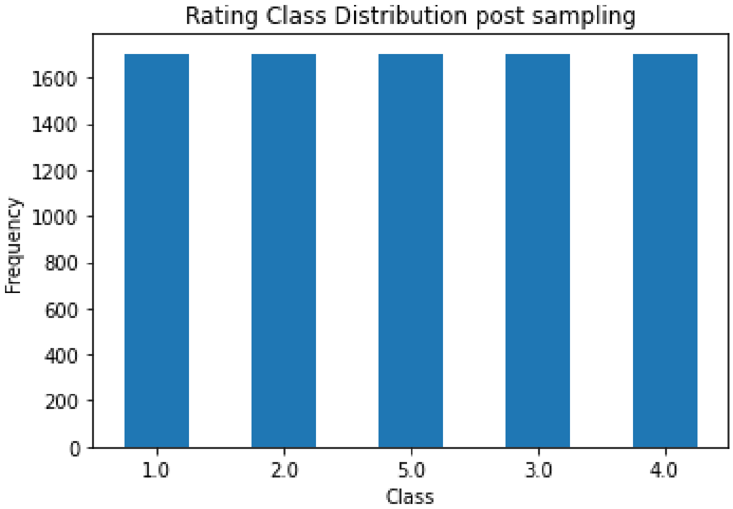


Figure 2: Frequency distribution of class labels, post-sampling

1.2 K-Means Clustering on Word2Vec embeddings

K means clustering is a method of dividing n observations in k clusters such that each observation belongs to one particular cluster with the nearest mean. It tries to keep intra cluster observations as close as possible and inter cluster observations as far as possible. The observations are assigned cluster such that the sum of the squared distance between the observation and clusters centroid is minimum.

1.2.1 Why KMeans?

One of the primary reasons KMeans can be a better choice of model for this task is that KMeans clustering guarantees convergence and completes with k clearly defined set of data points. Another advantage of KMeans clustering is that since it generalizes to cluster of different shapes fairly well than other clustering models, the high dimensional word embeddings of the reviews can have complex structures and hence can be well clusterized.

1.2.2 Determining optimal number of clusters

In order to determine the optimal number of clusters to be formed, we make use of Inertia which is the sum of squared errors for each cluster. Using the elbow method, we can see in the below diagram that the optimal number of clusters should be 3 because after that point, the reduction in inertia fluctuates.

1.2.3 Clustering

The following diagram shows the 2 components of each data points clusterized:

The following diagram shows some sample sentences assigned to each clusters. The samples can be hard to read because of cleaning and preprocessing but it shows that the sentences assigned to Cluster 0 show positive sentiment, Cluster 1 show neutral sentiment and Cluster 2 show negative sentiment.

