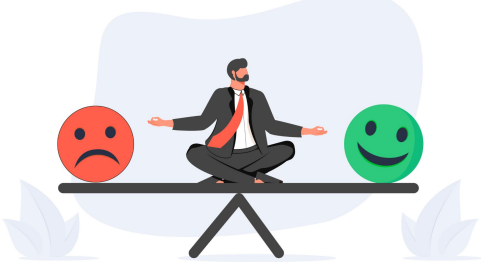
Book Review Sentiment Analysis

Binary Classification Task

Grishma Bhattarai
Spring 2022

MACHINE LEARNING TASK



Context: This project is a sentiment analysis ML project on Amazon Kindle book reviews. Individuals can use results from this analysis to further enrich their book collection with books that the majority love. If public sentiment towards a book is not so good, they may also decide to avoid such books. Additionally, local bookstores and vendors can use such analysis to understand general demand and offer good products. Thus, sentiment analysis can help people in their decision-making process.

Task: To train a **Binary Classification Model** to predict whether an Amazon Kindle book review is positive or non-positive based on the contents of the review. Each row in the data is an instance of a person's review related to a book.

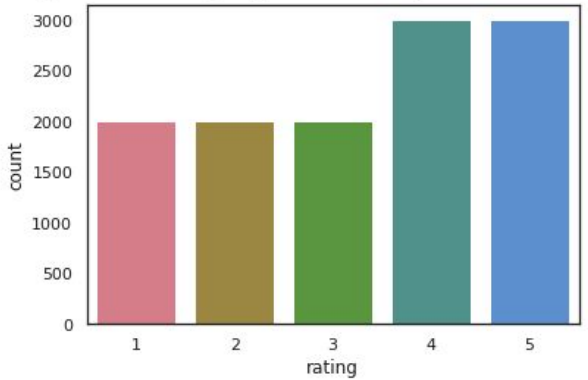
Data

The source for these reviews is a massive data repository that contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

	Unnamed: 0	rating	reviewText	summary
0	0	5	This book was the very first bookmobile book I bought when I was in the school book club. I loved the story then and I bet a dollar to a donut I will love it again. If my memory serves, I bought this book in 5th grade. That would have been about 1961. I am looking forward to reliving the memories.	50 + years ago...
1	1	1	When I read the description for this book, I couldn't wait to read it. Once I downloaded it to my Kindle, I found it extremely hard to keep reading it. To be honest, I stopped reading halfway through the book. It began slow and remained a slow, uninteresting read. It lacked passion; not making love passion, but passion for life. Neither Jada or Aaron were interesting characters and the story was too, too `everything is perfect'. Everybody is just so understanding and accommodating--the bit of drama with his father and her grandmother was blah. To give an example of what I mean (and there are MANY), Aaron finally finds out almost half way into the book that he fathered a child 5 years ago--keep in mind that Aaron and Jada kept in contact for a few months after he left and she never mentioned to him that she was pregnant. When she finally tells him he has a son, Aaron becomes overwhelmed with emotions (misty eyed) knowing that he's a father. WTH! I think most men would be upset/angry to know that they had a child and the woman they love never told them! Not in this book; it's all good--all is well; NO PASSION. OMG, don't let me get started on the scene when father and son finally meet. It was so over the top, it made me want to barf. Maybe if this scene had been towards the beginning of the book, it would have been touching. In the middle of the book after just too much blah it made me thrown down my kindle. It was like eating a sweet dessert that was so sweet it made your teeth hurt. This book had a good theme, but no follow-through. Also, if we're to believe that Jada graduated from Harvard, then let it be reflected in her speech and her job. We're told that she has this wonderful personality which caused Aaron to fall in love with her--never saw it. If the author had described her as depressed and foolish--Yes! I would believe it in a heartbeat. She was boring. This book was so frustrating to read. I truly struggled to read it as far as I did. I finally stopped torturing myself because I realized it wasn't going to get any better.	Boring! Boring! Boring!

