Customer Review Analysis of Amazon Products

Kavya Janga 94814
Pavani Eupuri 94850
Kiran Mai Bandaru 94966

Prof : ALex WU

Introduction

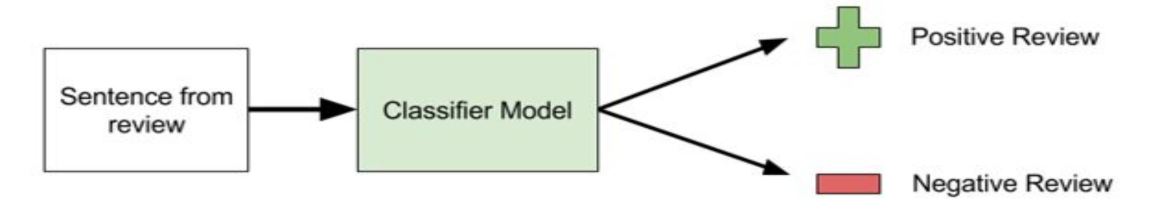
- Amazon is world widely known Ecommerce website. Initially it is known for huge collection of books but later it was expanded for other items and now it sells products too.
- Customer satisfaction and opinion is important for ecommerce websites. This gave rise “User Reviews”.
- User Reviews are customer suggestions which help other customers to make decision about that product.

Dataset

- Dataset consists a list of customer reviews for amazon products as of 2019.
- It contains over 28000 records and 24 attributes.

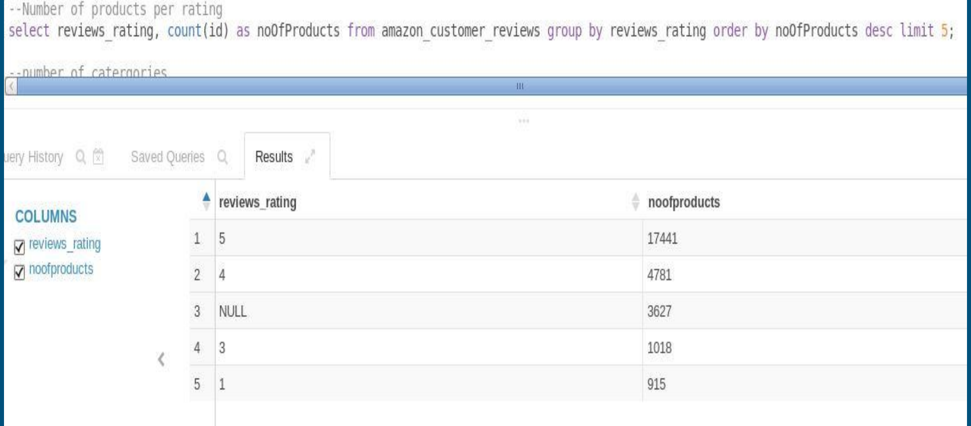
Source:https://data.world/datafiniti/consumer-reviews-of-amazon-products/workspace/file?filename=Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19.csv

Objective: Classifying amazon product reviews based on customer ratings.



Data Analysis

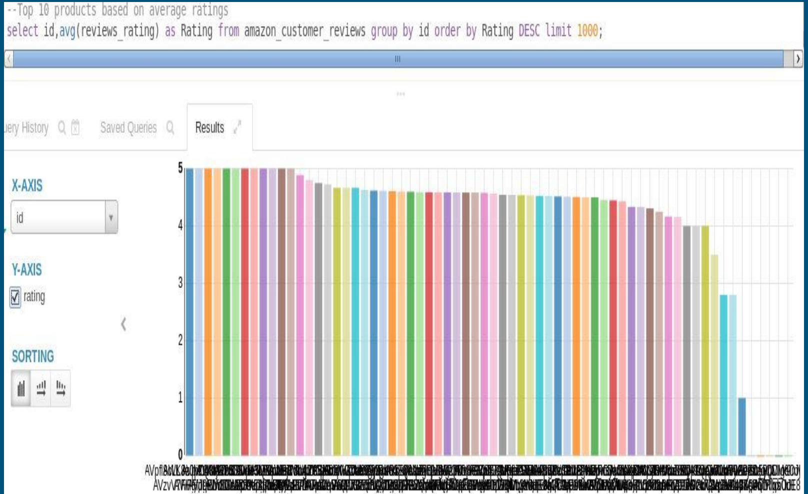
Overall ratings count among the reviews given.



```
--Number of products per rating
select reviews_rating, count(id) as noOfProducts from amazon_customer_reviews group by reviews_rating order by noOfProducts desc limit 5;

--number of catergories
```