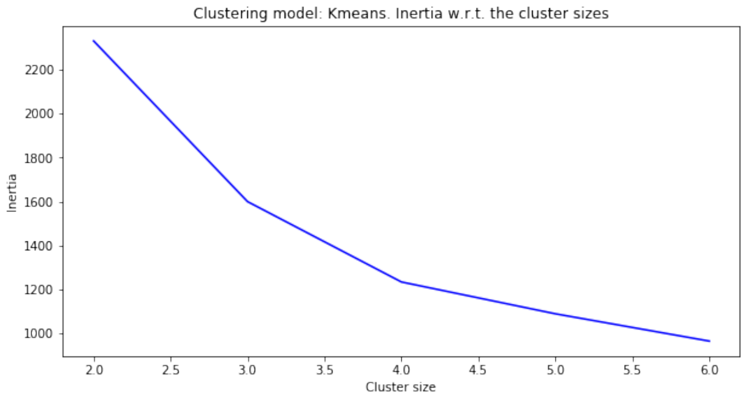


Figure 3: Inertia vs. cluster sizes

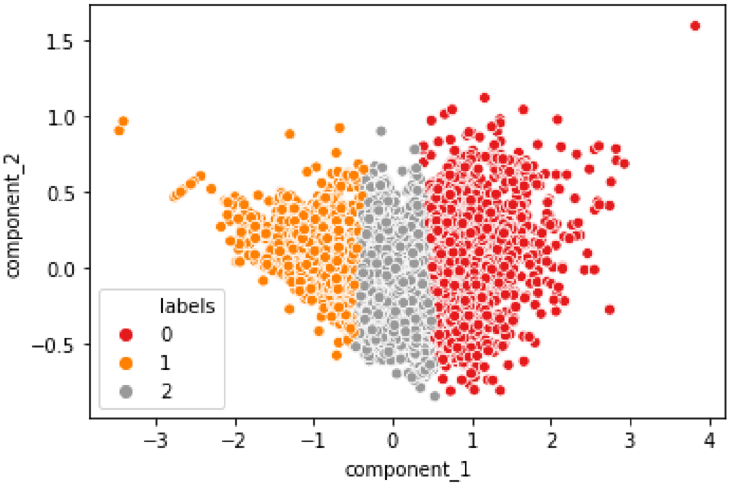
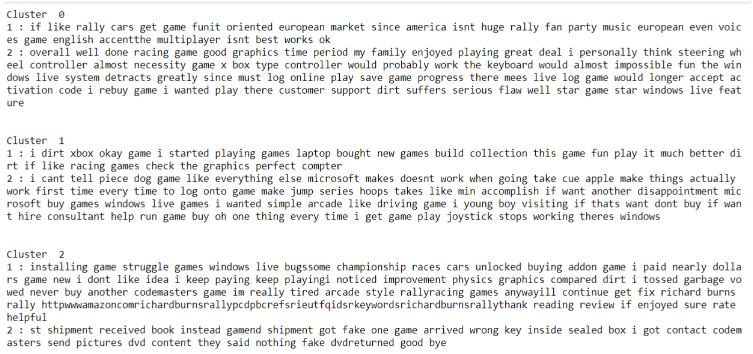


Figure 4: Clusters formed by KMeans clustering using Word2Vec



```
Cluster  0
1 : if like rally cars get game funit oriented european market since america isnt huge rally fan party music european even voic
es game english accentthe multiplayer isnt best works ok
2 : overall well done racing game good graphics time period my family enjoyed playing great deal i personally think steering wh
eel controller almost necessity game x box type controller would probably work the keyboard would almost impossible fun the win
dows live system detracts greatly since must log online play save game progress there mees live log game would longer accept ac
tivation code i rebuy game i wanted play there customer support dirt suffers serious flaw well star game star windows live feat
ure


Cluster  1
1 : i dirt xbox okay game i started playing games laptop bought new games build collection this game fun play it much better di
rt if like racing games check the graphics perfect compter
2 : i cant tell piece dog game like everything else microsoft makes doesnt work when going take cue apple make things actually
work first time every time to log onto game make jump series hoops takes like min accomplish if want another disappointment mic
rosoft buy games windows live games i wanted simple arcade like driving game i young boy visiting if thats want dont buy if wan
t hire consultant help run game buy oh one thing every time i get game play joystick stops working theres windows


Cluster  2
1 : installing game struggle games windows live bugssome championship races cars unlocked buying addon game i paid nearly dolla
rs game new i dont like idea i keep paying keep playingi noticed improvement physics graphics compared dirt i tossed garbage vo
wed never buy another codemasters game im really tired arcade style rallyracing games anywayill continue get fix richard burns
rally httpwwwamazoncomrichardburnsrallypcdpbcrefsrieutfqidsrkeywordsrichardburnsrallythank reading review if enjoyed sure rate
helpful
2 : st shipment received book instead gamend shipment got fake one game arrived wrong key inside sealed box i got contact codem
asters send pictures dvd content they said nothing fake dvdreturned good bye
```

Figure 5: Sample cluster-wise sentences for KMeans clustering using Word2Vec

1.3 Agglomerative clustering on Word2Vec embeddings

Agglomerative clustering is a type of hierarchical clustering technique which builds hierarchy of clusters by grouping the data points based on their similarity. In Agglomerative clustering, also known as AGNES, the clustering process starts by assigning each data points into their own clusters. Therefore at the beginning, the number of clusters are equal to the number of data points to be clustered. The Agglomerative clustering takes a bottom approach, which combines to smaller clusters which are similar to each other and forms a bigger cluster of the two. It repeats this process until there is all the data points are under one root ie. a big cluster.

