Customer Sentiment Analysis

BY

KSHITIJ MAMGAIN

python

Contents

1. Introduction
2. Project Objectives
3. About Data Set
4. Text Preprocessing
5. Data Exploration
6. NLP ML Models
7. BERT
8. Comparative Results
9. Conclusion
10. Recommendations

Introduction

- Data is the new oil!
- Over 2.5 quintillion bytes of **data** are **created every** single **day.**
- Data as – Structured, Semi-Structured and Unstructured.
- There are 2 trillion searches **per year** worldwide. That is over 40,000 search queries per second!
- Worldwide over 100 million messages are sent **every** minute via SMS and in-app messages! 26 billion **texts** were sent each **day by** 27 million people **in the** US.
- It is this text or human expression that machines endeavor to understand humans with **NLP.**

Project Objective

Our objective is to:

- Create Machine Learning Model for text classification
- Test the performance on accuracy, precision and recall

About Data Set

- Data Collection: The data set is available .csv file
- The dataset has customer reviews from e-commerce website on various products that they bought.
- There are total 10000 number of observations with no missing value
- Each review is classified into positive and negative label.

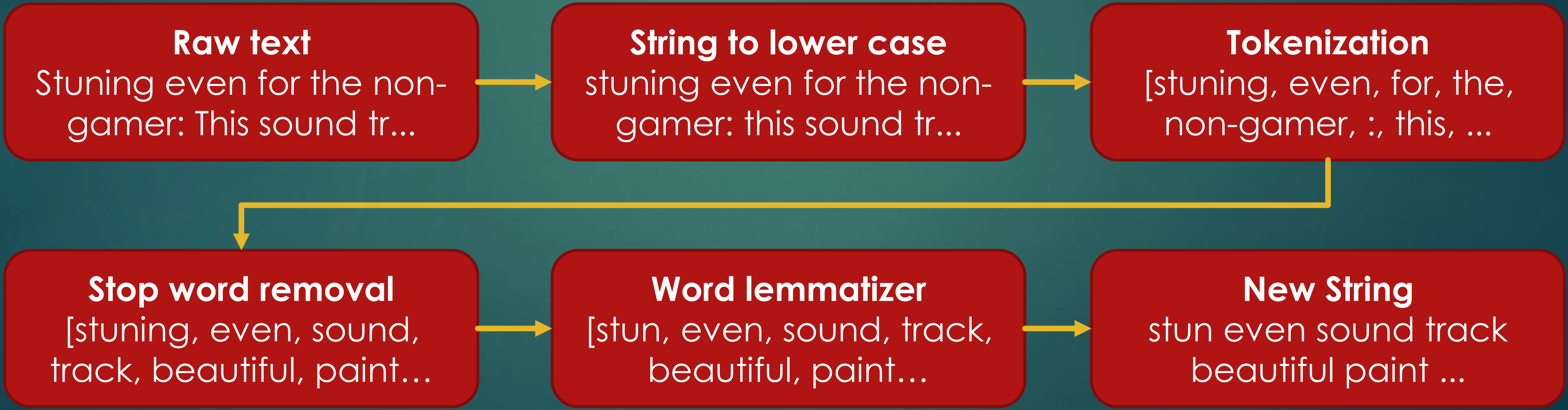
Not the shape I was going for....: I basically felt as if it squeezed my waist in and pushed all my "excess" down into the lower half of my torso. Great product in concept but didn't work at all for me.

Wow!: All and all this is a very nice product..I brought one with straps and bought the strapless one from Amazon......much cheaper than a boutique.Cher!

Actually, this book is not worth one star. I lost my interest in Jack Higgins Sean Dillion's serial after finishing this ridicules book.

Text Preprocessing

- Text preprocessing is *traditionally* an important step for natural language processing (NLP) tasks.
- It transforms text into a more digestible form so that machine learning algorithms can perform better.



Raw text
Stuning even for the non-gamer: This sound tr...

String to lower case
stuning even for the non-gamer: this sound tr...

Tokenization
[stuning, even, for, the, non-gamer, :, this, ...