	reviews_rating	noofproducts
1	5	17441
2	4	4781
3	NULL	3627
4	3	1018
5	1	915

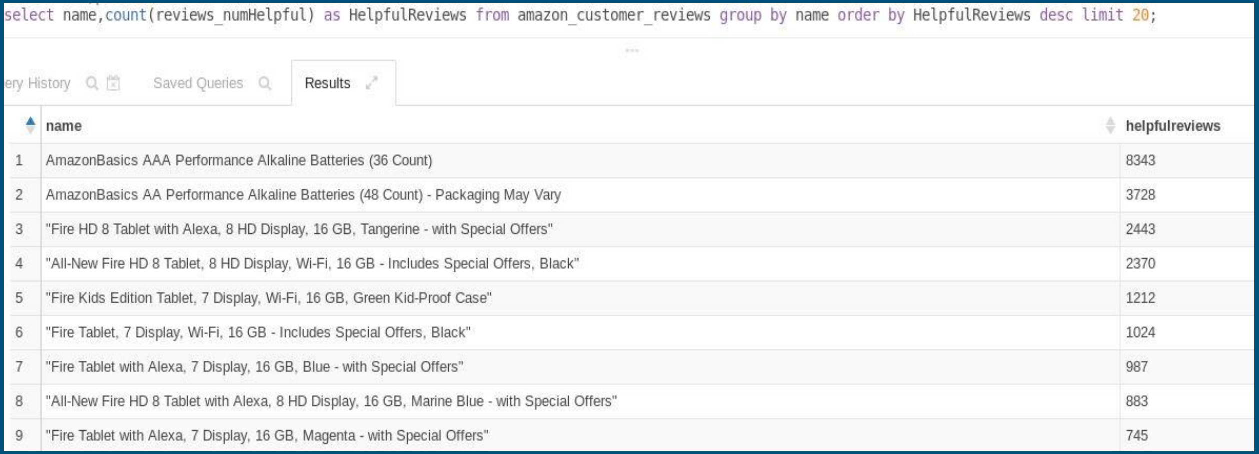


```
--Top 10 products based on average ratings
select id,avg(reviews_rating) as Rating from amazon_customer_reviews group by id order by Rating DESC limit 1000;
```

Top 1000 ratings in an order to check how ratings are distributed in the dataset and analysed that the dataset is unbalanced

Data Analysis

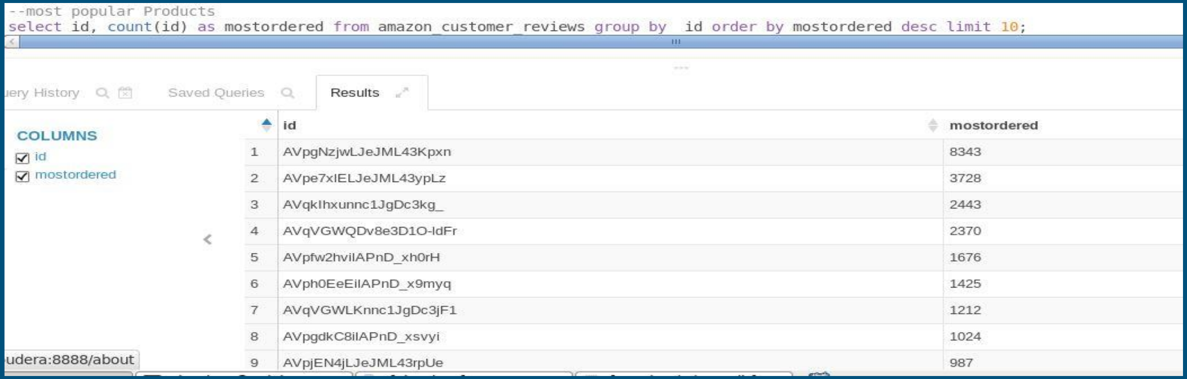
Products which has most reviews or most reviewed products in the data.



```
select name,count(reviews_numHelpful) as HelpfulReviews from amazon_customer_reviews group by name order by HelpfulReviews desc limit 20;
```

	name	helpfulreviews
1	AmazonBasics AAA Performance Alkaline Batteries (36 Count)	8343
2	AmazonBasics AA Performance Alkaline Batteries (48 Count) - Packaging May Vary	3728
3	"Fire HD 8 Tablet with Alexa, 8 HD Display, 16 GB, Tangerine - with Special Offers"	2443
4	"All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Black"	2370
5	"Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Case"	1212
6	"Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black"	1024
7	"Fire Tablet with Alexa, 7 Display, 16 GB, Blue - with Special Offers"	987
8	"All-New Fire HD 8 Tablet with Alexa, 8 HD Display, 16 GB, Marine Blue - with Special Offers"	883
9	"Fire Tablet with Alexa, 7 Display, 16 GB, Magenta - with Special Offers"	745

Most selling Amazon Products.

