Sentiment Analysis on reviews from Amazon

For the project you will perform sentiment analysis on reviews from amazon. The data to use is located here:http://deepyeti.ucsd.edu/jianmo/amazon/categoryFilesSmall/Cell_Phones_and_Accessories_5.json.gz . Note it is a gziped json file. If you'd like to download and extract it directly into colab, this can be done using the following lines: !curl http://deepyeti.ucsd.edu/jianmo/amazon/categoryFilesSmall/Cell_Phones_and_Accessories_5.json.gz -o reviews.json.gz

!gunzip reviews.json.gz

At this point it should be easily ingested into a panda's data frame. A few relevant fields are

reviewText: The text from the review

summary: The summary text from the review

overall: The score the reviewer gave the item.

Your task:

Perform the necessary transformations to train both regression and classification models to predict the 'overall' field in the data set. This should include creating the correctly sized training and test sets.

When performing the classification task, use overall less than 3 as negative, greater than 3 as positive, and 3 as neutral. If you'd prefer a numeric value 0 should be negative, 1 neutral and 2 positives.

You may wish to drop the columns image, style and votes. You may also wish the drop duplicate data.

There are several options for using the summary and reviewText together, such as concatenating the strings, training separate models on both and feeding those results into another models, etc. You may find that you don't want to use all fields (other than the ones I suggested dropping). I will let you experiment with this, just explain what you did and why.

You should certainly apply vectorization and perhaps a pca or nmf as well. Try at least three different classifiers/regressors. Attempt to get the best possible result, remember the different metrics we discussed for evaluating models. Discuss which metric you this you should optimize for and why. Pipelines and grid search will certainly help in optimizing your results!

Write a 5 page or less paper describing your work and the performance you were able to achieve.

1. General Description of the Data:

The data has 1,128,437 rows and 12 columns. The size of the data is 624 MB. The columns are Overall, Reviews and summary by the customer, the verification of the review, date of verification, reviewer id, reviewer name, asin, review time, style, vote and image of the product. We will be using only these columns-'reviewText', 'summary' and 'overall'.

2. Data Pre-Processing:

- Dropped the columns image, style and votes as it was not useful in our analysis.
- Dropped null (~1000 rows) and duplicate data (~4000 rows) as it was negligible compared to the huge size of the data.
- Changed the data type of 'reviewtime' column to datatime format.
- Merged the reviewText and summary columns into one column Reviews to simplify process.
- Created a new column 'target' for classification whose values are:
    - 0, if 'overall' < 3 (negative)
    - 1, if 'overall' = 3 (neutral)
    - 2, if 'overall' > 3 (positive)

3. Text Pre-Processing:

The following operations were performed on Reviews

- Tokenized the text into words
- Converted all words to lower case.
- Words containing anything other than alphabets are removed (punctuations, links, numbers etc. are removed).
- Removed stop words.

4. Text Vectorization:

- Word2Vec model is used to convert the text to vectors.
- The vector size for each review is 50.
- The Word2Vec model vectorizes words which repeats multiple times in the whole dataset to abstain from vectorizing words which have spelling botches, and so on.
- The vectors are placed in a data frame, which will be the input to train and test the Machine Learning models.

5. Classification Models:

These three models are applied:

Model:	**Logistic Regression**	**Linear SVC**	**Decision Tree Classifier**
Accuracy Score (on test data)	84.1%	84.5%	83.9%
F1 Score	0.843	0.836	0.807
Optimal Parameters	C = 0.01	C=10	Max Depth = 10
Confusion Matrix	30122 4677 6649 8080 10439 10696 11492 11973 242799	30851 2497 8100 10301 5969 12945 12912 5502 247850	21227 605 19601 5414 3003 20912 7247 536 258382

6. Regression Models:

A. Linear Regression:

- Accuracy = 72.375 %
- Coefficient of determination R2 score = 0.463
- Mean square error = 0.815

B. Decision Tree Regressor:

- Accuracy = 74.247 %
- Coefficient of determination R2 score = 0.449
- Mean square error= 0.837
- Optimal parameters max depth = 8