Text sentiment analysis using deep learning

Vanita Kalaichelvan

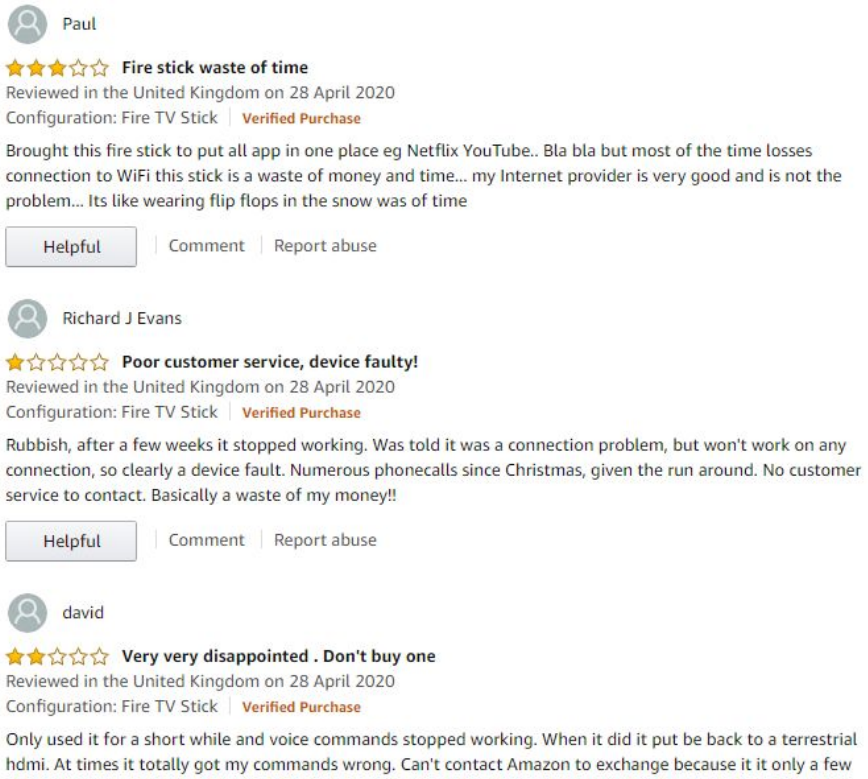
Project Aim

- To create a sentiment analysis tool for unlabelled texts
- Help businesses identify bad reviews so that they can react faster
- To learn how neural networks work

Dataset

Amazon reviews

- Rating from 1 to 5
- Review body



Paul

Fire stick waste of time

Reviewed in the United Kingdom on 28 April 2020

Configuration: Fire TV Stick | Verified Purchase

Brought this fire stick to put all app in one place eg Netflix YouTube.. Bla bla but most of the time losses connection to WiFi this stick is a waste of money and time... my Internet provider is very good and is not the problem... Its like wearing flip flops in the snow was of time

Helpful | Comment | Report abuse

Richard J Evans

Poor customer service, device faulty!

Reviewed in the United Kingdom on 28 April 2020

Configuration: Fire TV Stick | Verified Purchase

Rubbish, after a few weeks it stopped working. Was told it was a connection problem, but won't work on any connection, so clearly a device fault. Numerous phonecalls since Christmas, given the run around. No customer service to contact. Basically a waste of my money!!

Helpful | Comment | Report abuse

david

Very very disappointed . Don't buy one

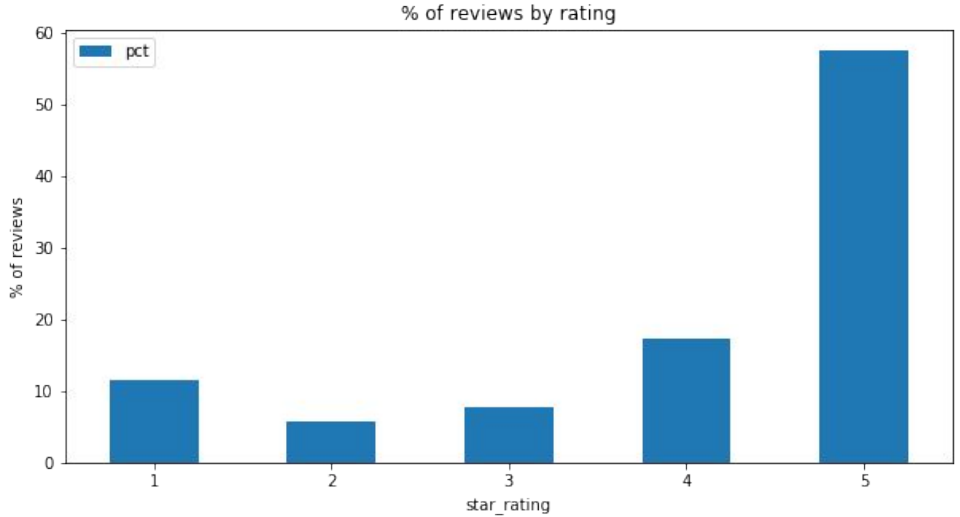
Reviewed in the United Kingdom on 28 April 2020

