RECOMMENDER SYSTEM- REVIEW BASED

TEAM 10

NANDINI JAMPALA

Dataset:

Source: https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

This is a list of over 34,000 consumer reviews for Amazon products like the Kindle, Fire TV Stick, The dataset includes basic product information, rating, review text and more for each product

Checking for null values:

```
reviews.rating 33
reviews.text 1
reviews.title 5
reviews.username 2
dtype: int64
```

	reviews.rating	reviews.text	reviews.title	reviews.username
0	5.0	This product so far has not disappointed. My c...	Kindle	Adapter
1	5.0	great for beginner or experienced person. Boug...	very fast	truman
2	5.0	Inexpensive tablet for him to use and learn on...	Beginner tablet for our 9 year old son.	DaveZ
3	4.0	I've had my Fire HD 8 two weeks now and I love...	Good!!!	Shacks
4	5.0	I bought this for my grand daughter when she c...	Fantastic Tablet for kids	explore42

PREPROCESSING AND FEATURE EXTRACTION:

The goal of this project is to classify reviews. Also, after removing the missing values in the for ‘reviewerID’, ‘asin’, ‘reviews.text’ and ‘reviews.rating’ features. Preprocessing and Dimensionality reduction are to be done only for the ‘reviews.text’ feature

- we need to vectorize our input variable.

-we use the count vectorizer function which returns a vector array.

DIMENSIONALITY REDUCTION:

- Remove all the stop words from the reviews as they do not provide any important information for a review's sentiment. Stop words include words like 'is', 'at', 'which', 'on' etc. These are the most commonly occurring words in the English Language.
- Remove punctuations.
- Remove words occurring less that 1% of the time. This is needed as there might occur cases when a person mentions a brand of the product or uses some word not common to sentiment analysis. If we include such words we might throw off our sentiment classifier.

gave christmas gift inlaws husband uncle loved easy use fantastic features

I gave this as a Christmas gift to my inlaws, husband and uncle.
They loved it and how easy they are to use with fantastic features!

	feature	coef
42481	terrible	-20.352050
48058	will definitely	-19.495217
10646	done great	-19.284680
38808	slow	-18.334396
18449	great year	-17.885817
36300	returning	-17.515921
32516	price isn	-16.317201
45890	using firestick	-16.302561
16576	gift three	-16.147438
36261	return	-16.097260

Models Used:

The algorithms used were specifically used to classify the reviews as positive or negative. As a result only classification algorithms were used and the best one from 5 of them was selected.

1. Naïve Bayes: The reason for choosing Naïve Bayes is that empirical results have proven that Naïve Bayes performs a very good task at sentiment analysis. It is used in many email-Spam filters.

2. AdaBoost Classifier: Adaboost algorithm is adaptive and the subsequent weak learners are tweaked so that the previously misclassified points are classified correctly. It is also less susceptible to over fitting and hence a better choice for our problem.

3. Logistic Regression: Logistic Regression in our problem gives us the probability of a review being classified as positive or negative. Thus, every review can either be positive or negative with some probability.

4. Random Forest Classifier: The classifier creates many random decision trees at the training time and finally prediction is based on the mode or the mean of the classes predicted by those random trees.

5. SVM: The reason for trying out the support vector machine is that if the true distribution of the polarity of reviews is not linearly separable then an SVM with 'rbf' kernel would help in giving us nonlinear boundaries.

Performance Metrics:

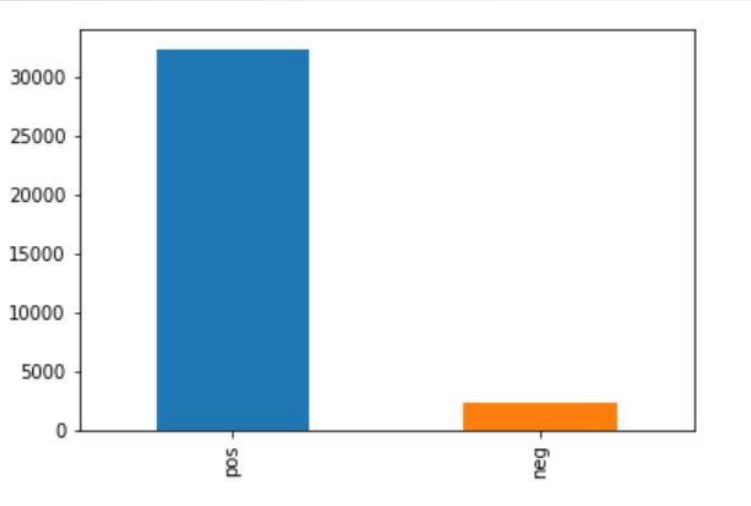
1. F-Beta Score:
a model's ability to precisely predict those reviews that have an actual rating of 4 or 5 is more important than the model's ability to recall those reviews. We can use F-beta score as a metric that considers both precision and recall.
when β=0.5, **more emphasis is placed on precision**. This is called the F0.5 score