1.3.1 Why Agglomerative clustering?

Agglomerative clustering, being a hierarchical clustering method can be useful where we do not want to pre-define the value of "k" clusters before hand. Even though it can be computationally expensive as compared to KMeans clustering, it produces intuitive results with hierarchy of clusters produced.

1.3.2 Clustering

The following diagram shows the 2 components of each data points clusterized:

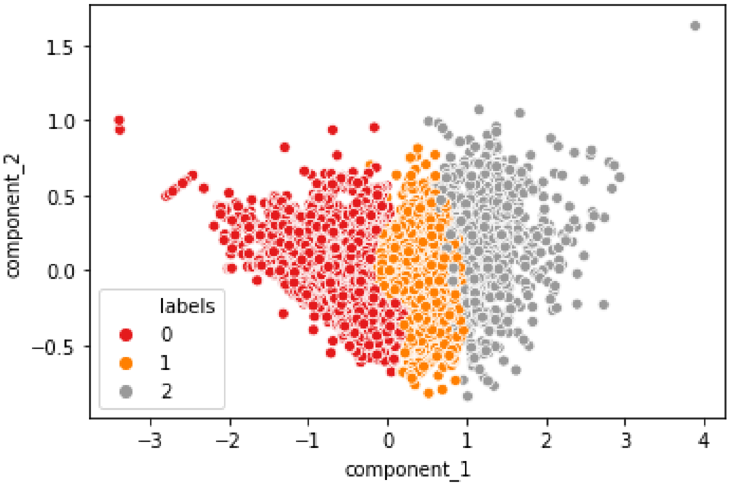
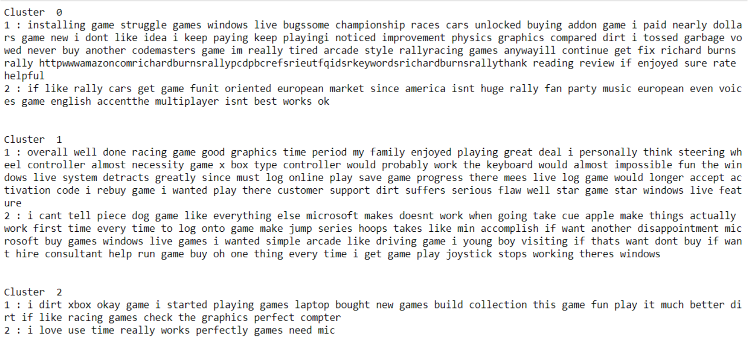


Figure 6: Clusters formed by Agglomerative clustering using Word2Vec

The following diagram shows some sample sentences assigned to each clusters.



```
Cluster  0
1 : installing game struggle games windows live bugssome championship races cars unlocked buying addon game i paid nearly dolla
rs game new i dont like idea i keep paying keep playingi noticed improvement physics graphics compared dirt i tossed garbage vo
wed never buy another codemasters game im really tired arcade style rallyracing games anywayill continue get fix richard burns
rally httpwwwamazoncomrichardburnsrallypcdpbcrefsrieutfqidsrkeywordsrichardburnsrallythank reading review if enjoyed sure rate
helpful
2 : if like rally cars get game funit oriented european market since america isnt huge rally fan party music european even voic
es game english accentthe multiplayer isnt best works ok


Cluster  1
1 : overall well done racing game good graphics time period my family enjoyed playing great deal i personally think steering wh
eel controller almost necessity game x box type controller would probably work the keyboard would almost impossible fun the win
dows live system detracts greatly since must log online play save game progress there mees live log game would longer accept ac
tivation code i rebuy game i wanted play there customer support dirt suffers serious flaw well star game star windows live feat
ure
2 : i cant tell piece dog game like everything else microsoft makes doesnt work when going take cue apple make things actually
work first time every time to log onto game make jump series hoops takes like min accomplish if want another disappointment mic
rosoft buy games windows live games i wanted simple arcade like driving game i young boy visiting if thats want dont buy if wan
t hire consultant help run game buy oh one thing every time i get game play joystick stops working theres windows


Cluster  2
1 : i dirt xbox okay game i started playing games laptop bought new games build collection this game fun play it much better di
rt if like racing games check the graphics perfect compter
2 : i love use time really works perfectly games need mic
```