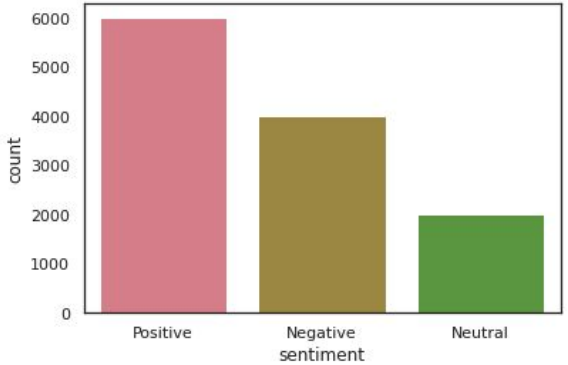
```
((12000, 4), None)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f7da502ec10>

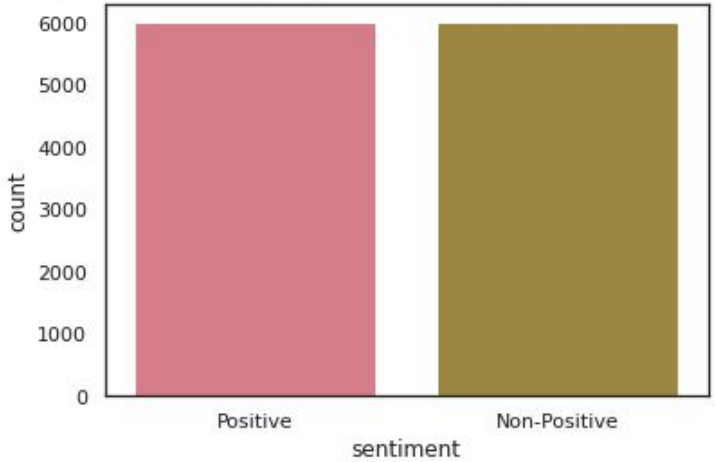


- Positive: Rating > 3
- Neutral: Rating == 3
- Negative: Rating < 3

<matplotlib.axes._subplots.AxesSubplot at 0x7f7da502eed0>

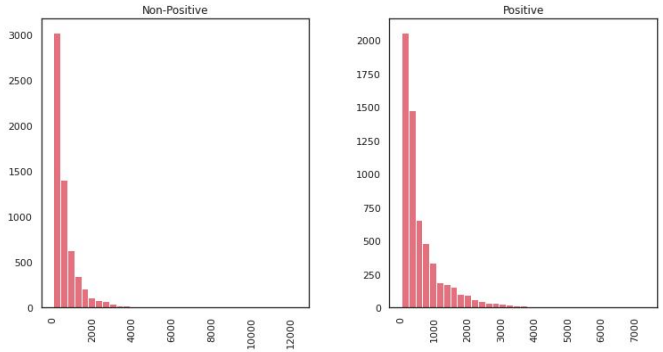


<matplotlib.axes._subplots.AxesSubplot at 0x7f7da8a09150>



Preprocessing Steps

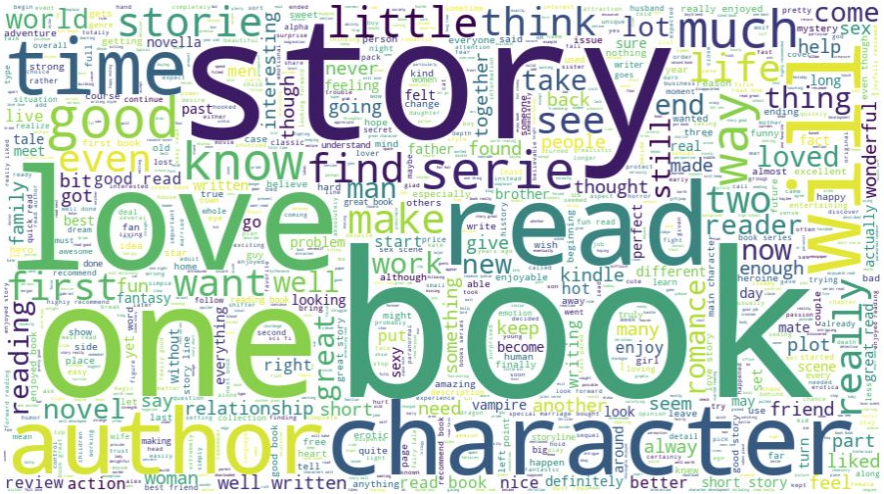
- Transformed ordinal ranking [1-5] to balanced Sentiment Labels [Non-Positive, Positive)
- Length Distribution Analysis: Similar Distributions



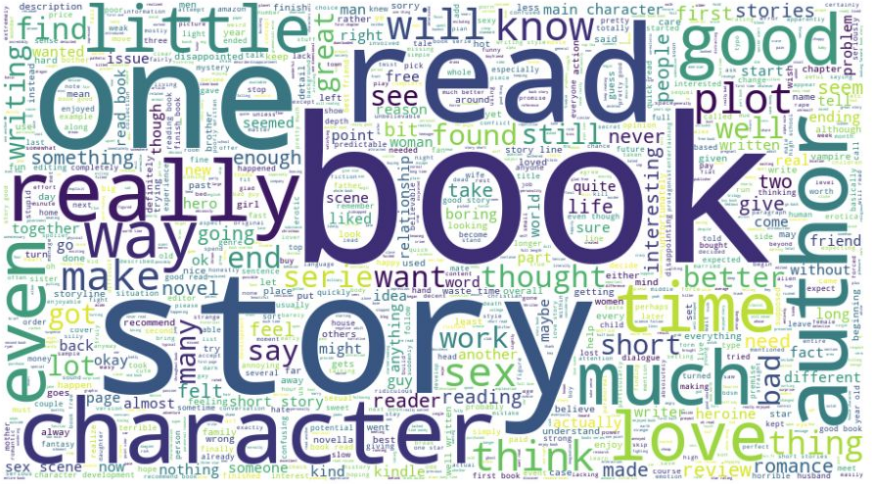
- Deconstruction & Cleaning: Removing punctuations, URLs, StopWords, Tokenization, Stemming
- Train Test Split: 70:30 ratio

```
("Where's the meat? Had to read certain passages twice--typos.
 'Where is the meat  Had to read certain passages twice  typos
```

Word Clouds

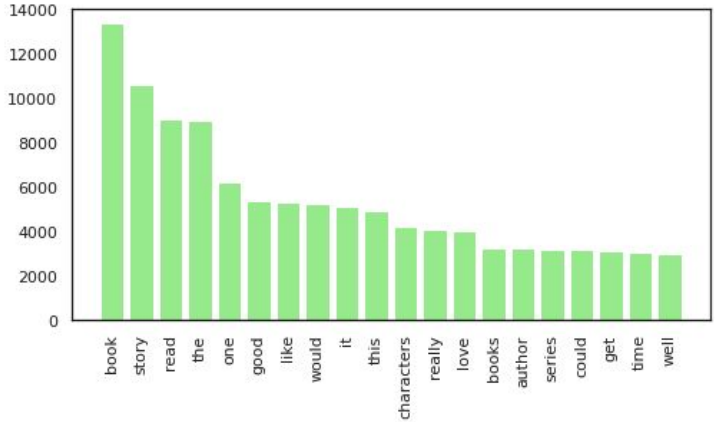


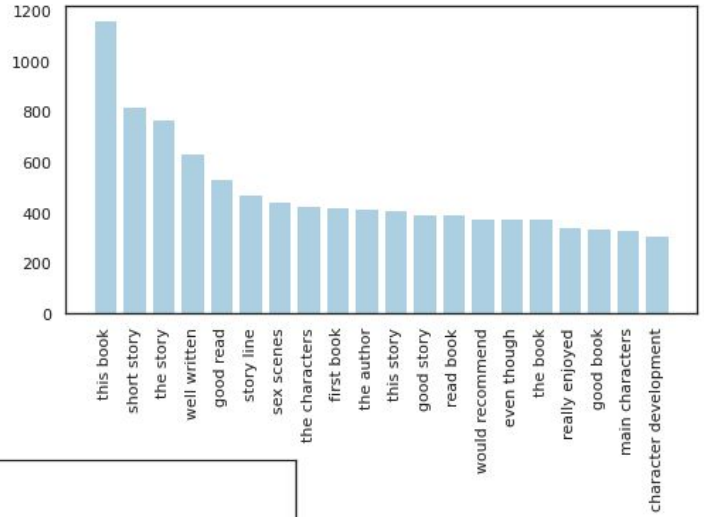
Positive Class

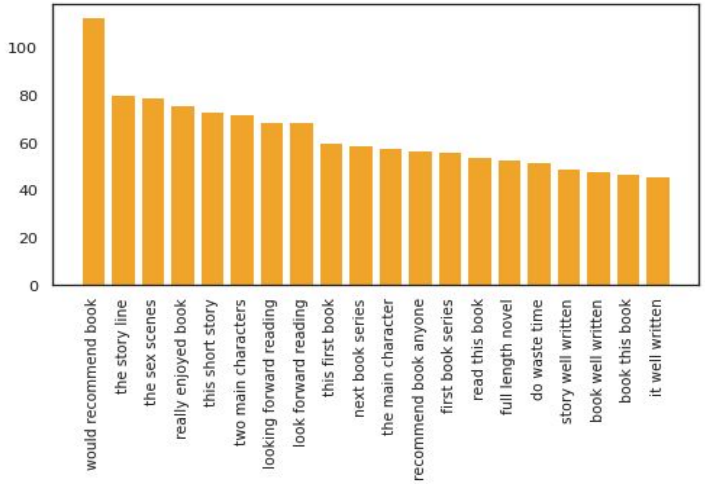


Non-Positive Class

N-gram Distributions





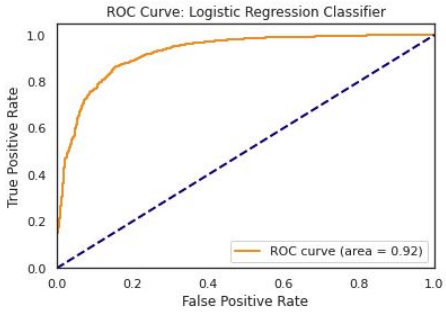


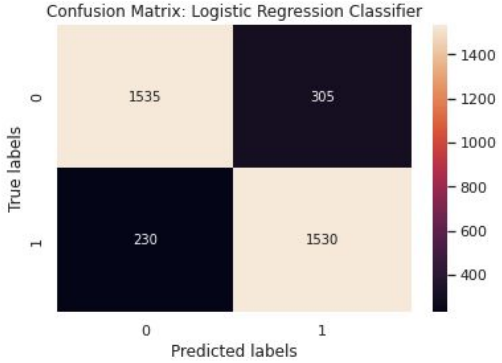
Logistic Regression Classifier Model

- CountVectorizer (lowercase=True, max_df=0.8, min_df=3, ngram_range=(1,2))
- TfidfTransformer()
- LogisticRegression(penalty='l2')

	precision	recall	f1-score	support
0	0.87	0.83	0.85	1840
1	0.83	0.87	0.85	1760
accuracy			0.85	3600
macro avg	0.85	0.85	0.85	3600
weighted avg	0.85	0.85	0.85	3600

Top Positive Features:	Top Non-Positive Features:
('loved', 5.982),	('okay', -3.945),
('great', 5.835),	('ok', -3.878),
('enjoyed', 4.806),	('boring', -3.534),
('love', 3.743),	('bad', -3.465),
('wonderful', 3.594),	('nothing', -3.408),
('hot', 3.243),	('waste', -3.067),
('well', 3.214),	('sex', -3.059),
('excellent', 2.915),	('finish', -2.839),
('good', 2.632),	('free', -2.834),
('read', 2.599)]	('sorry', -2.795)]



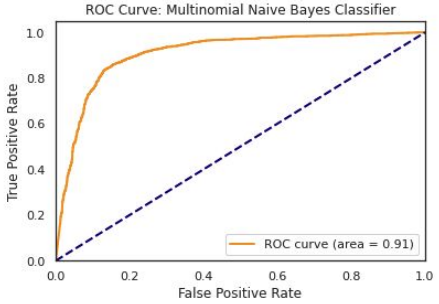


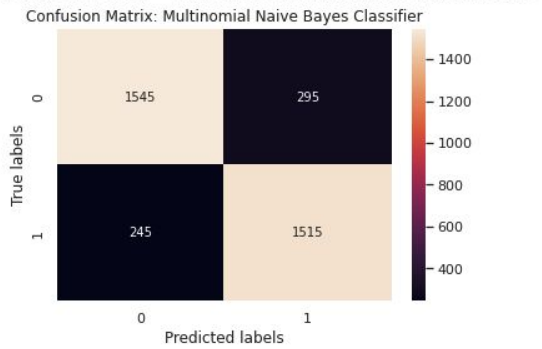
	cleaned_review	label	pred_label