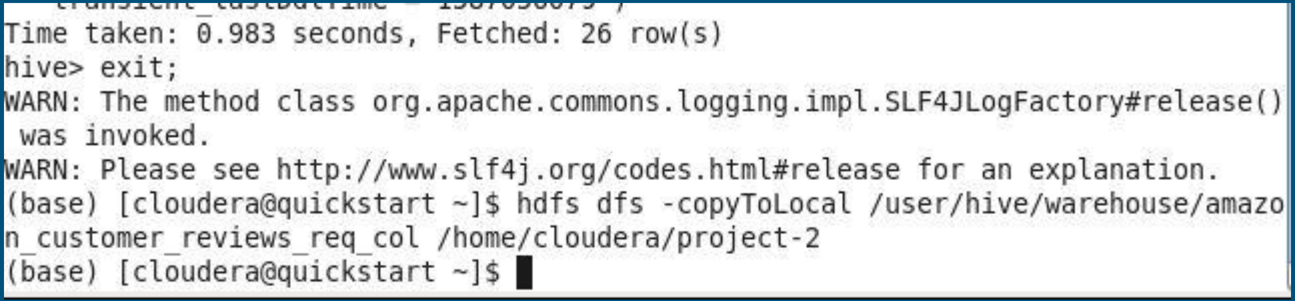


```
--most popular Products
select id, count(id) as mostordered from amazon_customer_reviews group by  id order by mostordered desc limit 10;
```

	id	mostordered
1	AVpgNzjwLJeJML43Kpxn	8343
2	AVpe7xlELJeJML43ypLz	3728
3	AVqkIhxunnc1JgDc3kg_	2443
4	AVqVGWQDv8e3D1O-ldFr	2370
5	AVpfw2hvilAPnD_xh0rH	1676
6	AVph0EeEilAPnD_x9myq	1425
7	AVqVGWLKnnc1JgDc3jF1	1212
8	AVpgdkC8ilAPnD_xsvyi	1024
9	AVpjEN4jLJeJML43rpUe	987

Data Cleaning

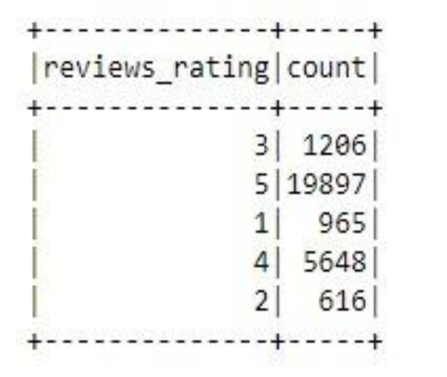
After removing the null values from the selected features.Copied the hive table to local filesystem and then loaded into spark.



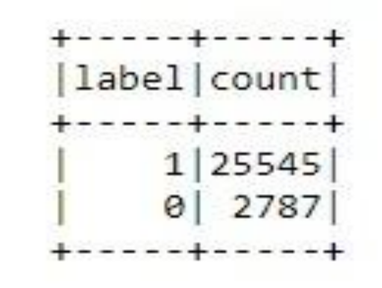
```
Time taken: 0.983 seconds, Fetched: 26 row(s)
hive> exit;
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release()
 was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
(base) [cloudera@quickstart ~]$ hdfs dfs -copyToLocal /user/hive/warehouse/amazo
n_customer_reviews_req_col /home/cloudera/project-2
(base) [cloudera@quickstart ~]$ 
```

Target Variable

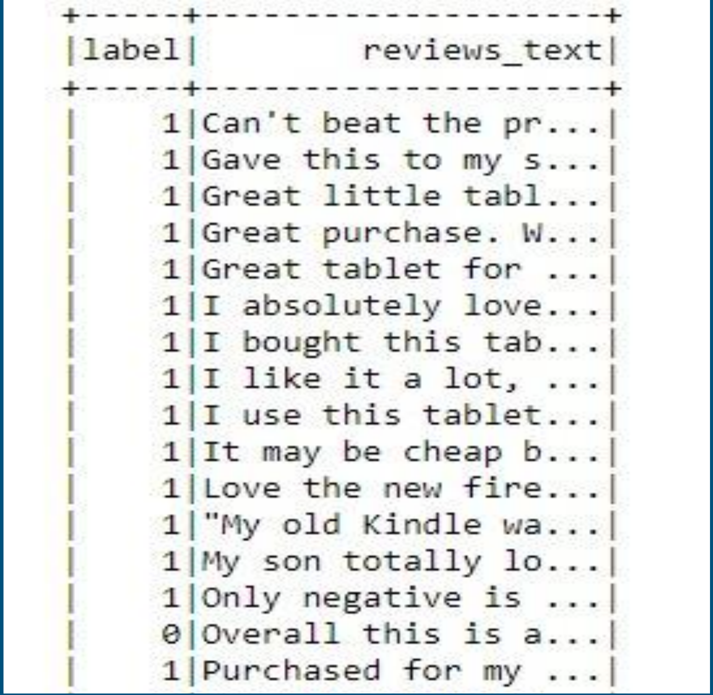
- Assigned a positive sentiment as ‘1’ for ratings >= 4 and otherwise a negative sentiment as ‘0’.



```
+--------------+-----+
|reviews_rating|count|
+--------------+-----+
|             3| 1206|
|             5|19897|
|             1|  965|
|             4| 5648|
|             2|  616|
+--------------+-----+
```