Figure 7: Sample cluster-wise sentences for Agglomerative clustering using Word2Vec

2 Unsupervised learning on BERT embeddings

BERT (Bidirectional Encoder Representations from Transformers) made by researchers at Google AI Language has changed NLP by presenting state-of-the-art results in a wide variety of tasks. It essentially uses cross-encoder networks that take 2 sentences as input to the transformer network and then predict a target value and conceptually involves encoding a sentence or short text paragraphs into a fixed length vector (dense vector space) and then the vector is used to evaluate how well their cosine similarities mirror human judgments of semantic relatedness.

2.1 KMeans Clustering

2.1.1 Determining optimal number of clusters

The value of k is either provided as an input. Here, we determine the optimal value of k by iterating k over a range of 2 to 20, and calculating the **Silhouette Scores** of K means clustering for each k, and choose k for which the said score is maximum. The resulting graph is shown in Figure 8.

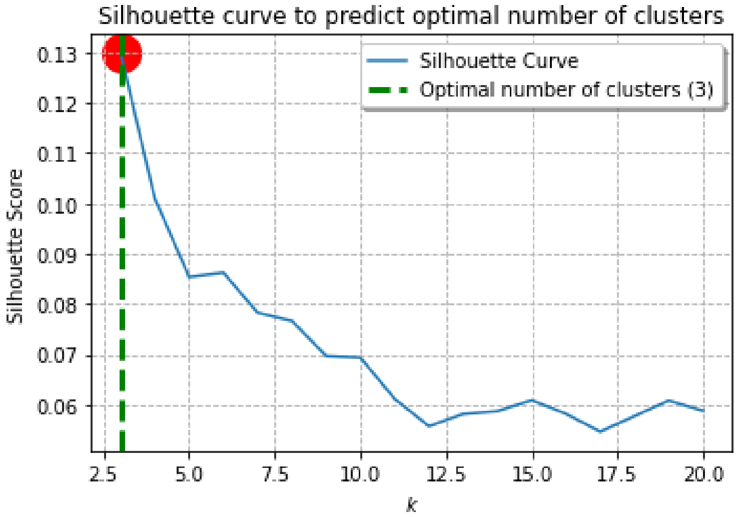


Figure 8: Silhouette score vs Number of class

2.1.2 Clustering

Once we determined the optimal number of clusters (here, 3), we go ahead with implementing K Means clustering algorithm on our sampled dataset. For doing this, we use KMeans library from *sklearn.cluster* package.