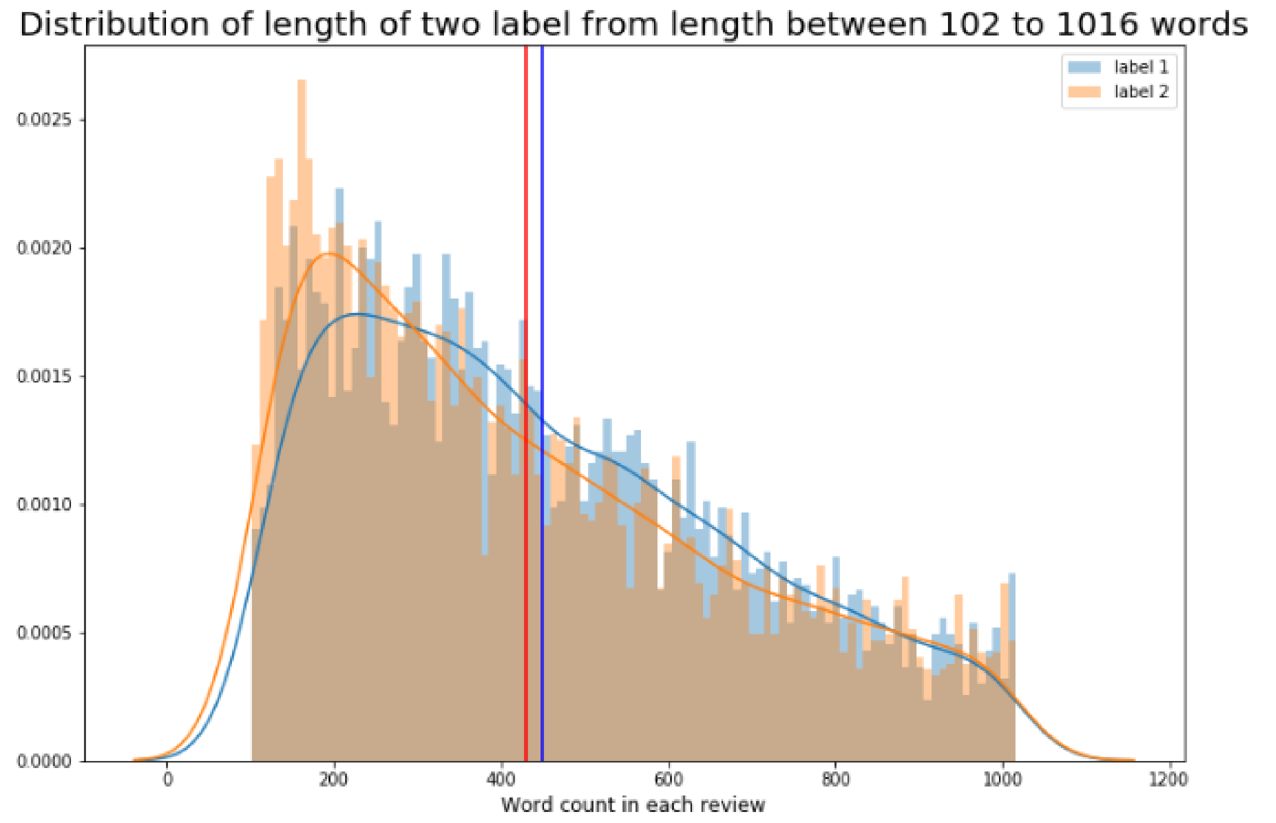
Stop word removal
[stuning, even, sound, track, beautiful, paint…

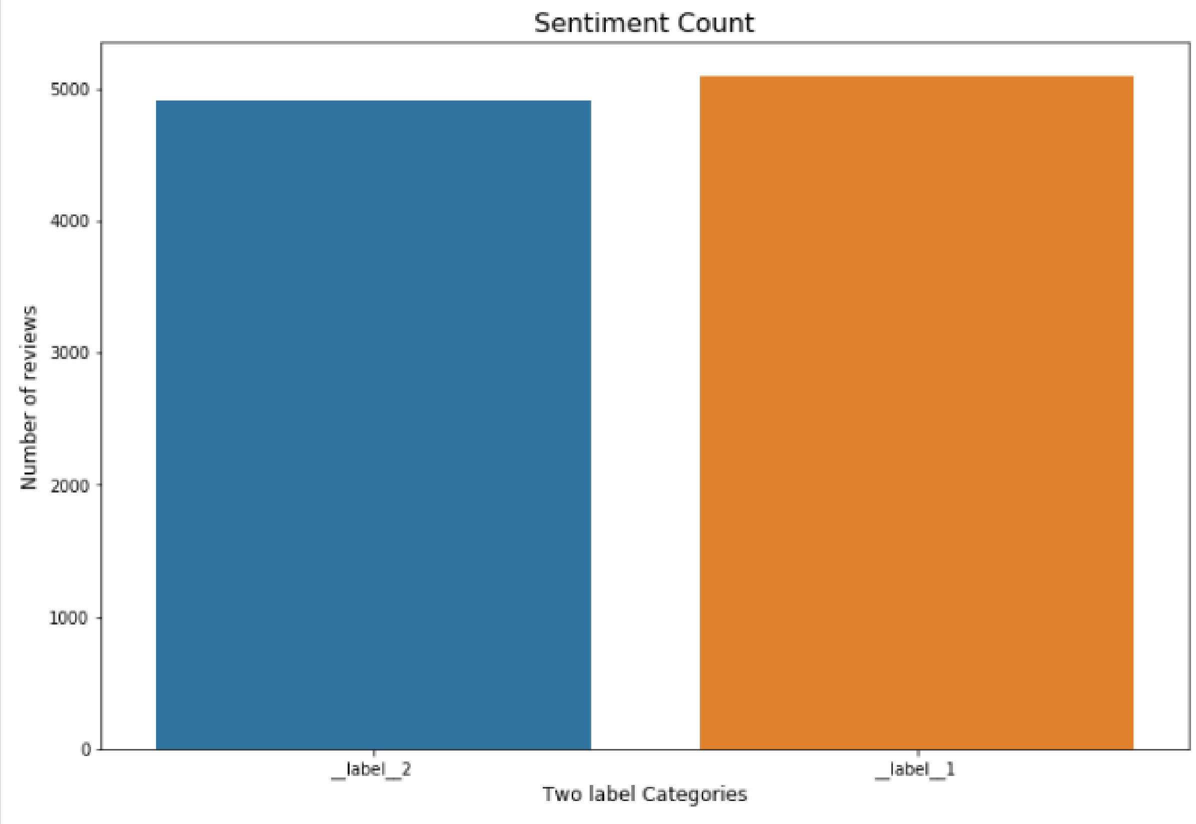
Word lemmatizer
[stun, even, sound, track, beautiful, paint…

New String
stun even sound track beautiful paint ...