$$F\beta = (1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+recall}$$

2. ROC_AUC:
AUC(area under the curve of the Receiver Operating characteristics) is used because since the data is highly biased towards one review class positive, our classifier might decide to classify all the reviews of the test set to be in the positive class, and we will still get a very high accuracy, but such a classifier would be of no use for a new data. However, when we use F0.5 score and AUC we get the insights of the True Positives and False Positives through F0.5 score and The True Positive rate and False Positive rate through the ROC and AUC.

Model Selection

step 1:

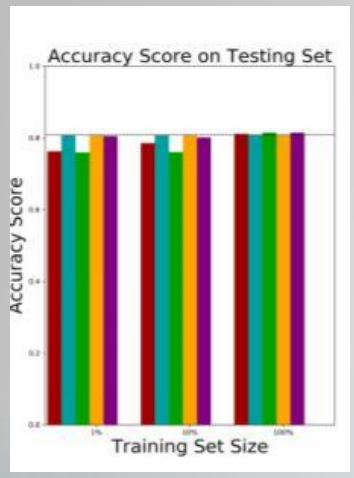


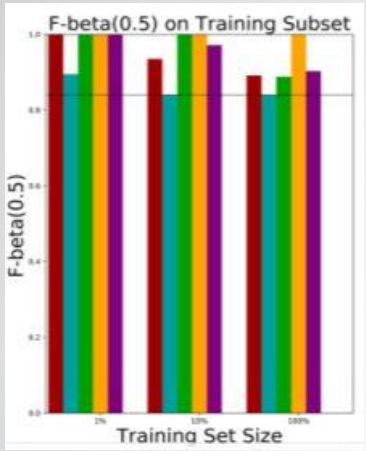
Naïve bayes :

accuracy	precision	recall	F score
0.8078	0.8078	1	0.8401

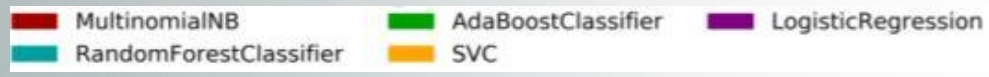
we should build a model that is better than this Naïve model and has F0.5 score better than this Naïve predictor.

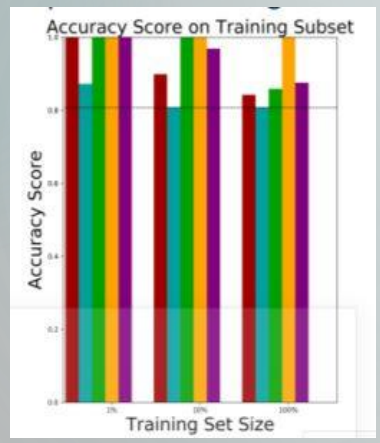
STEP-2(Model Selection):