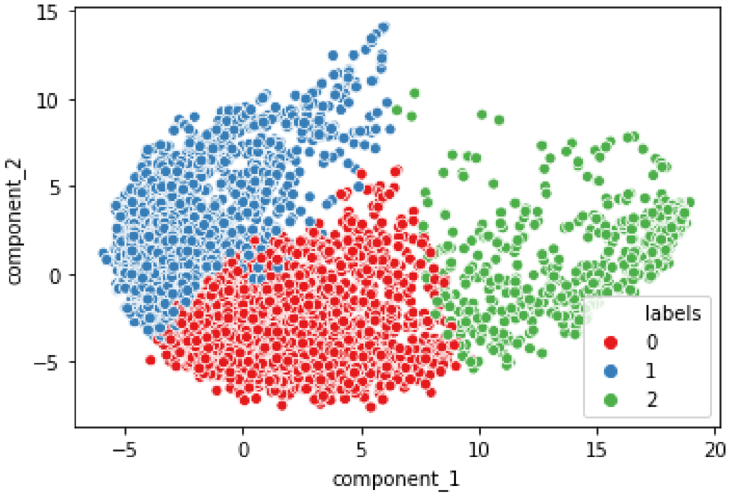
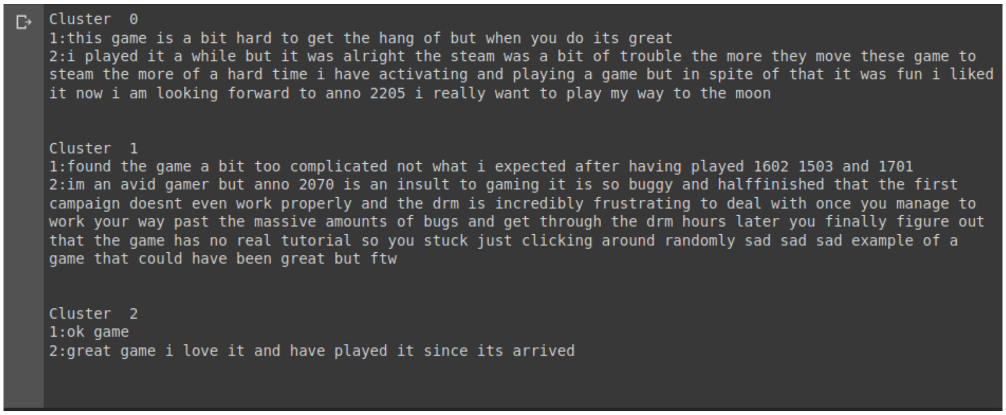


Figure 9: KMeans Clustering on BERT



```
Cluster  0
1:this game is a bit hard to get the hang of but when you do its great
2:i played it a while but it was alright the steam was a bit of trouble the more they move these game to
steam the more of a hard time i have activating and playing a game but in spite of that it was fun i liked
it now i am looking forward to anno 2205 i really want to play my way to the moon


Cluster  1
1:found the game a bit too complicated not what i expected after having played 1602 1503 and 1701
2:im an avid gamer but anno 2070 is an insult to gaming it is so buggy and halffinished that the first
campaign doesnt even work properly and the drm is incredibly frustrating to deal with once you manage to
work your way past the massive amounts of bugs and get through the drm hours later you finally figure out
that the game has no real tutorial so you stuck just clicking around randomly sad sad sad example of a
game that could have been great but ftw


Cluster  2
1:ok game
2:great game i love it and have played it since its arrived
```

Figure 10: Sample Clustering

2.2 Agglomerative clustering on BERT

2.2.1 Determining optimal number of clusters

Similar to KMeans, we have used silhouette score to determine optimal number of clusters for Agglomerative clustering. We iterate through 2 to 20 clusters and plot a graph of silhouette score vs. number of clusters, and we get the following graph.

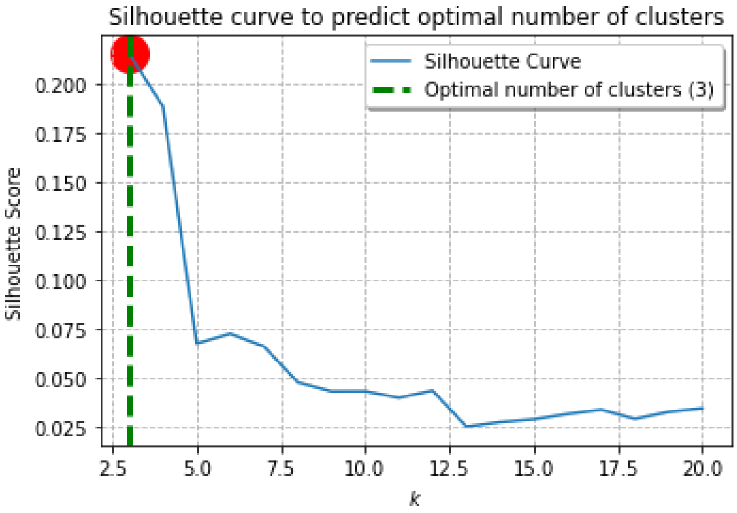


Figure 11: Silhouette Score plot

As expected, the optimal number of clusters is again 3.

