Sentiment Analysis of Amazon

1. The columns in the train dataset are : ['name', 'brand', 'categories', 'primaryCategories', 'reviews.date', 'reviews.text', 'reviews.title', 'sentiment']
2. The shape of the train dataset is (4000, 8)
3. The shape of the test dataset is (1000, 7)
4. There are 10 null values in reviews.title in training and 3 null values in reviews.title in testing dataset. We can drop those rows from the datasets
5. The shape of the train dataset after removing null values is (3990, 8)
6. The shape of the test dataset after removing null values is (997, 7)
7. We observe that there is no other brand in the dataset other than amazon hence we can drop brand column
8. Secondly categories column is generalized to primary categories hence we can drop that column
9. Name column can be generalized to categorical column that describe the product type like:
    a. Kindle E-redear and other Tablets as tablets
    b. TV
    c. Amazon Tap and Amazon Echo as Bluetooth Speaker
    d. Battery Charger
10. We can also drop the reviews.title and reviews.date as that does not make any affects to model compare to reviews.text
11. The shape of the train dataset after removing ['name', 'brand', 'categories','reviews.date', 'reviews.title'] col is (3990, 4)
12. The shape of the test dataset after removing b ['name', 'brand', 'categories','reviews.date', 'reviews.title'] col is (997, 3)
13. We observe that out of 3980 data 3739 data is itself of the positive sentiment. And neutral and negative data are less 10% of the data. Hence our data is too imbalance.
14. The shape of Train Data TF-IDF Score array (3990, 3928)
15. The shape of Test Data TF-IDF Score array (997, 3928)
16. The shape of Train padded tokenised_seq data(3990, 1559)
17. The shape of Test padded tokenised_seq data(997, 1559)
18. Based on MulitnomialNB classifer accurracy is good enough but the precision, f1 and recall score on consider equal weights to each label is too low.
19. Based on RandomForest Classifer the accuracy is very good of the model also the f1 score for the mode has increase quite high comparatively
20. Based on XGBoost Classifer the accuracy is good of the model also the f1 score for the model has increase quite high comparatively.
21. Based on SVM Classifer the accuracy is very good of the model also the f1 score of the model is quite like XGBoost.
22. NN with LSTM and GRU has good accuracy but its f1 scores are not good enough compared to other good models.
23. Based on comparison we can conclude that SVM classifier has the best accuracy and f1 score is also good comparatively
24. Based on NMF model generated word arrays, here are the topics that can be identified:
    a. Topic 1: Evaluating the Amazon Fire Tablet

b. Topic 2: Gifting Amazon Fire Tablets
c. Topic 3: Positive Opinions about Smart Home Devices
d. Topic 4: Simplicity and Usability of Kindle E-readers
e. Topic 5: Comparing Kindle and Other E-readers
f. Topic 6: Overall Opinions about Tablets This
g. Topic 7: Tablet Purchasing and Recommendations
h. Topic 8: Positive Opinions about Kindle E-books
i. Topic 9: Tablet Durability and Age-Appropriateness
j. Topic 10: Parental Control and Content Management

25. Used the NMF (Non-negative Matrix Factorization) instead LDA (Latent Dirichlet Allocation) because as NMF assumes that each review is a mixture of topics and LDA assumes that each review is a mixture of these hidden topics, and each word in the review is generated from a specific topic with a certain probability. Hence NMF generate generalized topics based on all reviews for this sentiment analysis data.