KINDLE BOOK REVIEWS

DSO 560 – Final Project

Team Purple

AMAN AGARWAL

CHINMAYI BENGALURU PRAKASH

FALAK JAIN

NAVEEN KUMAR MANJUNATHA

Problem Statement

- The Amazon Kindle reviews dataset is studied using machine learning models to extract key insights
- In this project, we leverage the user reviews data to perform text analysis exploring the following:
  - Topic modeling: To generate key themes user reviews include, how they are rated and
  - Sentiment Analysis: To Analyze the sentiment of reviews and classify them further into positive, negative and neutral classes
- The Kindle reviews dataset with over 980K reviews is used after being filtered to about 115K of the most recent reviews to keep analysis accurate to current trends

Business Scope

- Topic modelling allows us to identify themes across reviews and discover hidden areas of concern and dissatisfaction. Reviews are segmented across different ratings to determine what themes are common across positive reviews. This helps understand key areas of product improvement and identify key positive themes to further monetize on
- Human emotions can be studied by analyzing text. By creating a classification model to determine sentiment, unlabeled & labeled text data can be used to determine sentiment against products, improving customer experience
- These analyses will help Amazon determine which genre of products on its platform is driving traffic, user interest, criticism and it can then make targeted product recommendation and informed marketing decisions accordingly. These strategies will have a direct impact on the revenue

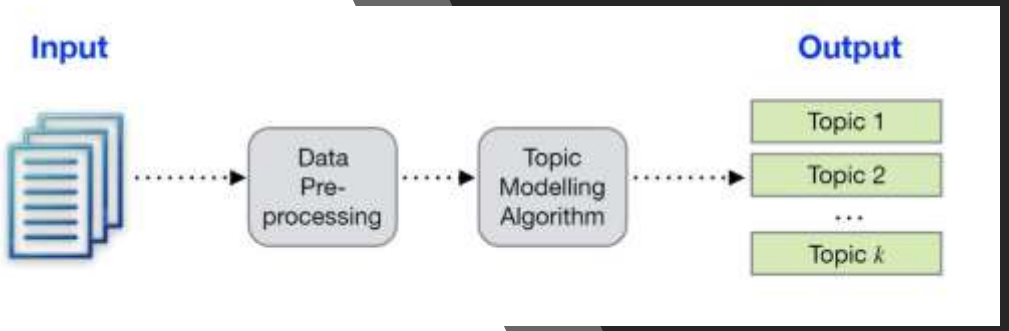
About the Kindle dataset

- Amazon Kindle is a series of e-readers designed and marketed by Amazon
- Amazon Kindle devices enable users to browse, buy, download, and read e-books, newspapers, magazines and other digital media via wireless networking to the Kindle Store
- Amazon Kindle platform consists of millions of user reviews over time, spread across wide range of products rich in textual data
- The raw data[1] set in this analysis contains product reviews from the Amazon Kindle store category by customers between June – July 2014
- It comprises of 11 fields and 115K records. The fields include *product id, review id, rating, user review, review summary, review timestamp* among others
- The field *rating* is used as labels for sentiment analysis. *ReviewText* consists of the text which is analyzed for both topic modelling and sentiment analysis. The other columns are not considered for further analysis

[1]*source : https://www.kaggle.com/datasets/bharadwaj6/kindle-reviews*

TOPIC MODELLING

What is Topic Modelling?



- *Topic Modeling* is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents
  - *It* refers to the process of dividing a corpus of documents in two:
  - A list of the topics covered by the documents in the corpus
  - Several sets of documents from the corpus grouped by the topics they cover
- *Unsupervised Model* - No labelled data required

- **Why Topic Modelling?**
  - Quick and easy to implement
  - Gives good understanding and insights into our data

How does Topic Modelling work?

