Predicting Customer's Feedback using Random Forest

Problem Statement

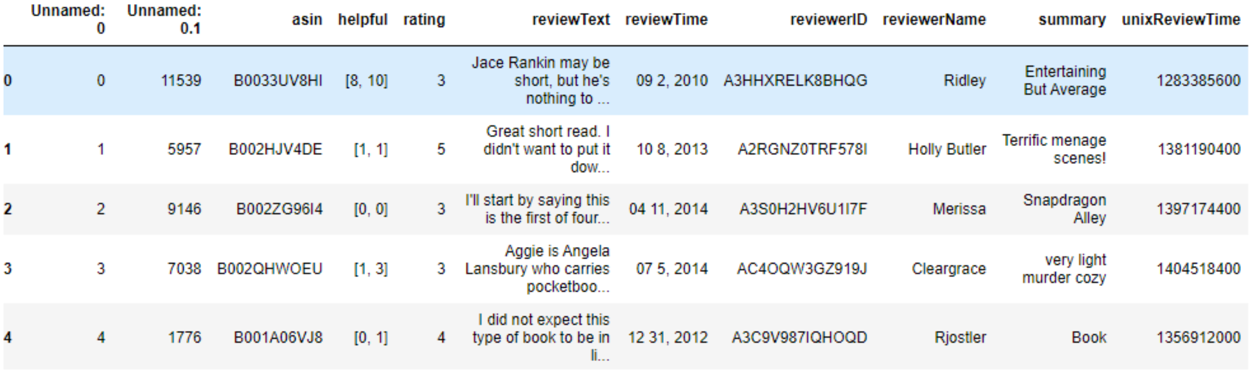
Building a system that helps Cairo Book Fair visitors to know more about the customers' feedback on the books by using Kindle reviews data to train a model that can detect the sentiment based on the review given to the book.

We will be using Random Forest Algorithm for this task to classify whether the review is negative or positive.

Data Wrangling

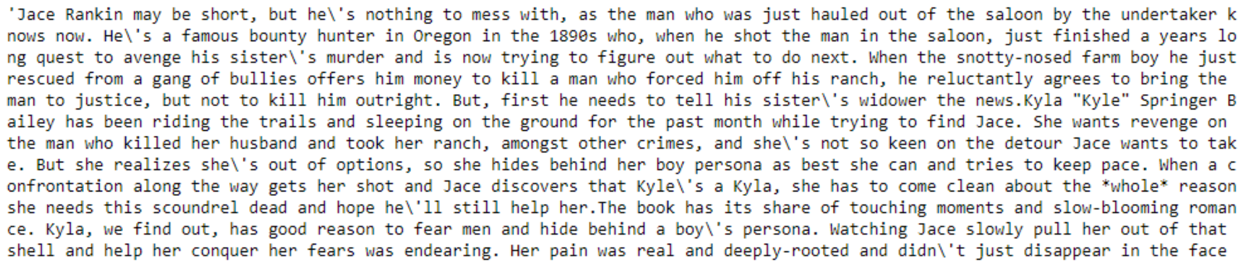
We have done some of data wrangling steps in order to make the data cleansed and ready to be fitted in our algorithm to assure best results

Here are samples of our data before vs after wrangling:



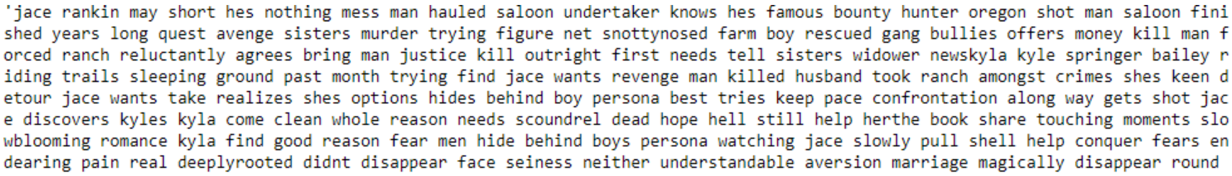
	Unnamed: 0	Unnamed: 0.1	asin	helpful	rating	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	0	11539	B0033UV8HI	[8, 10]	3	Jace Rankin may be short, but he's nothing to ...	09 2, 2010	A3HHXRELK8BHQG	Ridley	Entertaining But Average	1283385600
1	1	5957	B002HJV4DE	[1, 1]	5	Great short read. I didn't want to put it dow...	10 8, 2013	A2RGNZ0TRF578I	Holly Butler	Terrific menage scenes!	1381190400
2	2	9146	B002ZG96I4	[0, 0]	3	I'll start by saying this is the first of four...	04 11, 2014	A3S0H2HV6U1I7F	Merissa	Snapdragon Alley	1397174400
3	3	7038	B002QHWOEU	[1, 3]	3	Aggie is Angela Lansbury who carries pocketboo...	07 5, 2014	AC4OQW3GZ919J	Cleargrace	very light murder cozy	1404518400
4	4	1776	B001A06VJ8	[0, 1]	4	I did not expect this type of book to be in li...	12 31, 2012	A3C9V987IQHOQD	Rjostler	Book	1356912000

Sample text before cleansing



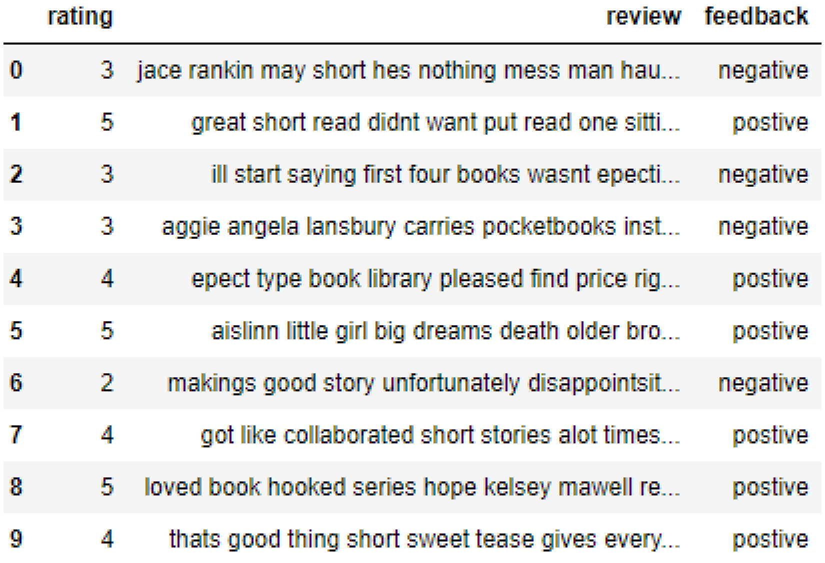
```
'Jace Rankin may be short, but he\'s nothing to mess with, as the man who was just hauled out of the saloon by the undertaker k
nows now. He\'s a famous bounty hunter in Oregon in the 1890s who, when he shot the man in the saloon, just finished a years lo
ng quest to avenge his sister\'s murder and is now trying to figure out what to do next. When the snotty-nosed farm boy he just
rescued from a gang of bullies offers him money to kill a man who forced him off his ranch, he reluctantly agrees to bring the
man to justice, but not to kill him outright. But, first he needs to tell his sister\'s widower the news.Kyla "Kyle" Springer B
ailey has been riding the trails and sleeping on the ground for the past month while trying to find Jace. She wants revenge on
the man who killed her husband and took her ranch, amongst other crimes, and she\'s not so keen on the detour Jace wants to tak
e. But she realizes she\'s out of options, so she hides behind her boy persona as best she can and tries to keep pace. When a c
onfrontation along the way gets her shot and Jace discovers that Kyle\'s a Kyla, she has to come clean about the *whole* reason
she needs this scoundrel dead and hope he\'ll still help her.The book has its share of touching moments and slow-blooming roman
ce. Kyla, we find out, has good reason to fear men and hide behind a boy\'s persona. Watching Jace slowly pull her out of that
shell and help her conquer her fears was endearing. Her pain was real and deeply-rooted and didn\'t just disappear in the face
```