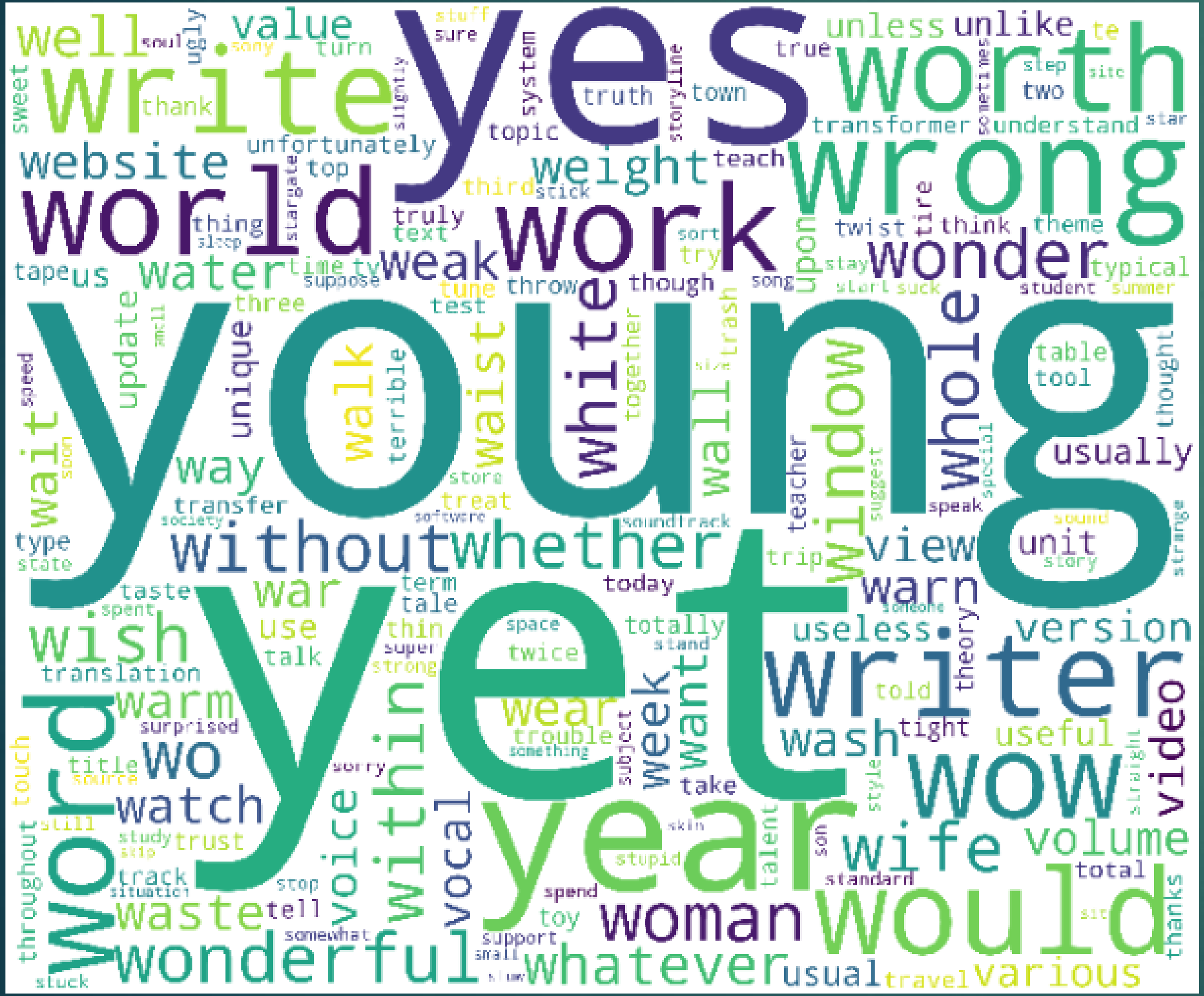
Data Exploration

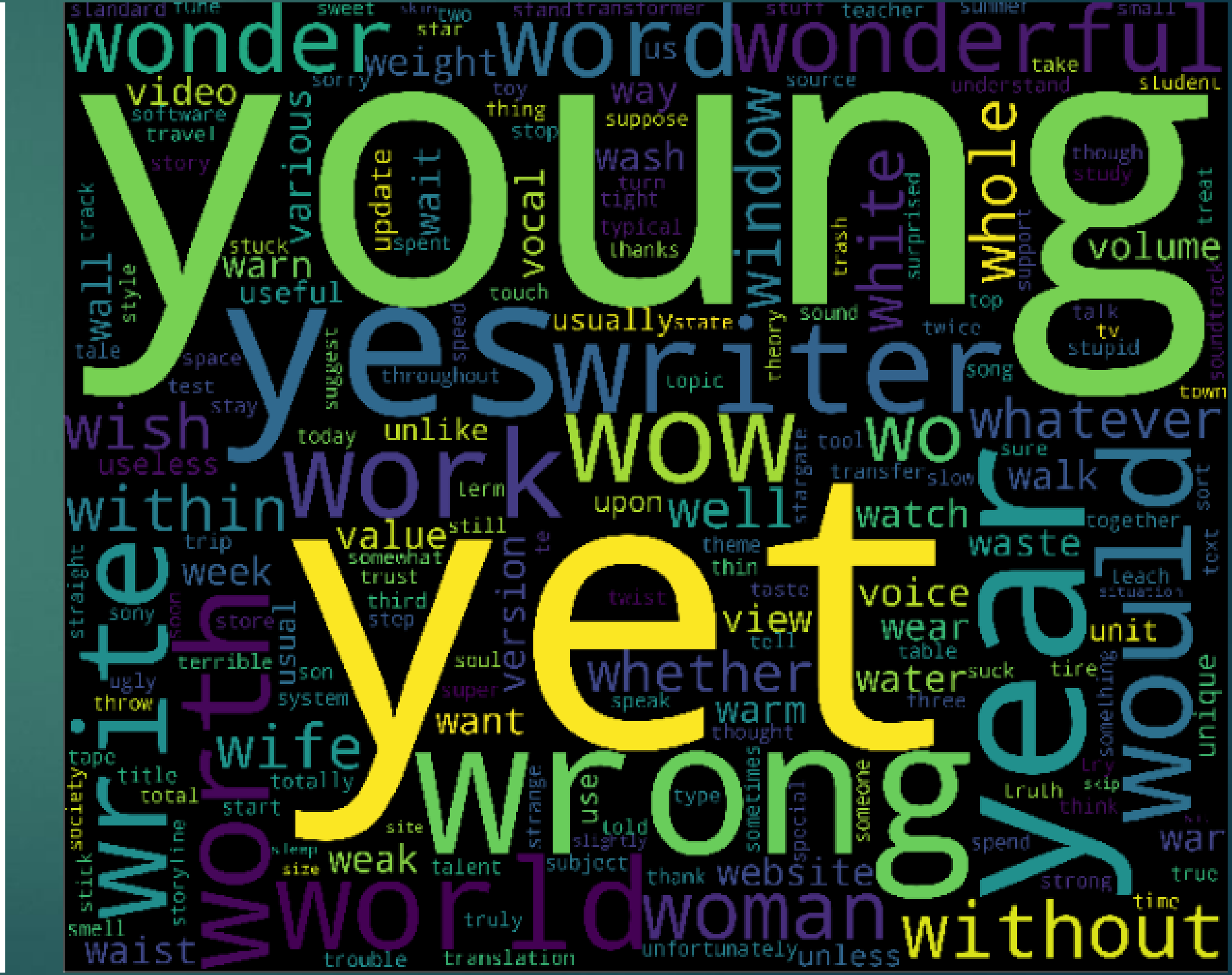
- The data is a pure NLP problem since we