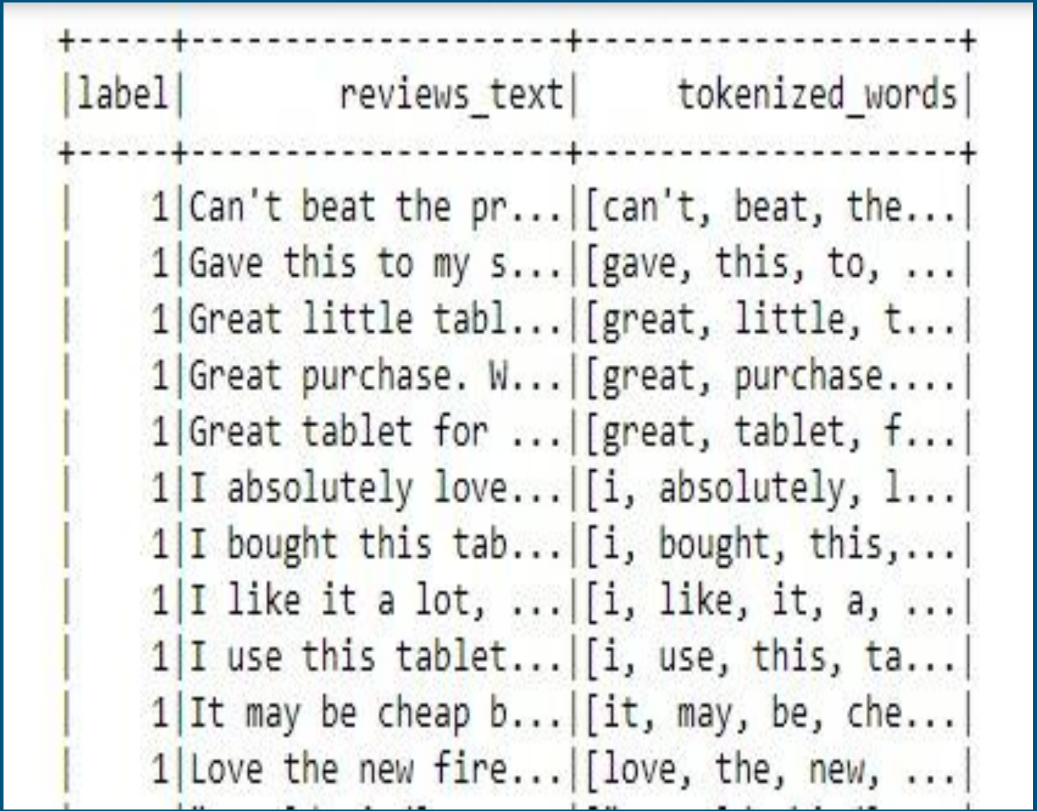
```
+-----+-----+
|label|count|
+-----+-----+
|    1|25545|
|    0| 2787|
+-----+-----+
```



```
+-----+--------------------+
|label|        reviews_text|
+-----+--------------------+
|    1|Can't beat the pr...|
|    1|Gave this to my s...|
|    1|Great little tabl...|
|    1|Great purchase. W...|
|    1|Great tablet for ...|
|    1|I absolutely love...|
|    1|I bought this tab...|
|    1|I like it a lot, ...|
|    1|I use this tablet...|
|    1|It may be cheap b...|
|    1|Love the new fire...|
|    1|"My old Kindle wa...|
|    1|My son totally lo...|
|    1|Only negative is ...|
|    0|Overall this is a...|
|    1|Purchased for my ...|
```

Tokenization

- Each review is tokenized or transformed into an ordered list of words.



```
+-----+--------------------+--------------------+
|label|        reviews_text|     tokenized_words|
+-----+--------------------+--------------------+
|    1|Can't beat the pr...|[can't, beat, the...|
|    1|Gave this to my s...|[gave, this, to, ...|
|    1|Great little tabl...|[great, little, t...|
|    1|Great purchase. W...|[great, purchase....|
|    1|Great tablet for ...|[great, tablet, f...|
|    1|I absolutely love...|[i, absolutely, l...|
|    1|I bought this tab...|[i, bought, this,...|
|    1|I like it a lot, ...|[i, like, it, a, ...|
|    1|I use this tablet...|[i, use, this, ta...|
|    1|It may be cheap b...|[it, may, be, che...|
|    1|Love the new fire...|[love, the, new, ...|
```