Configuration: Fire TV Stick | Verified Purchase

Only used it for a short while and voice commands stopped working. When it did it put be back to a terrestrial hdmi. At times it totally got my commands wrong. Can't contact Amazon to exchange because it it only a few

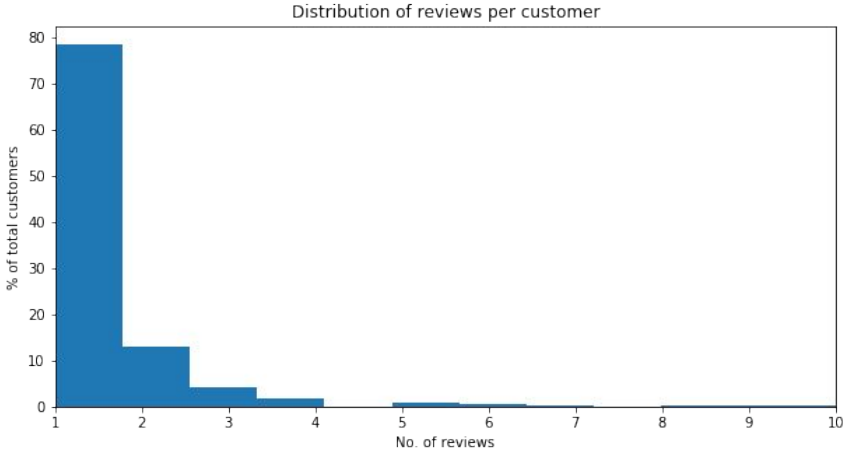
Data Exploration

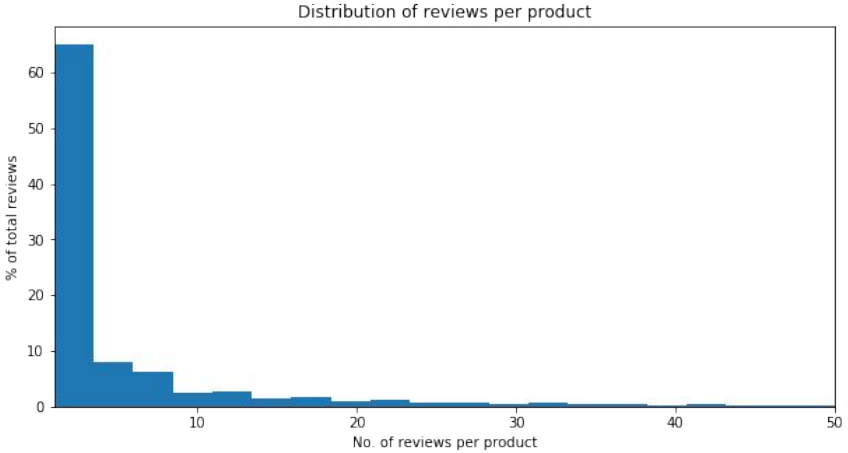
- 3 Million + reviews
- From 2000 to 2016
- Most reviews are positive - marketplace feedback loop