notice no data imbalance or significant difference in the review length





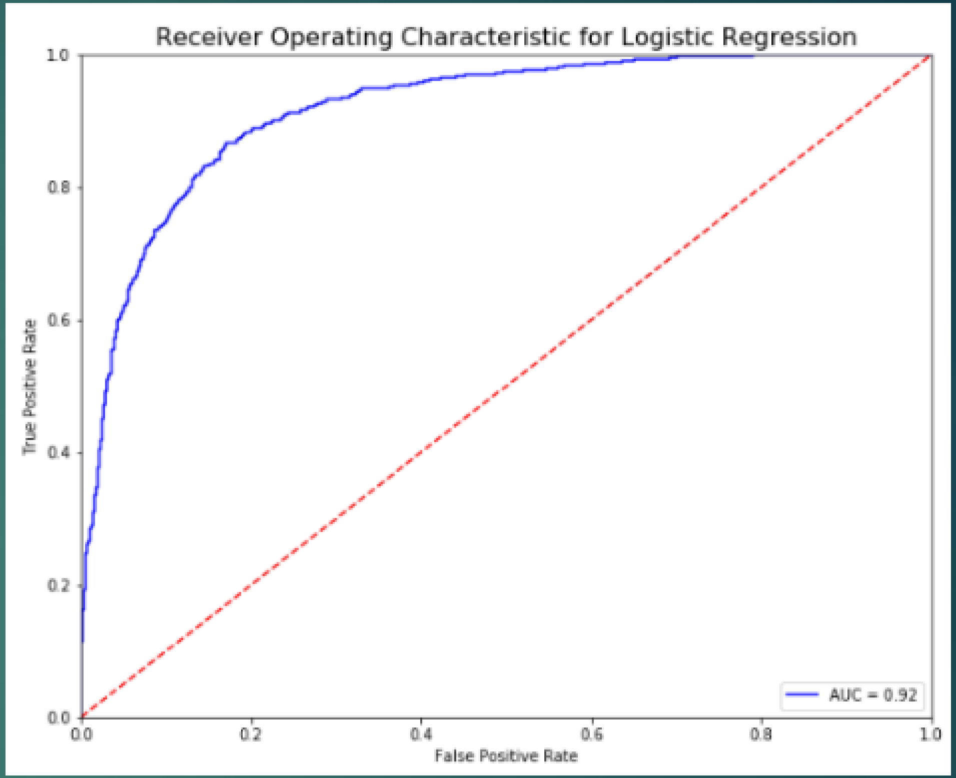
Data Exploration (contd.)

