COMP 6950 - Final Project

Name - **Mohammad Shehabul Islam**
ID - **202196528**

Review Analysis of Amazon using Text Classification

Objective: To analyze the feedback and classify it as Positive and Negative review.

Dataset Overview:

	reviewerID	asin	reviewerName	helpful/0	helpful/1	reviewText	overall	summary	unixReviewTime	reviewTime
0	A1F2H80A1ZNN1N	B00GDM3NQC	Connie Correll	0	0	I bought both boxed sets, books 1-5. Really a...	5	Can't stop reading!	1390435200	01 23, 2014
1	AI3DRTKCSK4KX	B00A5MREAM	Grandma	0	0	I enjoyed this short book. But it was way way ...	3	A leaf on the wind of all hallows	1399593600	05 9, 2014
2	A3KAKFHY9DAC8A	0446547573	toobusyreading "Inspired Kathy"	1	1	I love Nicholas Sparks. I’ve read everyt...	4	Great writing from Nicholas Sparks.	1404518400	07 5, 2014

We are focused on the feedback of the customers that bought the books. Therefore we drop unnecessary columns and keep only the ‘reviewText’ and ‘overall’ column.

Data Preprocessing: After removing the columns, our dataframe looks like this:

Out[5]:

	feedback	rating	label
0	I bought both boxed sets, books 1-5. Really a...	5	Positive
1	I enjoyed this short book. But it was way way ...	3	Neutral
2	I love Nicholas Sparks. I’ve read everyt...	4	Positive
3	I really enjoyed this adventure and look forwa...	4	Positive
4	It was a decent read.. typical story line. Not...	3	Neutral
...	...	...	...
9995	The whole series was great! Melody is a fanta...	5	Positive
9996	I didn't thing that much of this book. I am a...	3	Neutral
9997	It is an emotional TRIP to the past with Trip ...	5	Positive
9998	This definitely got under my veins whereby I h...	5	Positive
9999	Highly recommend this entire trilogy. It is ve...	4	Positive

10000 rows × 3 columns

Since these were all user feedback data, it was not clean data and for that I further preprocessed it:

1. Calculate text length
2. Calculate punctuation %
3. Remove punctuation
4. Tokenize the text
5. Removed Stop Words
6. Used Stemming and Lemmatization
7. Tf-Idf Vectorizer

Libraries:

- ➢ Numpy
- ➢ Pandas
- ➢ Matplotlib
- ➢ Scipy
- ➢ Scikit learn
- ➢ Nltk

Out[48]:

	feedback	rating	label	text_length	punc_%	clean_text	tokenized_clean_text	text_no_stopword	ps_stem	wn_lemmatize	processed_text
0	I bought both boxed sets, books 1-5. Really a...	5	Positive	425	2.352941	I bought both boxed sets books 15 Really a gr...	[i, bought, both, boxed, sets, books, 15, real...	[bought, boxed, sets, books, 15, really, great...	[bought, box, set, book, 15, realli, great, se...	[bought, boxed, set, book, 15, really, great, ...	i bought boxed set book really great series st...
1	I enjoyed this short book. But it was way way ...	3	Neutral	123	4.878049	I enjoyed this short book But it was way way t...	[i, enjoyed, this, short, book, but, it, was, ...	[enjoyed, short, book, way, way, short, see, e...	[enjoy, short, book, way, way, short, see, eas...	[enjoyed, short, book, way, way, short, see, e...	i enjoyed short book but way way short i see e...
2	I love Nicholas Sparks. I’ve read everyt...	4	Positive	1723	3.250145	I love Nicholas Sparks I8217ve read everything...	[i, love, nicholas, sparks, i8217ve, read, eve...	[love, nicholas, sparks, i8217ve, read, everyt...	[love, nichola, spark, i8217v, read, everyth, ...	[love, nicholas, spark, i8217ve, read, everyth...	i love nicholas sparks i read everything writt...

Machine Learning Model: After processing the data, I have used '*Multinomial Naive Bayes*' and '*Random Forest Classifier*' to predict the future comment/feedback. The models were trained on 70% of the data and tested on the rest 30%. Both the models performed well and achieved an accuracy around **83%**.