Data Exploration

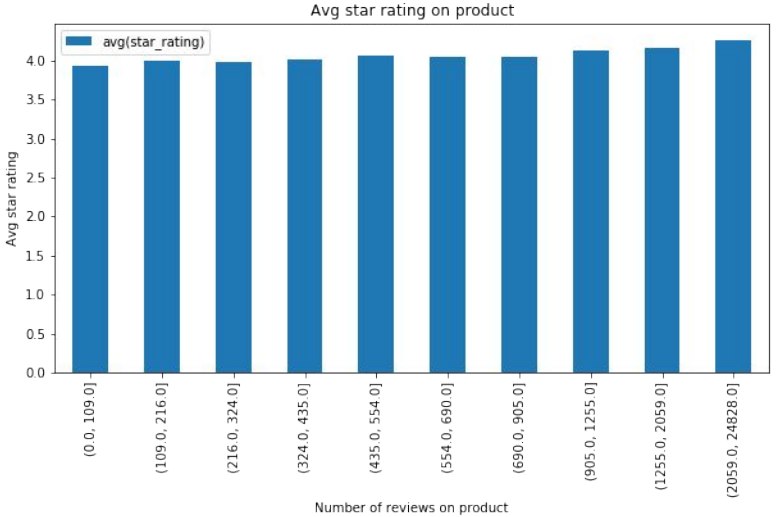
- Most customers only post 1 review
- Most products have less than 10 reviews
- A few products have many reviews - marketplace feedback loop

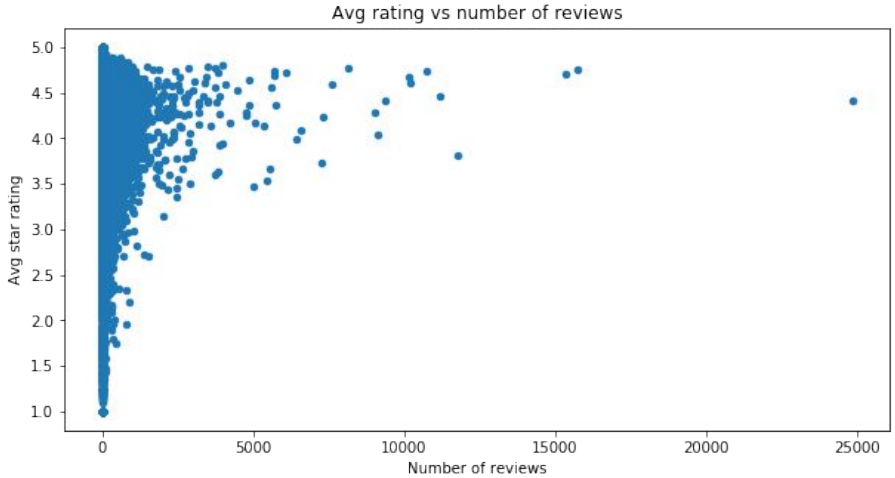




Data Exploration

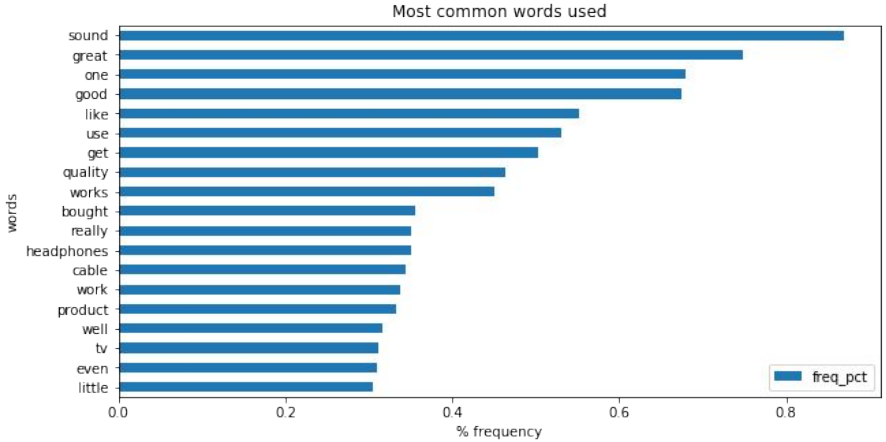
- Popular products with more reviews have higher ratings
- Marketplace dynamics - why buy a product with low ratings?