6494	Not well known today readers old character When I child parents loved stories made several movies course I thought boring old fashioned Well still old fashioned comforting old world way The stories set British Isles part WWI WWII Bulldog Drummond makes habit saving damsel distress flair dash There bad language course innuendo explicit sex violence interesting narrative story lines Great quiet rainy Sunday	0	1
9290	Liked quick read I going three stars though things pulled four star range times I LOVED first half book I loved get see like little snippets years together gradual progression things I liked Adam I mean sweetie cooks come rough existence pathetic parents I felt He showed strength loyalty good work ethic I would liked little interaction Rhone I enjoyed inner dialogue get Rhone good character though times I wanted like Dude Can really see There oh God super like embarrassing moments Overall I enjoyed story I read Dane work I read	0	1
7238	BOOK PLEASE I actually started reading book least times actually reading cover cover one day I might add In fact I decided start reading nothing else seemed interesting Call reader block On rare ocasions I actually force feed eyes brain I held hope book would break current block I thought What heck read page throughout day finally something hit senses hours later I looking sequel No spoilers Touching love story two people thought set lives futures seeing first time Over extremely short period time come know therefore wanted remember As one learns strength rediscovers hope living tragedy looms head Ms McCay details main characters past discover possible futures little humor thrown pleasure My regret finding sequel I left imagination help decide happens Michael Regina etc This book dated I think time find sacrifices made worth give tale happy ending	1	0

Multinomial NB Classifier Model

- CountVectorizer (lowercase=True, max_df=0.8, min_df=1, ngram_range=(1,2))
- MultinomialNB(penalty='l2')

	precision	recall	f1-score	support
0	0.86	0.84	0.85	1840
1	0.84	0.86	0.85	1760
accuracy			0.85	3600
macro avg	0.85	0.85	0.85	3600