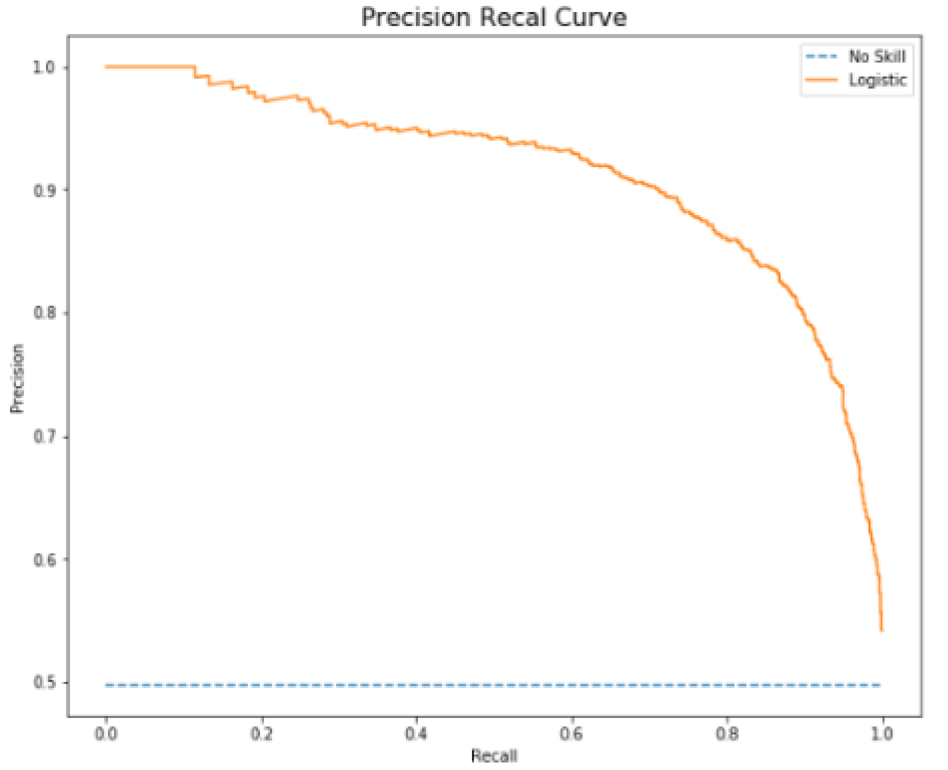




NLP Machine Learning Model

Testing on Logistic Regression provided an overall accuracy of 85.05% with Cross Validation