Sample text after cleansing



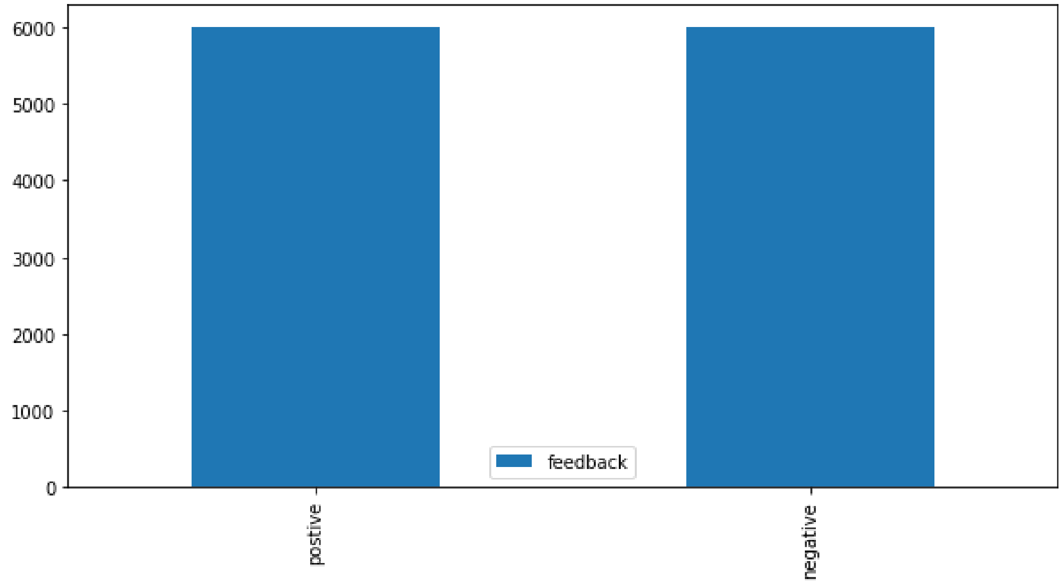
```
'jace rankin may short hes nothing mess man hauled saloon undertaker knows hes famous bounty hunter oregon shot man saloon fini
shed years long quest avenge sisters murder trying figure net snottynosed farm boy rescued gang bullies offers money kill man f
orced ranch reluctantly agrees bring man justice kill outright first needs tell sisters widower newskyla kyle springer bailey r
iding trails sleeping ground past month trying find jace wants revenge man killed husband took ranch amongst crimes shes keen d
etour jace wants take realizes shes options hides behind boy persona best tries keep pace confrontation along way gets shot jac
e discovers kyles kyla come clean whole reason needs scoundrel dead hope hell still help herthe book share touching moments slo
wblooming romance kyla find good reason fear men hide behind boys persona watching jace slowly pull shell help conquer fears en
dearing pain real deeplyrooted didnt disappear face seiness neither understandable aversion marriage magically disappear round
```

Here is our data frame after dropping the unnecessary columns and creating a new column “Feedback” that classifies the rate more than 3 is positive and below than 4 is negative to be our criteria.



	rating	review	feedback
0	3	jace rankin may short hes nothing mess man hau...	negative
1	5	great short read didnt want put read one sitti...	postive
2	3	ill start saying first four books wasnt epecti...	negative
3	3	aggie angela lansbury carries pocketbooks inst...	negative
4	4	epect type book library pleased find price rig...	postive
5	5	aislinn little girl big dreams death older bro...	postive
6	2	makings good story unfortunately disappointsit...	negative
7	4	got like collaborated short stories alot times...	postive
8	5	loved book hooked series hope kelsey mawell re...	postive
9	4	thats good thing short sweet tease gives every...	postive

Coincidently that the data is balanced and the samples of negative reviews equal to the positive ones



Splitting Data & Data modeling

In this stage we have used TF-IDF to count the frequency of the term in the data inverse the document frequency. Then we trained the Random Forest Algorithm on the data and tested the accuracy of the model on the test date and it has an accuracy 83%

So the model correctly predicted 83% of the instances that was tested on

Limitations

We can increase the accuracy of the model by:

1- Collecting more date to train our model on

2- Using Neural Network Algorithm such as “LSTM” can deliver better results being trained and tested on text data