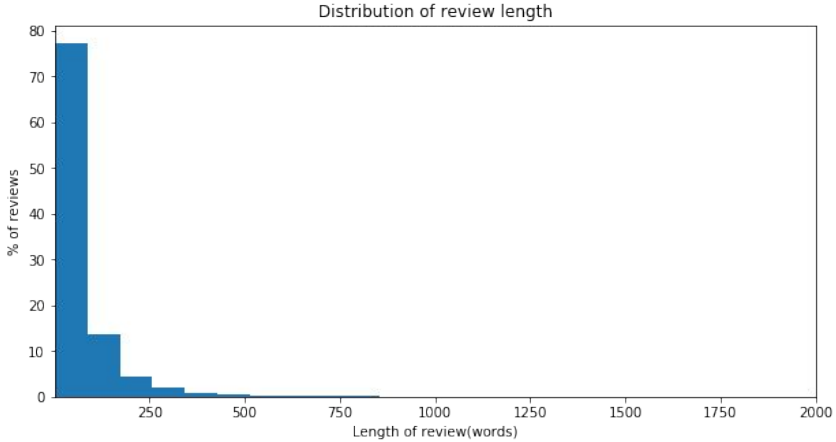




Data Exploration

- Most reviews are less than 250 words long
- Popular words are either positive or category related (within electronics)





Data Wrangling

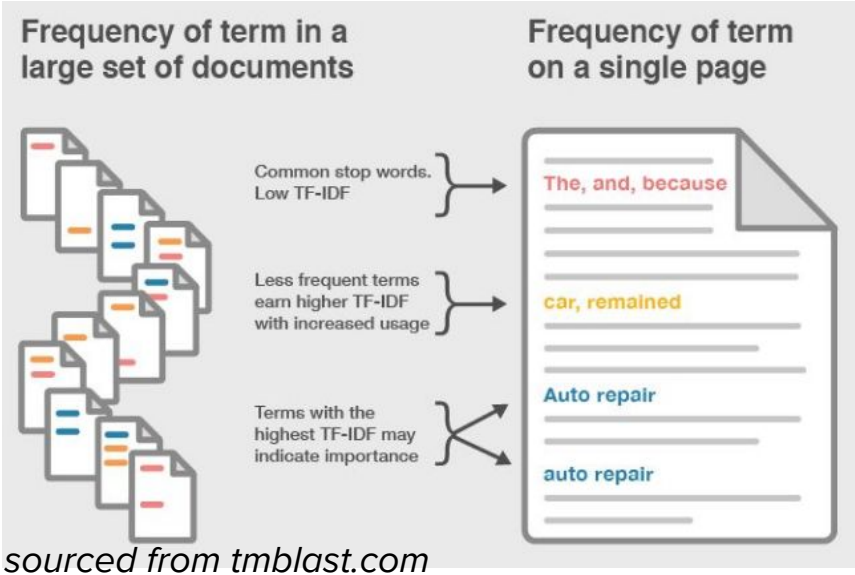
- Remove empty reviews and any null values
- Remove punctuation from text and lowercase all letters
- One-hot encode ratings for multi-class classification problem
- Feature Engineering - transforming words to numbers

Feature Engineering

- TF-IDF Tokenization
- N-grams
- Word Embeddings

TF-IDF

- TF - number of times a word appears in a review
- IDF - inverse of the number of times a word appears in all the reviews



sourced from tmblast.com

N-grams

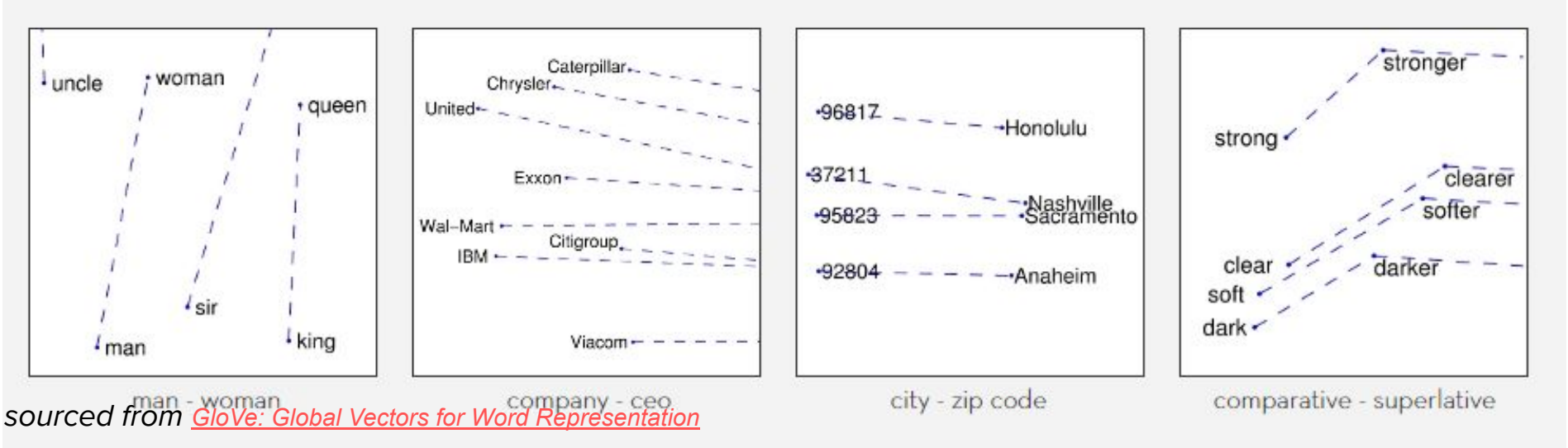
- Cannot always treat each word independently e.g. great, not great