Receiver Operating Characteristic for Logistic Regression

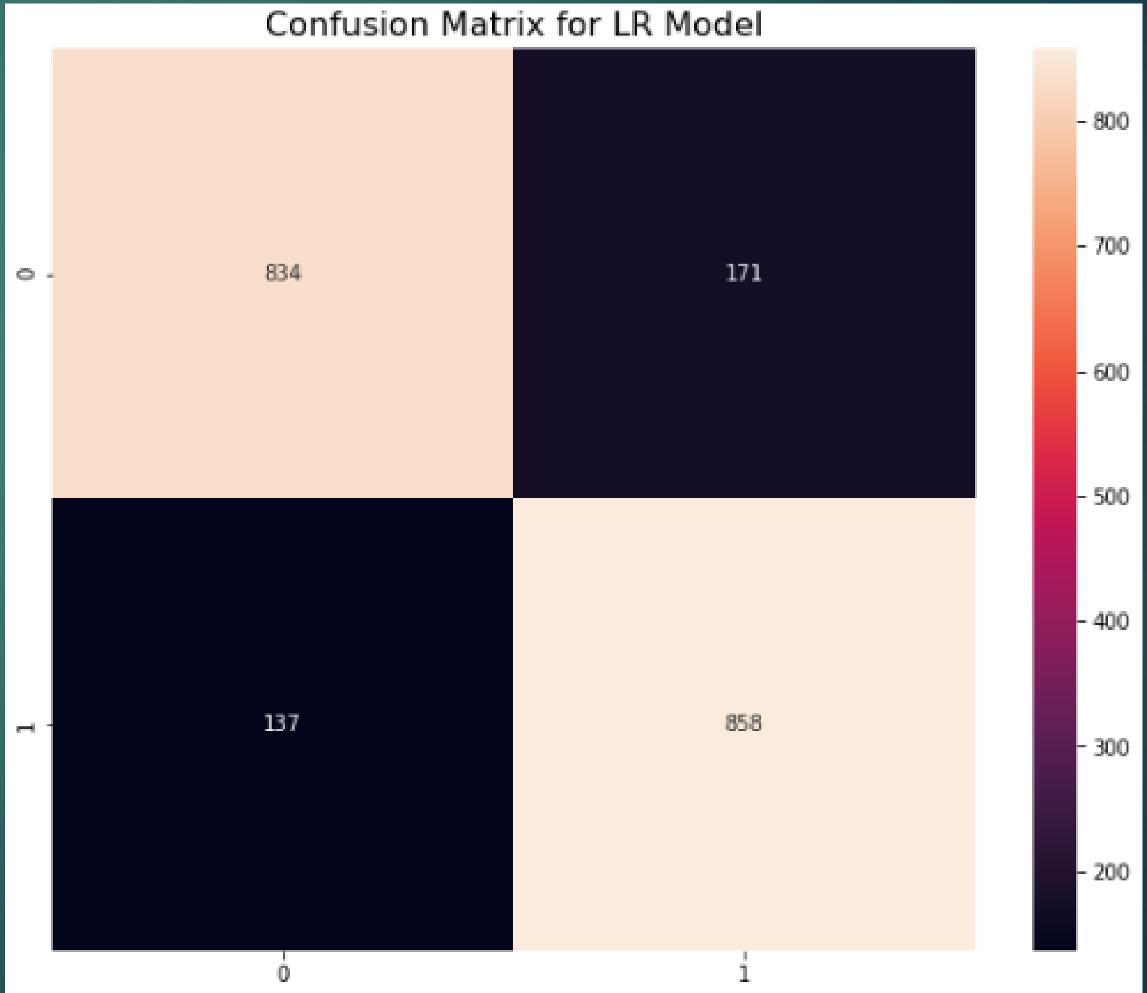

Precision Recal Curve

	precision	recall	f1-score	support
0	0.86	0.83	0.84	1005
1	0.83	0.86	0.85	995
accuracy			0.85	2000
macro avg	0.85	0.85	0.85	2000
weighted avg	0.85	0.85	0.85	2000

Mislabeled Data

'clean story appropriate part really funny horrible sickest humor do movie see definately kid expect well deed adam sandler…

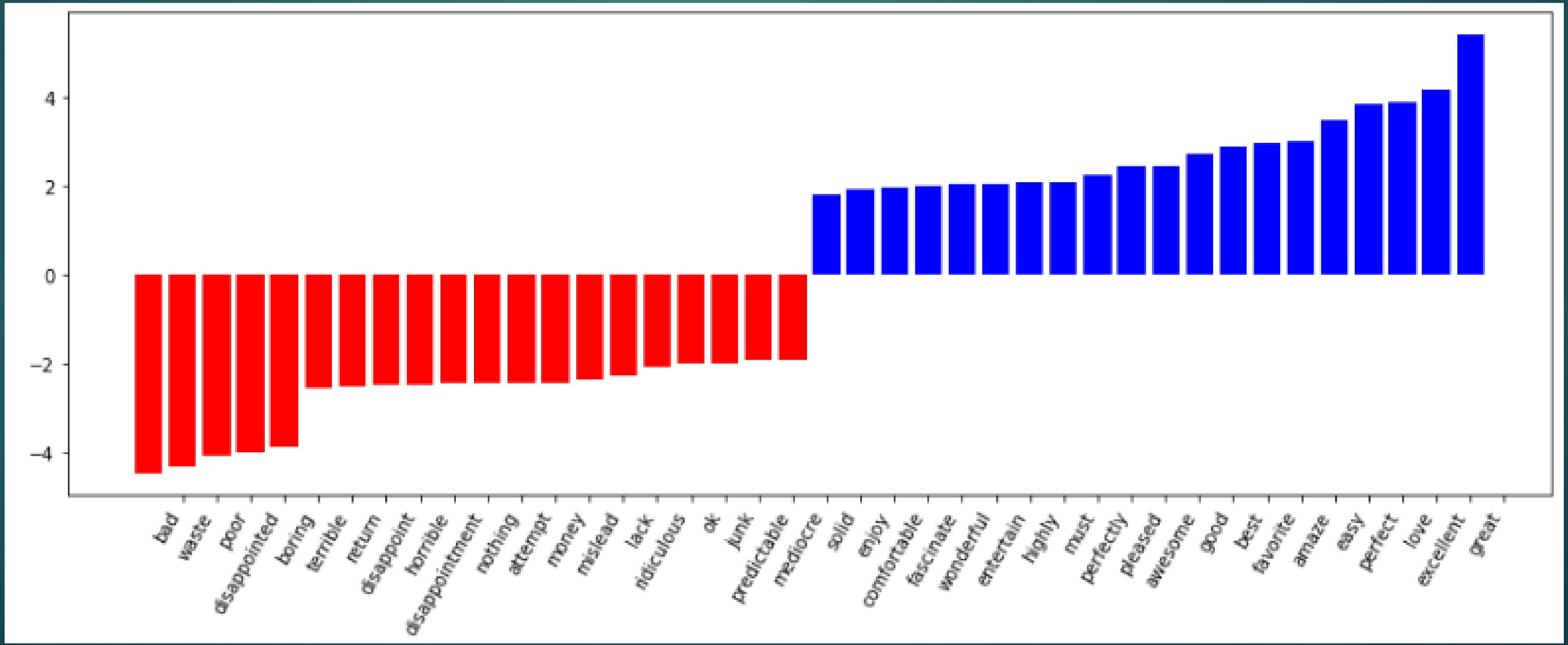
'…tough music less aggressive drum less furious pace overall calmer get wrong limp still spark energy creativity become mature know anything band say find shellac tortoise arty improvisation people call'



Support Vector Machine Classifier

- SVM is considered to be a great classifier for text-analysis
- The accuracy core with CV was 85.12% better than LR

- The results work best with linear kernel since in text classification it treats each word as a feature dimension and calculates the widest distance or broadest street.