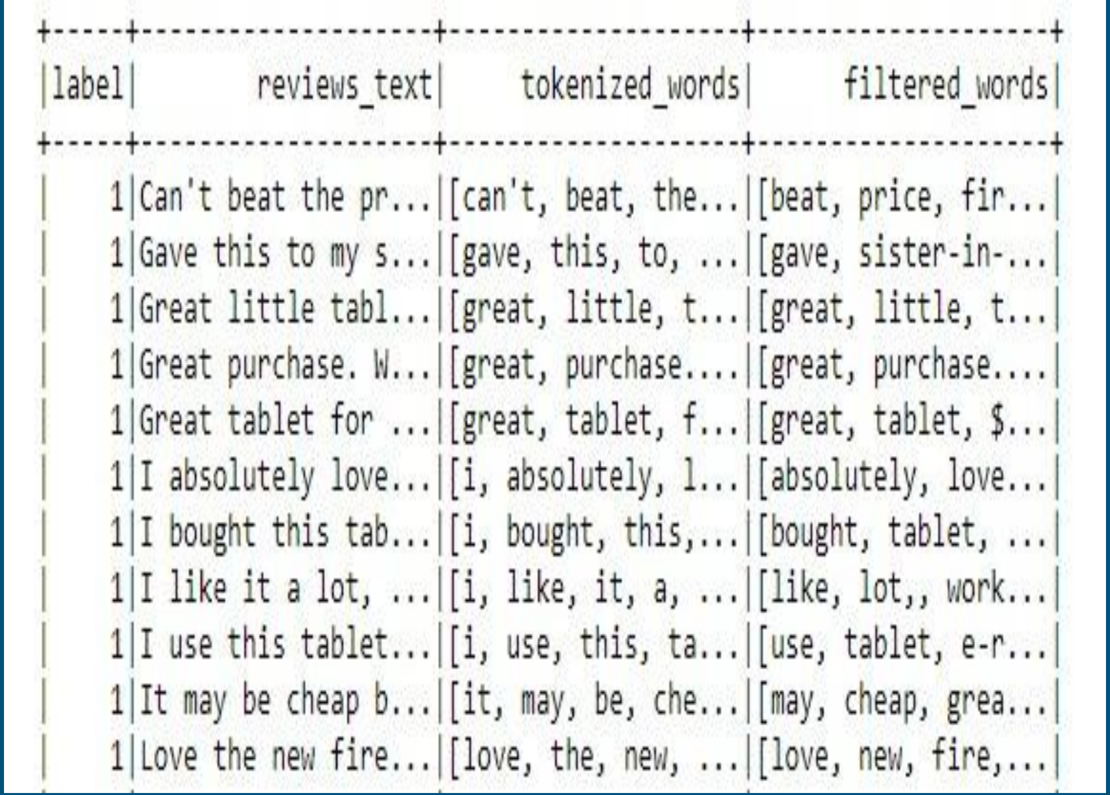
Removal Of Stop Words

- Stop words consist of most commonly used words that include:

Pronouns (us, she, their)