weighted avg	0.85	0.85	0.85	3600





Top Positive Features:
[('book', -5.411),
('read', -5.533),
('story', -5.552),
('one', -6.048),
('good', -6.096),
('love', -6.21),
('great', -6.392),
('would', -6.412),
('series', -6.424),
('characters', -6.434)]

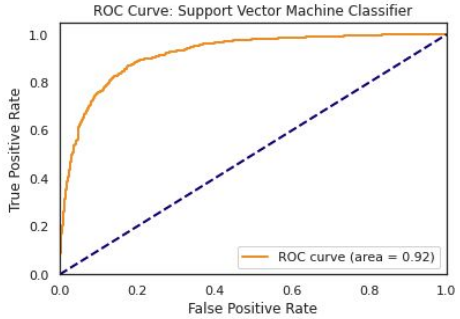
Top Non-Positive Features:
[('aa', -13.743),
('aa done', -13.743),
('aaaaaannnnd', -13.743),
('aaaaaannnnd nothing', -13.743),
('aaboulet', -13.743),
('aand', -13.743),
('aand nate', -13.743),
('aarggg', -13.743),
('aaron', -13.743),
('aaron also', -13.743)]

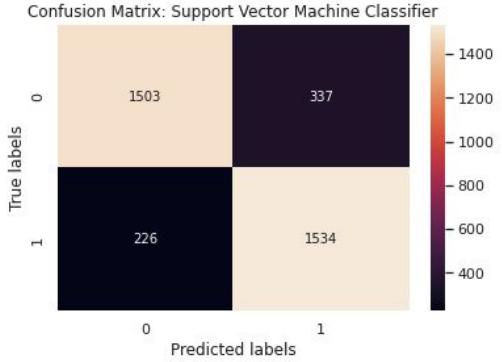
	cleaned_review	label	pred_label
6494	Not well known today readers old character When I child parents loved stories made several movies course I thought boring old fashioned Well still old fashioned comforting old world way The stories set British Isles part WWI WWII Bulldog Drummond makes habit saving damsel distress flair dash There bad language course innuendo explicit sex violence interesting narrative story lines Great quiet rainy Sunday	0	1
4617	Great sci fi story I love book I even halfway thru I ready buy rest series It hard sci fi lot well researched mathematics enough keep hard sci fi fan happy To previous reviewer commented global warming real U S bad guy I get book Yes mentioned global warming turned real U S spend lots money get control I think author preaching either subject It simply small part storyline In fact far mentioned I recommend book sci fi fan	1	0
7238	BOOK PLEASE I actually started reading book least times actually reading cover cover one day I might add In fact I decided start reading nothing else seemed interesting Call reader block On rare ocasions I actually force feed eyes brain I held hope book would break current block I thought What heck read page throughout day finally something hit senses hours later I looking sequel No spoilers Touching love story two people thought set lives futures seeing first time Over extremely short period time come know therefore wanted remember As one learns strength rediscovers hope living tragedy looms head Ms McCay details main characters past discover possible futures little humor thrown pleasure My regret finding sequel I left imagination help decide happens Michael Regina etc This book dated I think time find sacrifices made worth give tale happy ending	1	0