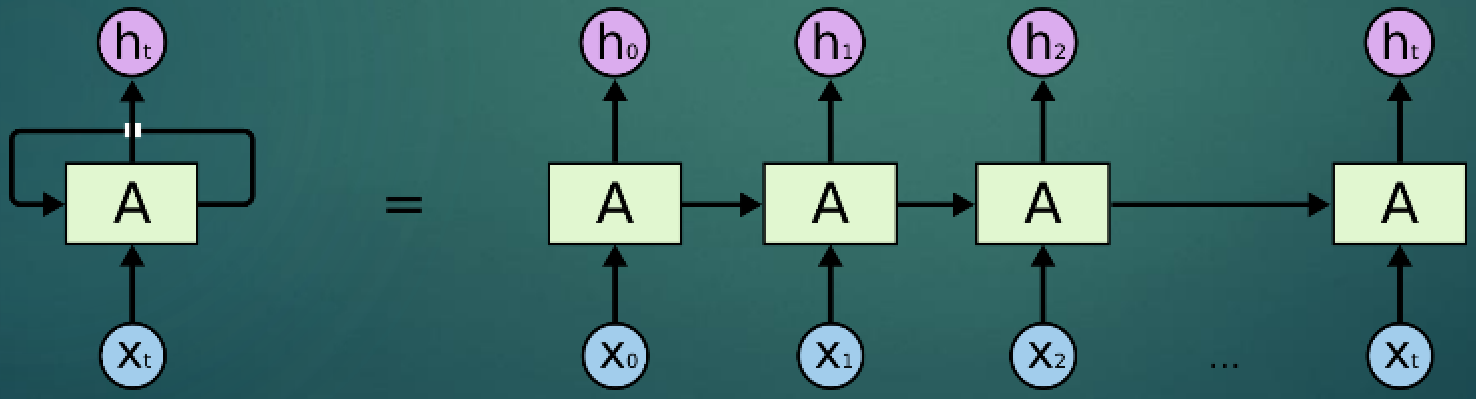


Limitation of simple models

- Key limitation of classic ML models is that these did not take the context of the word into account.
- One limitation of the embeddings in NLP was the use of very shallow Language Models. This meant there was a limit to the amount of information they could capture
- This motivated the use of deeper and more complex language models like LSTM and GRU

Long-short term memory (LSTM)

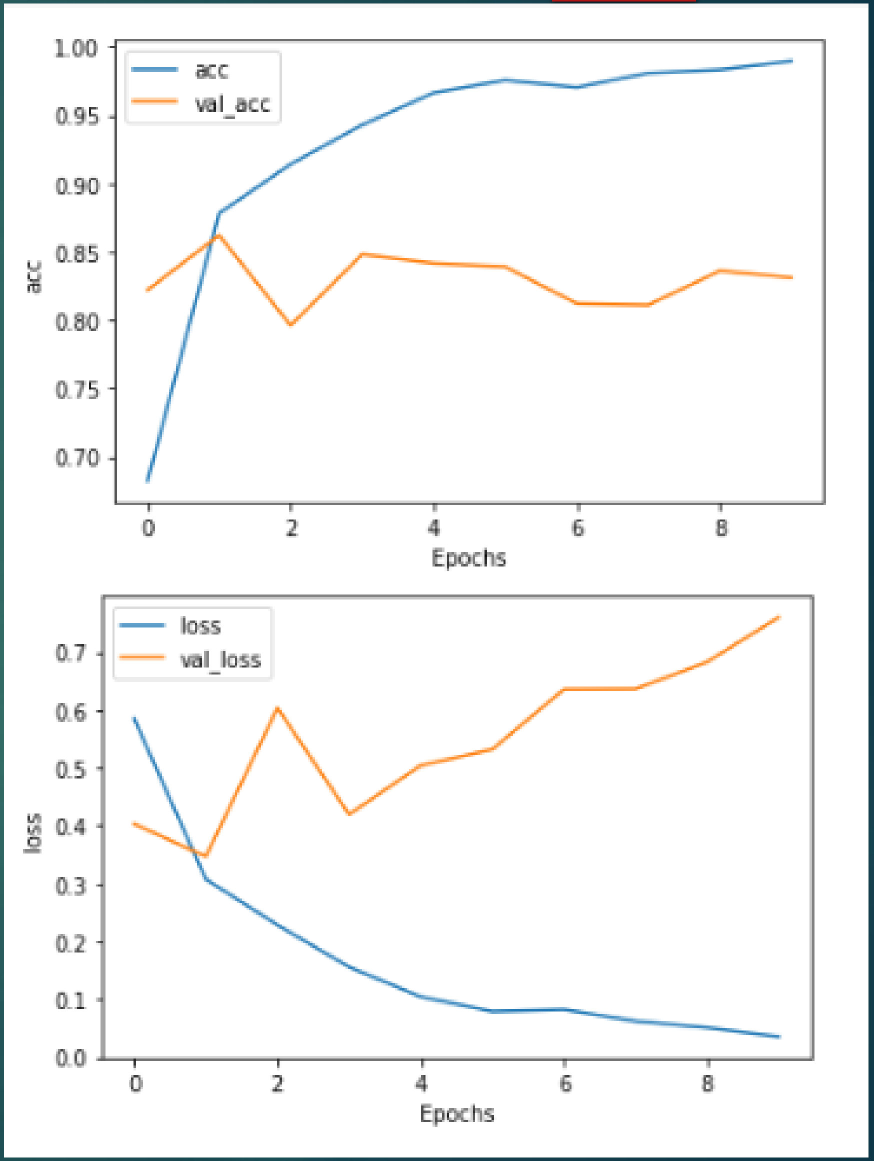
- RNN and LSTM and derivatives use mainly sequential processing over time.
- To understand the context LSTM module can bypass units and thus remember for longer time steps.