This is a sentiment analysis tool

Unigram	This	is	a	sentiment	analysis	tool
Bigram	This is	is a	a sentiment	sentiment analysis	analysis tool	
Trigram	This is a	is a sentiment	a sentiment analysis	sentiment analysis tool		

Word Embeddings

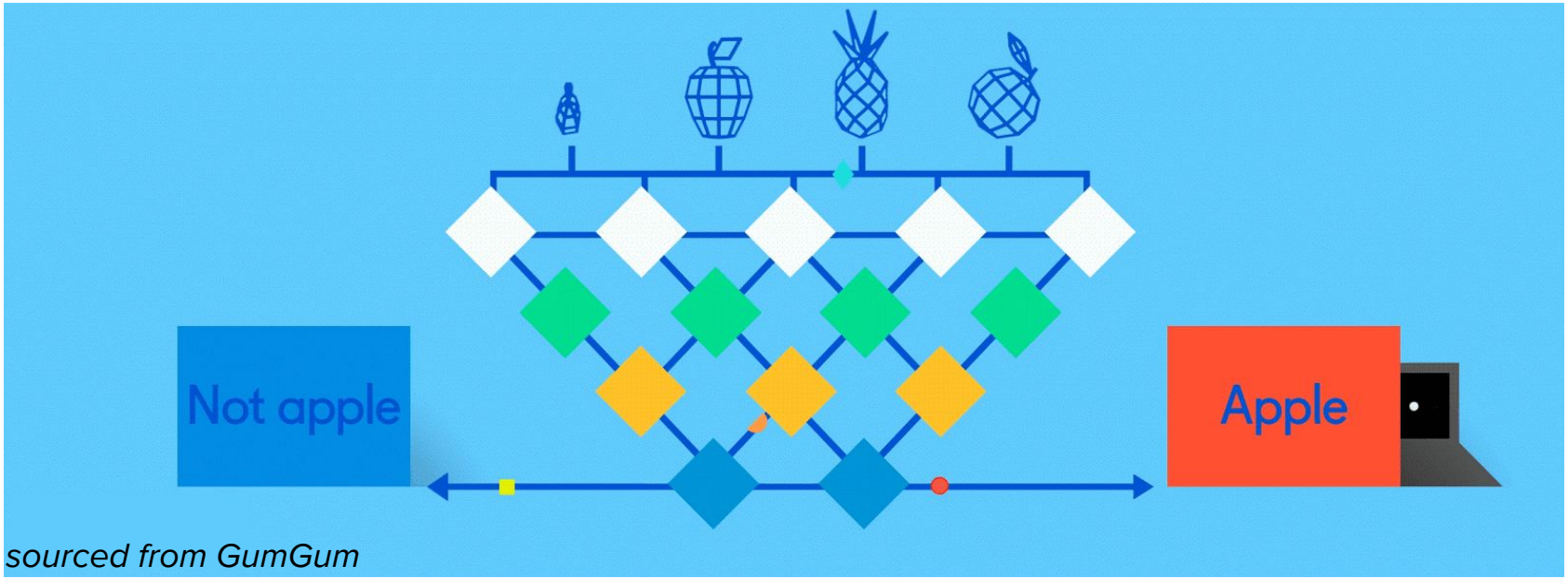
- Learning representations of words such that similar words have similar representations
- Can learn with the model or use pre-trained embeddings e.g. GloVe



sourced from [GloVe: Global Vectors for Word Representation]