Support Vector Machine Classifier Model

- CountVectorizer (lowercase=True, max_df=0.8, min_df=5, ngram_range=(1,2))
- TfidfTransformer()
- SGDClassifier(loss='hinge', penalty='l2', alpha=1e-3, random_state=42, tol=None)

	precision	recall	f1-score	support
0	0.87	0.82	0.84	1840
1	0.82	0.87	0.84	1760
accuracy			0.84	3600
macro avg	0.84	0.84	0.84	3600
weighted avg	0.85	0.84	0.84	3600





Top Positive Features:

('great', 2.947),	('loved', 2.721),
('enjoyed', 2.448),	('love', 2.158),
('well', 1.651),	('hot', 1.641),
('wonderful', 1.625),	('read', 1.575),
('series', 1.531),	('good', 1.477)]

Top Non-Positive Features:

[('ok', -2.098),	('bad', -1.97),
('okay', -1.819),	('sex', -1.653),
('nothing', -1.613),	('boring', -1.611),
('like', -1.564),	('free', -1.533),
('book', -1.42),	('plot', -1.415)]

	cleaned_review	label	pred_label
6494	Not well known today readers old character When I child parents loved stories made several movies course I thought boring old fashioned Well still old fashioned comforting old world way The stories set British Isles part WWI WWII Bulldog Drummond makes habit saving damsel distress flair dash There bad language course innuendo explicit sex violence interesting narrative story lines Great quiet rainy Sunday	0	1
8128	The Heat Knight The author listed erotica however middle road erotica Not light erotica heavy erotica Basically lord family butlers daughter Christiana They liked younger However something happens never acknowledge fell Years later father dies Lord Beckett makes come work house protect Beckett still love wants forget happen years ago Christiana want work castle Some people mean wants get away feeling Beckett womanizing ways Christiana seems give easily becoming Beckett mistress decent short story A big misunderstanding almost costs happiness long couple things But always things work end Short story short really develop great story lot say Will come whether like erotic	1	0