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM (contd)

- LSTM training model after one epoch had the best results : loss: 0.3072 - acc: 0.8780 - val_loss: 0.3465 - val_acc: 0.8620
- However, training with greater epochs and regularizing did not yield great results.
- "…the biggest challenges in natural language processing is the shortage of training data. Because NLP is a diversified field with many distinct tasks, a few thousand or a few hundred thousand human-labelled training examples." – Google AI



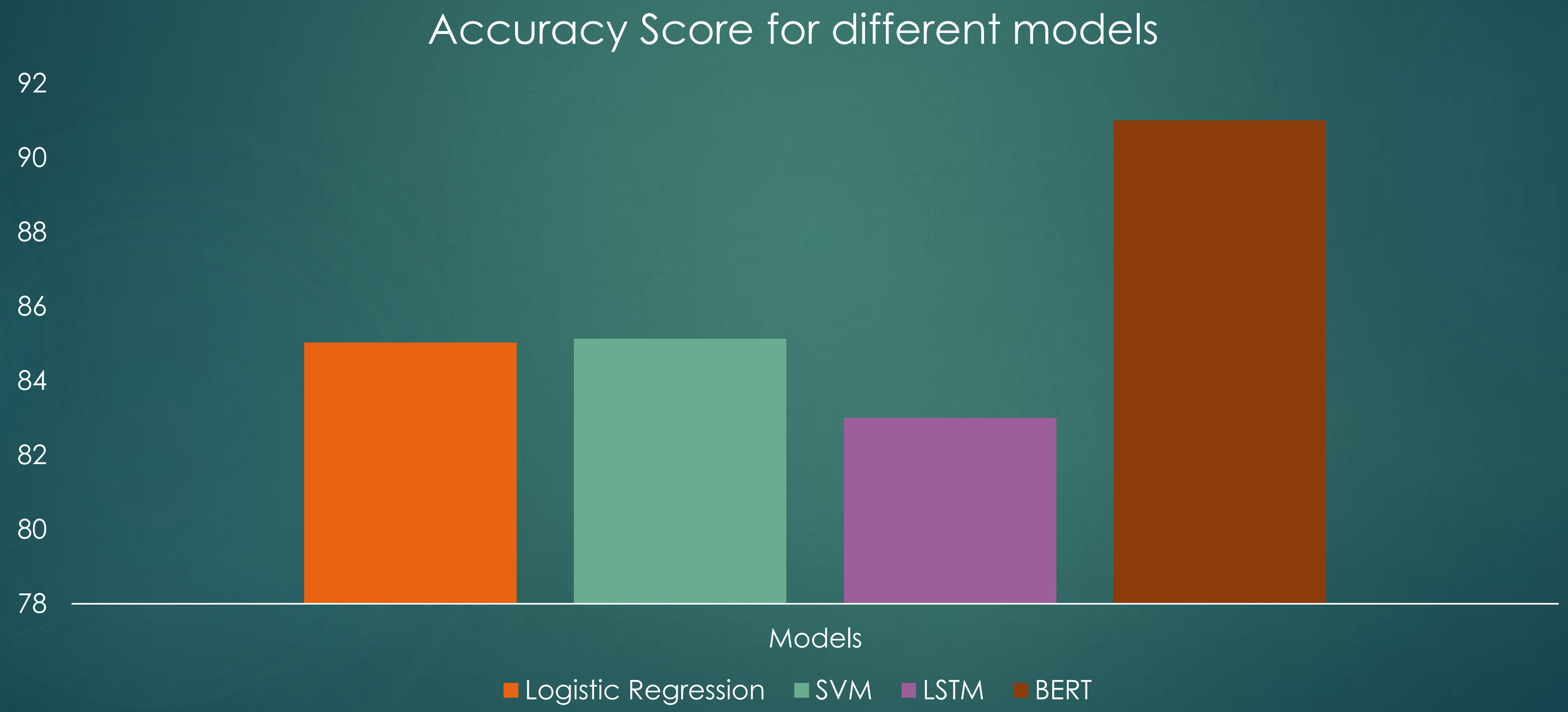
New Approach to NLP - BERT

- ***B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) is a 2 step process:*

1. Train a language model on a large unlabeled text corpus (unsupervised or semi-supervised)
2. Fine-tune this large model to specific NLP tasks to utilize the large repository of knowledge this model has gained (supervised)

- Embedding is a combination of 3 embeddings: Position Embeddings, Segment Embeddings and Token Embeddings
- BERT is pretrained on two NLP tasks: Masked Language Modeling and Next Sentence Prediction

Comparative Results



Conclusion

- The simple ML algorithm with right pre-processing provide a robust model to analyze the text.
- The SVM classifier with linear kernel is a great text classifier
- To improve the accuracy in sentiment analysis models like BERT have taken over earlier popular RNN
- The BERT model has the highest accuracy score in the balanced dataset

Recommendations

- A successful text classification model offers great possibility to address the automation of customer feedback.
- Retail, Finance and IT sector has much use of this technology to take feedback on the products, services and brands

The administrators staff give students a very hard time. They are not profes sional.

METRO COLLEGE
OF TECHNOLOGY

MCT has helped me increase my knowledge and be competitive in the market.

```
tensor([[0]]) tensor([[[0.0001]]],
grad_fn=<SigmoidBackward>)
```

```
tensor([[1]]) tensor([[[0.9999]]],
grad_fn=<SigmoidBackward>)
```

Thank You