Modelling

- Neural Networks
  - trains weights connecting layers to minimise final error



sourced from GumGum

Modelling

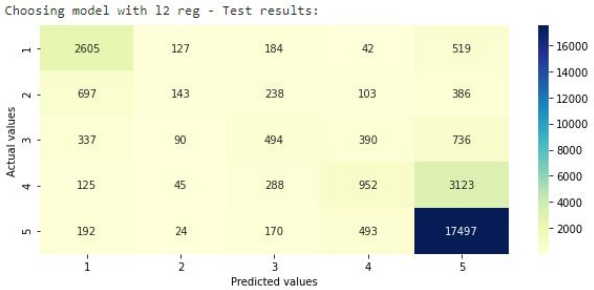
- Simple neural networks still look at each word independently
- Need neural networks that understand the concept of time and can process the whole sequence at each node
  - Recurrent neural network
  - Long Short Term Memory neural network
- RNN/LSTM carry information of words through time

Models

- TF-IDF tokenization with Simple NN
- TF-IDF + N-grams tokenization with Simple NN
- Word Embeddings with LSTM
- Metrics: Accuracy, Recall on bad reviews

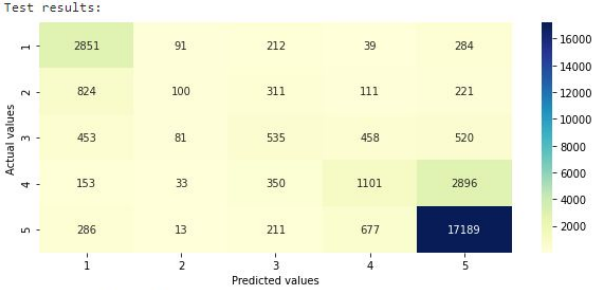
Evaluation

TF-IDF with Simple NN



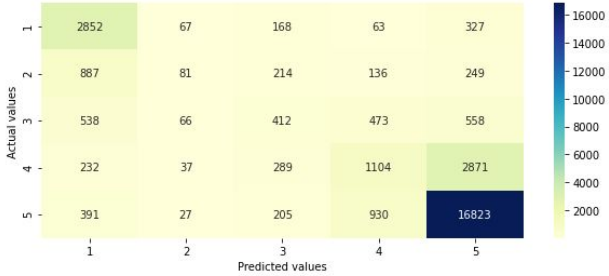
Accuracy for class 1 is 74.92%
Accuracy for class 2 is 9.13%
Accuracy for class 3 is 24.13%
Accuracy for class 4 is 21.00%
Accuracy for class 5 is 95.22%
Mean class accuracy is 44.88%
Test accuracy is 72.30%
The recall for bad reviews is 70.82%

TF-IDF + Ngrams with Simple NN



Accuracy for class 1 is 82.00%
Accuracy for class 2 is 6.38%
Accuracy for class 3 is 26.14%
Accuracy for class 4 is 24.29%
Accuracy for class 5 is 93.54%
Mean class accuracy is 46.47%
Test accuracy is 72.59%
The recall for bad reviews is 76.65%

Word embeddings with LSTM



Accuracy for class 1 is 82.02%
Accuracy for class 2 is 5.17%
Accuracy for class 3 is 20.13%
Accuracy for class 4 is 24.35%
Accuracy for class 5 is 91.55%
Mean class accuracy is 44.64%
Test accuracy is 70.91%
The recall for bad reviews is 77.06%

Evaluation

- Best performing model is the TF-IDF + N-grams with Simple NN
- Don’t always need the most complex model for the problem (e.g. RNN/LSTM)
- LSTM model has the best recall score

Main Takeaways

- Deep Learning models can be hard to tune due to their black box like nature
- Simpler models can perform better based on dataset
- Using deep learning means sacrificing a lot of interpretability (e.g. feature importance)

Future Work

- Reformulating problem as a binary one
- Implementing simpler models e.g. Naive Bayes, SVM
- Implementing more complex models e.g. BERT
- Tuning hyperparameters for current models
- Finding more balanced dataset to train on