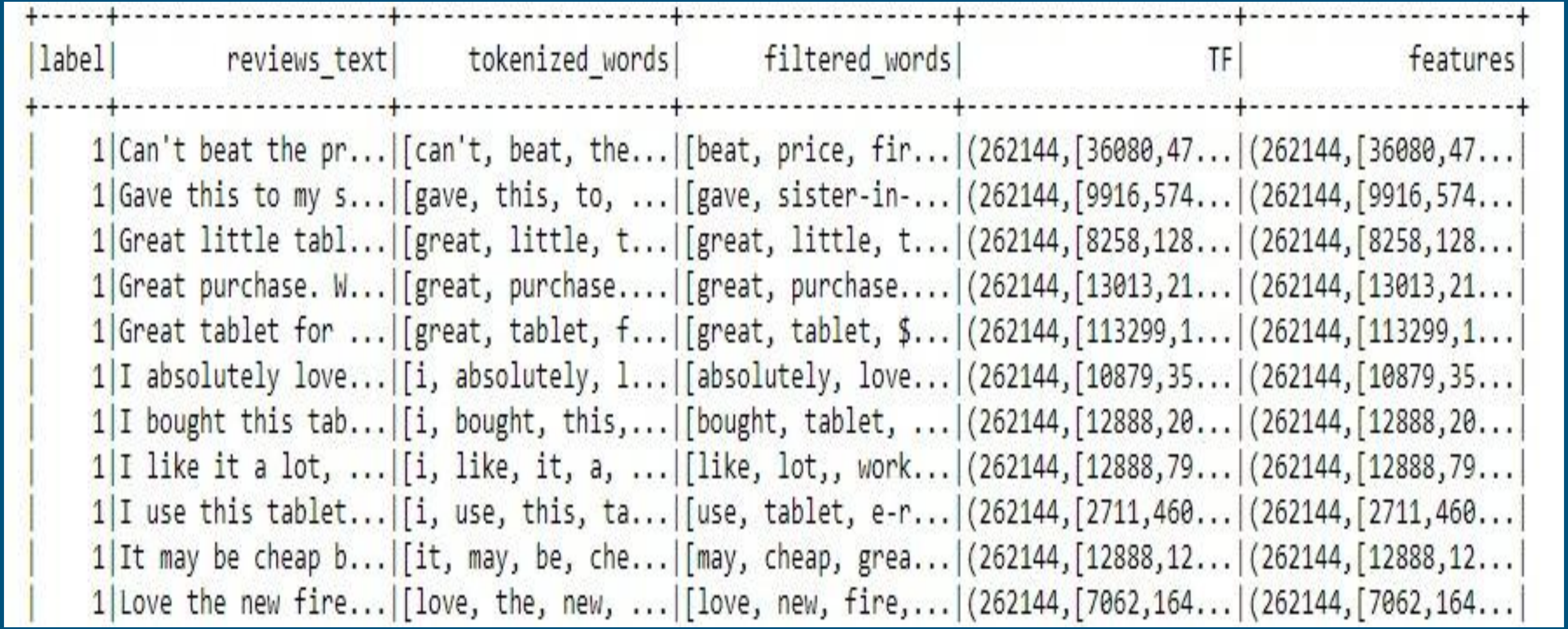
Articles (a, an, the)

Prepositions (under, from, off)



```
+-----+--------------------+--------------------+--------------------+
|label|        reviews_text|     tokenized_words|      filtered_words|
+-----+--------------------+--------------------+--------------------+
|    1|Can't beat the pr...|[can't, beat, the...|[beat, price, fir...|
|    1|Gave this to my s...|[gave, this, to, ...|[gave, sister-in-...|
|    1|Great little tabl...|[great, little, t...|[great, little, t...|
|    1|Great purchase. W...|[great, purchase....|[great, purchase....|
|    1|Great tablet for ...|[great, tablet, f...|[great, tablet, $...|
|    1|I absolutely love...|[i, absolutely, l...|[absolutely, love...|
|    1|I bought this tab...|[i, bought, this,...|[bought, tablet, ...|
|    1|I like it a lot, ...|[i, like, it, a, ...|[like, lot,, work...|
|    1|I use this tablet...|[i, use, this, ta...|[use, tablet, e-r...|
|    1|It may be cheap b...|[it, may, be, che...|[may, cheap, grea...|
|    1|Love the new fire...|[love, the, new, ...|[love, new, fire,...|
```

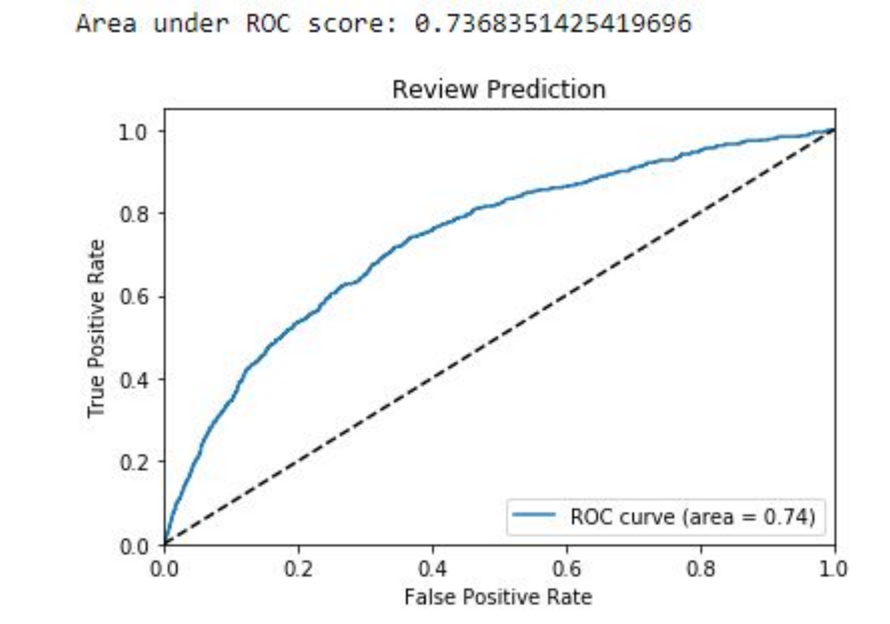
TF IDF Vectors



```
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+
|label|        reviews_text|     tokenized_words|      filtered_words|                  TF|            features|
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+
|    1|Can't beat the pr...|[can't, beat, the...|[beat, price, fir...|(262144,[36080,47...|(262144,[36080,47...|
|    1|Gave this to my s...|[gave, this, to, ...|[gave, sister-in-...|(262144,[9916,574...|(262144,[9916,574...|
|    1|Great little tabl...|[great, little, t...|[great, little, t...|(262144,[8258,128...|(262144,[8258,128...|
|    1|Great purchase. W...|[great, purchase....|[great, purchase....|(262144,[13013,21...|(262144,[13013,21...|
|    1|Great tablet for ...|[great, tablet, f...|[great, tablet, $...|(262144,[113299,1...|(262144,[113299,1...|
|    1|I absolutely love...|[i, absolutely, l...|[absolutely, love...|(262144,[10879,35...|(262144,[10879,35...|
|    1|I bought this tab...|[i, bought, this,...|[bought, tablet, ...|(262144,[12888,20...|(262144,[12888,20...|
|    1|I like it a lot, ...|[i, like, it, a, ...|[like, lot,, work...|(262144,[12888,79...|(262144,[12888,79...|
|    1|I use this tablet...|[i, use, this, ta...|[use, tablet, e-r...|(262144,[2711,460...|(262144,[2711,460...|
|    1|It may be cheap b...|[it, may, be, che...|[may, cheap, grea...|(262144,[12888,12...|(262144,[12888,12...|
|    1|Love the new fire...|[love, the, new, ...|[love, new, fire,...|(262144,[7062,164...|(262144,[7062,164...|
```