9290	Liked quick read I going three stars though things pulled four star range times I LOVED first half book I loved get see like little snippets years together gradual progression things I liked Adam I mean sweetie cooks come rough existence pathetic parents I felt He showed strength loyalty good work ethic I would liked little interaction Rhone I enjoyed inner dialogue get Rhone good character though times I wanted like Dude Can really see There oh God super like embarrassing moments Overall I enjoyed story I read Dane work I read	0	1

Summary & Conclusion

With the above project, we built 3 binary sentiment classifier models for Amazon Kindle Book Reviews Dataset. The 3 models used were a Logistic Regression Classifier, a Multinomial NB Classifier and a Support Vector Classifier. The following was the pipeline of our project:

- **Data Preprocessing**: We first explored and cleaned the data. We also balanced our data and explored word clouds and n-gram plots.
- **Model Selection and Parameter Tuning**: We then built 3 pipelines and used GridSearchCV to explore the best parameter settings for CountVectorizer the respective model .
- **Model Training and Evaluation**: We then split the data into 70% training and 30% testing sets. We transformed our data using CountVectorizer and Tfidf Transformer. Then, we trained the 3 respective classifier models on the training data (with the best parameter setting). Finally, we we evaluated our 3 modes" performance on the testing data using classification report, ROC curve, confusion matrix.
- **Model Explanation and Error Analysis**: Finally, we checked the top coefficient words to explain the model performance and conducted an error analysis by explaining samples where the model(s) made errors.

	Model Type	F1-Score	AUC Metric
1	Logistic Regression Classifier	0: 0.85, 1: 0.85	0.92
2	Multinomial Naive Bayes Classifier	0: 0.85, 1: 0.85	0.91
3	Support Vector Machine Classifier	0: 0.84 , 1: 0.84	0.92