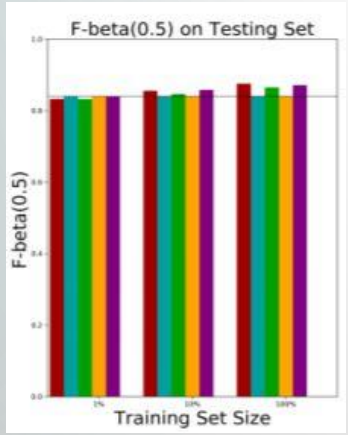


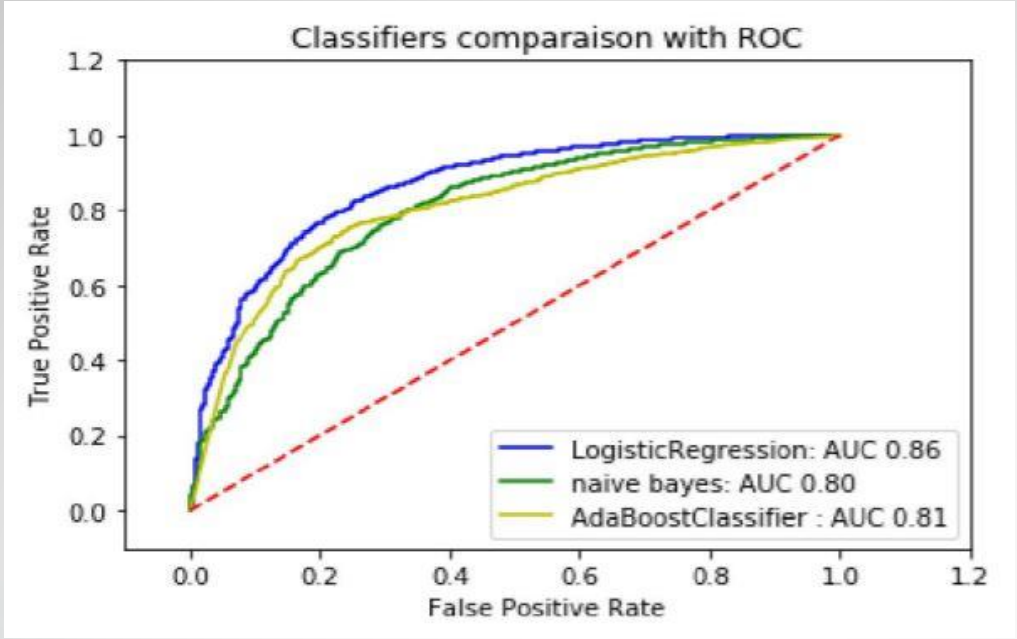


Models	Train acc	Test acc	Train F-score	Test F-score
Naïve bayes	0.84	0.81	0.90	0.87
Random forest	0.81	0.80	0.85	0.84
AdaBoost	0.85	0.81	0.89	0.87
SVM	0.99	0.80	0.99	0.84
Logistic Regression	0.87	0.81	0.90	0.87









from the above graph and table, Naïve Bayes, AdaBoost and Logistic regression perform the best.

since this is a huge dataset, and the time taken for training SVM is pretty high as compared to other models

the standard deviation of the distribution of AdaBoost classifier is less as compared to that of logistic regression.

Because of these reason I have selected AdaBoost classifier as my final model

Step 3:

Having selected the AdaBoost classifier, tune the Hyper parameters in order to get the best model.

Strengths: The main strength of AdaBoost is that it has very less hyper parameters to tune(namely **'base_estimator', 'n_estimator' and 'learning_rate'.** Due to this the AdaBoost algorithm is pretty fast as compared to other algorithms like neural networks and SVM.
Weakness: AdaBoost algorithm is highly affected by the quality of data that we are working with. If our data has quite some number of outliers, then such outliers would affect the performance of the Adaboost classifier.
but we have cleaned it to quite some extent. This means that when we removed the low frequency words we were actually removing the outliers. As a result I think that using the Adaboost algorithm would generate good results on the test data

AdaBoost Hyper-Parameter Tuning:

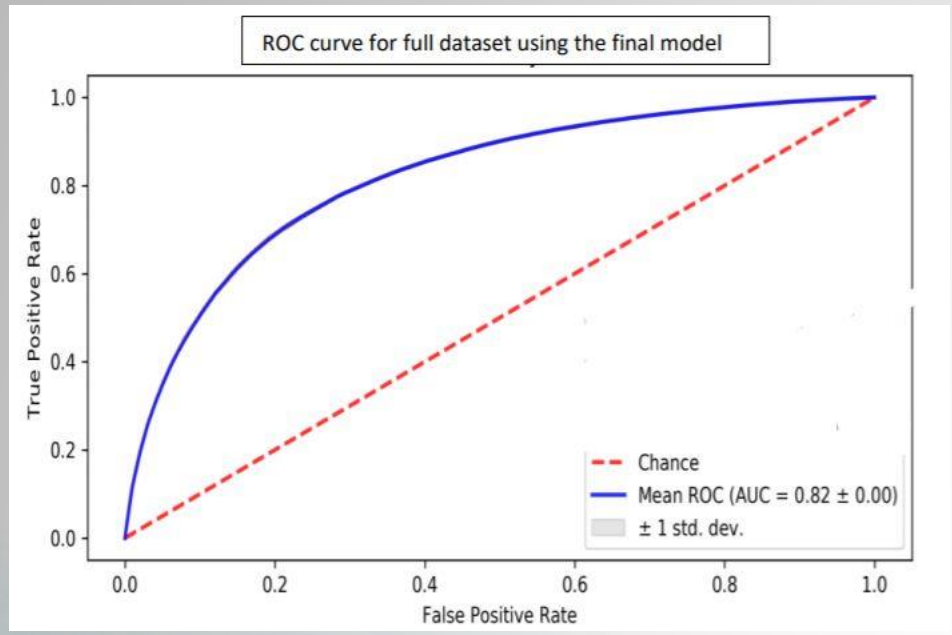
The hyper parameters that are to be tuned for the AdaBoost classifier are ‘n_estimators’ and the ‘learning rate’. The base estimator is set to be equal to a Decision stump.
The hyper-parameter tuning is done on the 80% training or development set.
->n_estimators: A total of 90 values starting from 100 and going till 1000 with a step of 10.
-> learning_rate: values starting from 0.5 to 2 with a step size of 0.02.
-> After doing the GridSearch, the final best estimator has been found.

Final model is:

Base estimator	n_estimators	Learning rate
Decision tree	500	1.5

	accuracy	F-score
training	0.8334	0.8712
testing	0.8325	0.8704

34630	4.0	My new Kindle DX2 graphite came yesterday and ...	Very pleased with cover	Bill Wood	1
34631	1.0	This cover is made for the Kindle DX and fits ...	Great protection if it doesn't damage the Kind...	K.A. Smith	0
34632	5.0	I have several covers, from simple rubber cove...	very good fit and some protection for Kindle DX	Martijn13Maart1970	1
34633	1.0	The reason people give it 5 stars is because t...	Eventually Your Kindle Will Crack	Alexander Rodriguez	0
34634	5.0	This cover is a must have for your Kindle read...	Love my Kindle	wawasee	1
34635	5.0	I have this cover in black and my cover does i...	Very classy cover!	David	1
34636	5.0	I found this cover to be quite handsome in app...	Product appears to be updated	Nicholas Sabalos, Jr.	1
34637	1.0	I, as well as other Kindle lovers, have notice...	Great at first, warped within 3 months	Karen Henning	0
34638	3.0	This cover is a replacement for the initial co...	Kindle Cover Issue	LoriC	1
34639	5.0	Having recently received a Kindle Fire HD as a...	Extremely Handy!	Johnny K. Young	1
34640	5.0	Surpassed my expectations it charges faster th...	Surpassed expectation	GIMBO2006	1



	Testing Accuracy	Testing F0.5
Base Model(Naïve pred)	0.8078	0.8401
Final Model	0.8325	0.8704

Reviews.rating	Reviews.text	Reviews.title	Reviews.username	Summary_clean	words	naive	Ada	Ran	log
NaN	The Kindle is my first e-ink reader. I own an ...	Worth the money. Not perfect, but very very go...	Jeffrey Stanley	the kindle is my first e ink reader i own an i...	[the, kindle, is, my, first, e, ink, reader, i...	neg	pos	neg	pos
NaN	I'm a first-time Kindle owner, so I have nothi...	I Wanted a Dedicated E-Reader, and That's What...	Matthew Coenen	i m a first time kindle owner so i have nothin...	[i, m, a, first, time, kindle, owner, so, i, h...	neg	pos	neg	pos
NaN	UPDATE NOVEMBER 2011:My review is now over a y...	Kindle vs. Nook (updated)	Ron Cronovich	update november my review is now over a year o...	[update, november, my, review, is, now, over, ...	neg	pos	neg	pos
NaN	I'm a first-time Kindle owner, so I have nothi...	I Wanted a Dedicated E-Reader, and That's What...	Matthew Coenen	i m a first time kindle owner so i have nothin...	[i, m, a, first, time, kindle, owner, so, i, h...	neg	pos	neg	pos