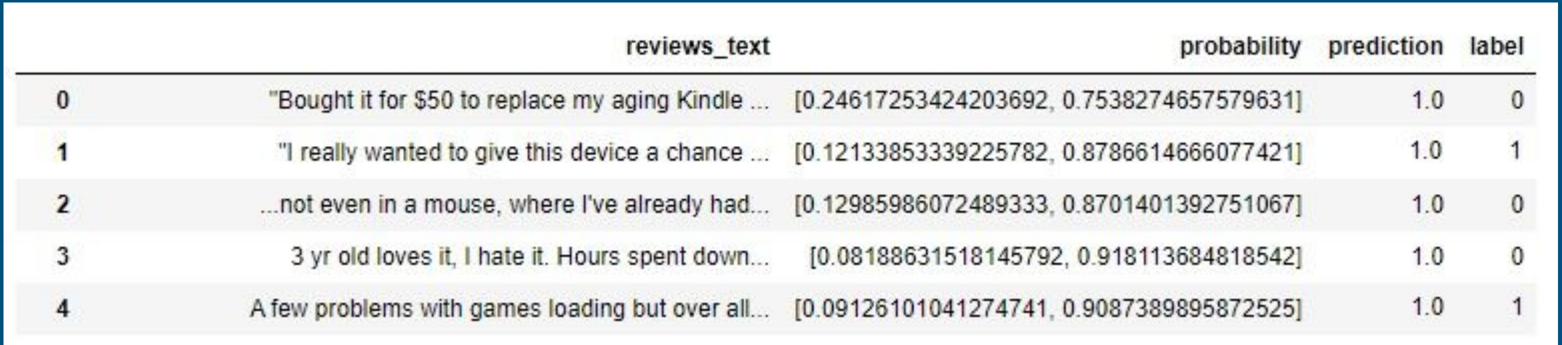
Logistic Regression Model

- Pipeline(stages=[tokenizer, remover, hashingTF, idfModel, lr])
- CrossValidator with estimator as pipeline and evaluator as BinaryClassificationEvaluator



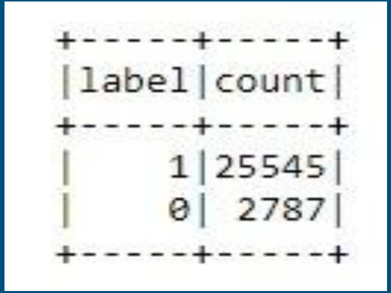
```
Area under ROC score: 0.7368351425419696
```

Logistic Regression Model



	reviews_text	probability	prediction	label
0	"Bought it for $50 to replace my aging Kindle ...	[0.24617253424203692, 0.7538274657579631]	1.0	0
1	"I really wanted to give this device a chance ...	[0.12133853339225782, 0.8786614666077421]	1.0	1
2	...not even in a mouse, where I've already had...	[0.12985986072489333, 0.8701401392751067]	1.0	0
3	3 yr old loves it, I hate it. Hours spent down...	[0.08188631518145792, 0.918113684818542]	1.0	0
4	A few problems with games loading but over all...	[0.09126101041274741, 0.9087389895872525]	1.0	1