2.2.2 Clustering

Once we have the optimal number of clusters, we implement Agglomerative clustering on our sampled dataset. We use AgglomerativeClustering library from *sklearn.cluster* package.
The resulting scatter plot is as follows:

3 Results

To compare the two models, we make use of CrossTab, which is a tabular representation, depicting the distribution of classes across the clusters. In this, the rows represent the clusters and the columns represent the class labels .These crosstabs are as follows:

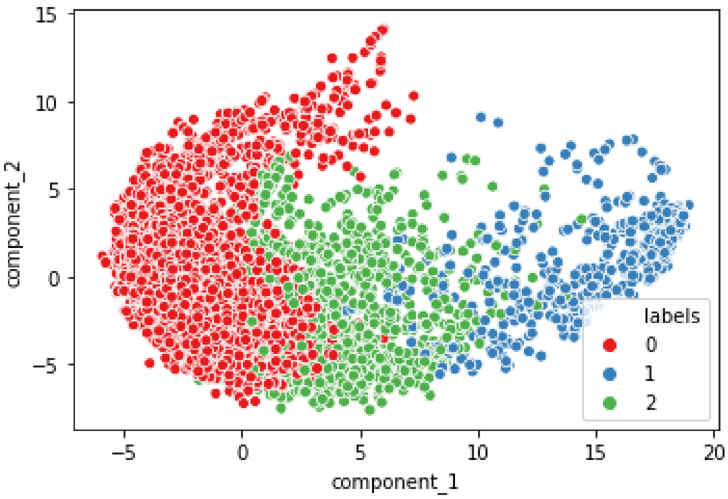
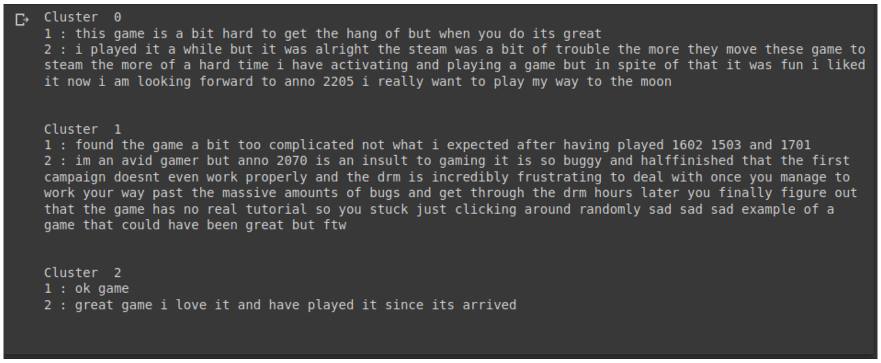
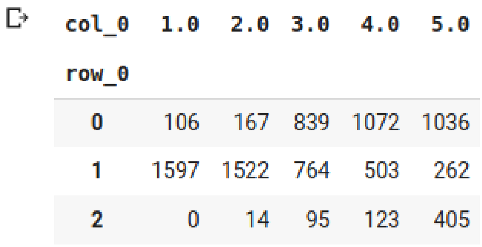