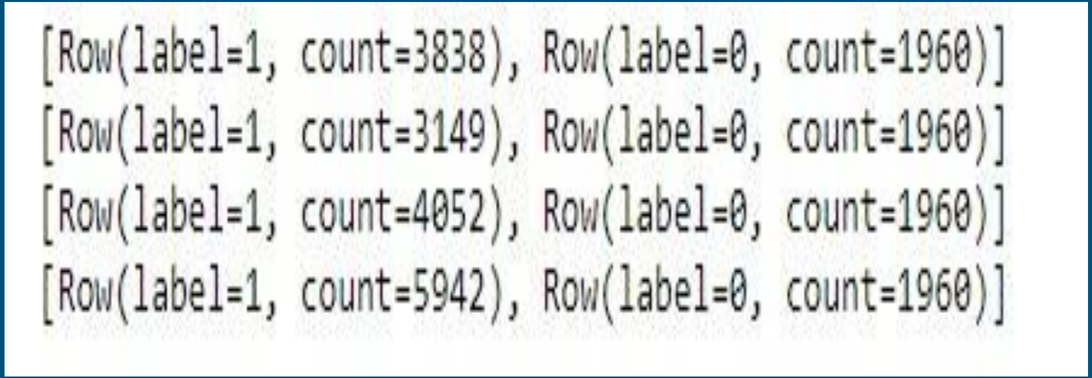
- The output is more biased to the positive reviews.
- Unbalanced dataset with 90% positive reviews and 10% negative reviews.



```
+-----+-----+
|label|count|
+-----+-----+
|    1|25545|
|    0| 2787|
+-----+-----+
```

Down-Sampling

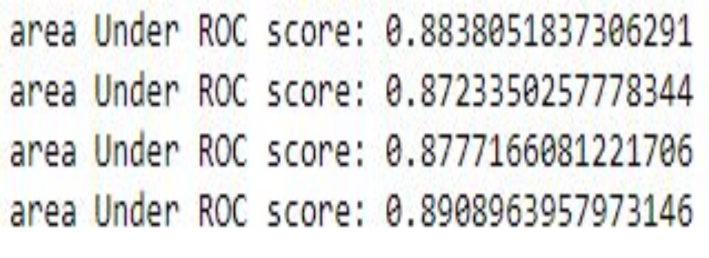
- Down sampling
- Ensemble of down samplings



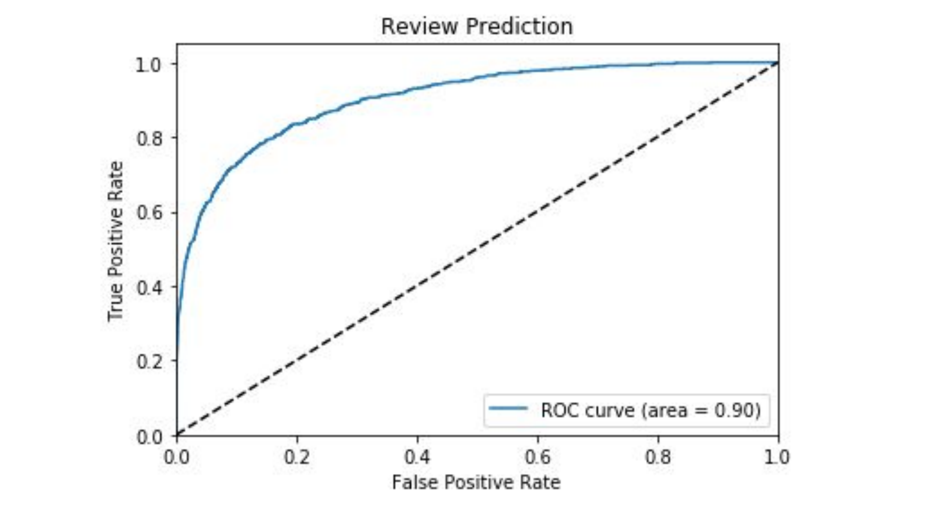
```
[Row(label=1, count=3838), Row(label=0, count=1960)]
[Row(label=1, count=3149), Row(label=0, count=1960)]
[Row(label=1, count=4052), Row(label=0, count=1960)]
[Row(label=1, count=5942), Row(label=0, count=1960)]
```

Random Forest Classifier

- Pipeline(stages=[tokenizer, remover, hashingTF, idfModel, rf])



```
area Under ROC score: 0.8838051837306291
area Under ROC score: 0.8723350257778344
area Under ROC score: 0.877166081221706
area Under ROC score: 0.8908963957973146
```



Conclusion

- Various NLP pre processing techniques and concepts were explored during this project.

Limitations

- Handling of incorrectly spelled words
- Text cannot be interpreted by underlying context

THANK YOU