Figure 12: Scatterplot of Agglomerative Clustering



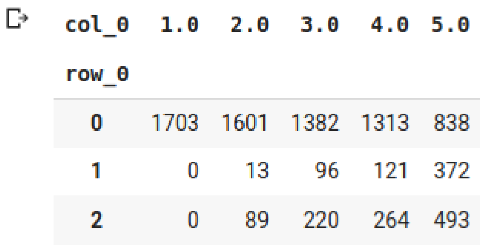
```
Cluster  0
1 : this game is a bit hard to get the hang of but when you do its great
2 : i played it a while but it was alright the steam was a bit of trouble the more they move these game to
steam the more of a hard time i have activating and playing a game but in spite of that it was fun i liked
it now i am looking forward to anno 2205 i really want to play my way to the moon


Cluster  1
1 : found the game a bit too complicated not what i expected after having played 1602 1503 and 1701
2 : im an avid gamer but anno 2070 is an insult to gaming it is so buggy and halffinished that the first
campaign doesnt even work properly and the drm is incredibly frustrating to deal with once you manage to
work your way past the massive amounts of bugs and get through the drm hours later you finally figure out
that the game has no real tutorial so you stuck just clicking around randomly sad sad sad example of a
game that could have been great but ftw


Cluster  2
1 : ok game
2 : great game i love it and have played it since its arrived
```

Figure 13: Sample clusters



col_0	1.0	2.0	3.0	4.0	5.0
row_0					
0	106	167	839	1072	1036
1	1597	1522	764	503	262
2	0	14	95	123	405

Figure a: K Means BERT



col_0	1.0	2.0	3.0	4.0	5.0
row_0					
0	1703	1601	1382	1313	838
1	0	13	96	121	372
2	0	89	220	264	493

Figure b: Agglomerative BERT

Figure (a) depicts how K means classifies the classes into 3 clusters. Cluster 0 consists mainly of neutral and positive classes (ratings 3, 4 and 5), cluster 1 consists mainly of negative and neutral classes (ratings 1, 2 and 3) and cluster 2 consists of positive classes (mostly rating 5). So, we can say that K Means on BERT does a decent job of classifying reviews into 3 classes - Positive, negative and neutral. Whereas, from Figure (b), we can infer that Agglomerative clustering on BERT embedding does a bad job when compared to K Means on BERT. Here, cluster 0 represents all classes and can be extremely ambiguous.

4 Conclusion and Future Work

4.1 Comparison with Supervised learning

In this milestone, we looked at some unsupervised learning tasks over the Amazon video games reviews dataset. The primary goal of the unsupervised learning techniques is to identify different patterns that exists among the different data points. In supervised learning milestone, we simply analyze the data and work on training an equation which can help us predict a particular variable by minimizing the error. For example, we worked on predicting the sentiment class (positive, negative or neutral) for the review. However in unsupervised learning milestone, we analyzed the data for pattern recognition and not for predicting a particular variable. For example, in our unsupervised learning, we analyzed the reviews without considering the sentiment classes and attempted to check if any common pattern is recognized amongst them.

4.2 Further improvements

In our unsupervised learning milestone, we did not use our complete dataset because agglomerative clustering is computationally expensive (quadratic space complexity and cubic time complexity). We can further experiment by choosing to use more data for our clustering models and see if there are any improvements in the results that we get. Experimentation by using different models such as top-down based divisive clustering model, can also help.

4.3 Combining Supervised and Unsupervised learning

We can combine the two techniques to perform complex tasks, which involves pattern recognition techniques of unsupervised learning and using it for predictive analysis involving supervised learning techniques. For example, if there is no proper structure to the raw dataset, we can create some structures in the dataset by forming clusters based on some features and then using the clusterized dataset for some supervised learning tasks such as classification or regression. Hence this kind of approach, involves both supervised and unsupervised learning.

References

[1] Ruining He and Julian McAuley. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. WWW, 2016.

[2] Julian McAuley, C. Targett, J. Shi and A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015

[3] Wanliang Tan, Xinyu Wang, Xinyu Xu: Sentiment Analysis for Amazon Reviews

[4] Y. Xu, X. Wu, and Q. Wang. Sentiment analysis of yelpsratings based on text reviews.