- Topic modeling involves counting words and grouping similar word patterns to infer topics within unstructured data
- By detecting patterns such as word frequency and distance between words, a topic model clusters feedback that is similar, and words and expressions that appear most often and thereon quickly deduce what each set of texts are talking about
- A software company wants to know what customers are saying about features of your product

✗ Spend hours going through heaps of feedback

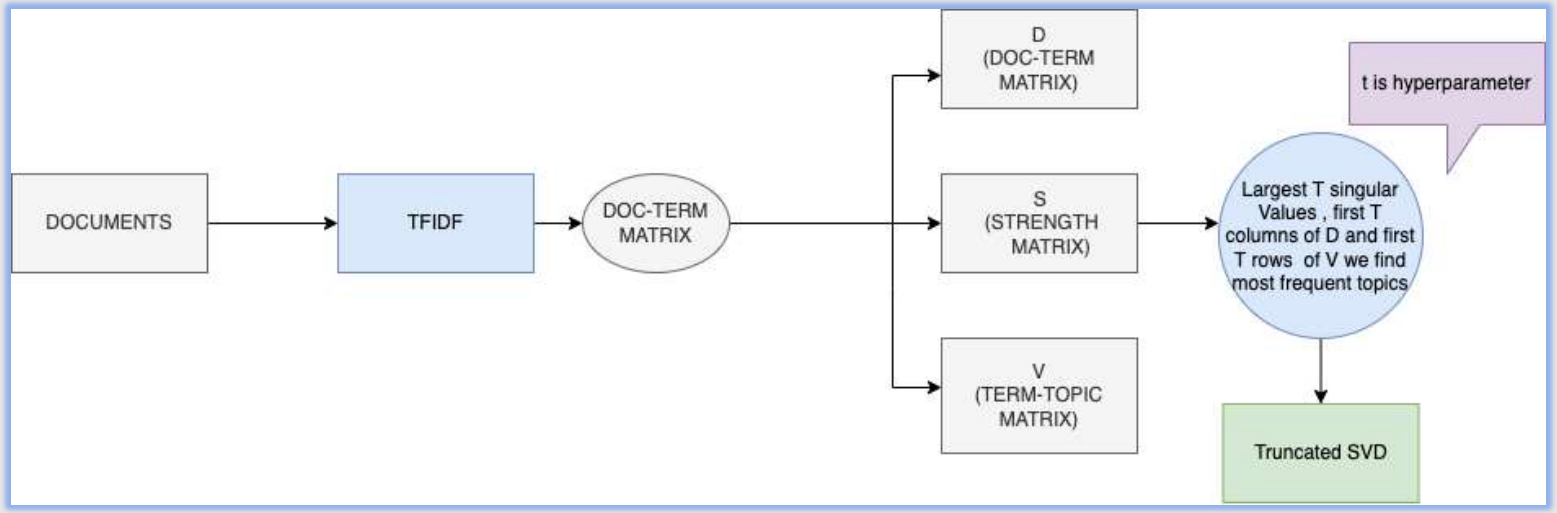
✓ Use t*opic modeling algorithm* to deduce that topics the texts are talking about

Latent Semantic Analysis (LSA)

- *LSA* is one of the most frequent topic modeling methods
- Based on distributional hypothesis which states that the semantics of words can be grasped by looking at the contexts the words appear in
- LSA computes how frequently words occur in the documents – and the whole corpus – and assumes that similar documents will contain approximately the same distribution of word frequencies for certain words
- Syntactic information (example: word order) and semantic information (example: the multiplicity of meanings of a given word) are ignored and each document is treated as a **bag of words**

LSA - Steps

1. Obtain *Document term matrix* using **TF-IDF** Vectorization
2. Decompose this matrix into the product of 3 matrices (USV) by using singular value decomposition ***(SVD)***
3. Keep the largest **t** singular values together with the first **t** columns of U and the first **t** rows of V, we can obtain the **t** more frequent topics found in our original *Document-term matrix*. We call this ***truncated SVD***
4. To use it for LSA, we set the value of **t** as a **hyperparameter**
5. Use matrices S &D to identify the best results



What is BERT?

- **BERT or BerTopic** is a topic modeling technique that uses transformers (BERT embeddings) and class-based TF-IDF to create dense clusters
- Better for contextual analysis, understands the context of the sentence

- **Why is BERT better than LSA?**
  1) More interpretable clusters
  2) More accurate results
  3) More efficient representation

BERT Topic algorithm - Steps

The Bert topic algorithm contains **3 stages :**

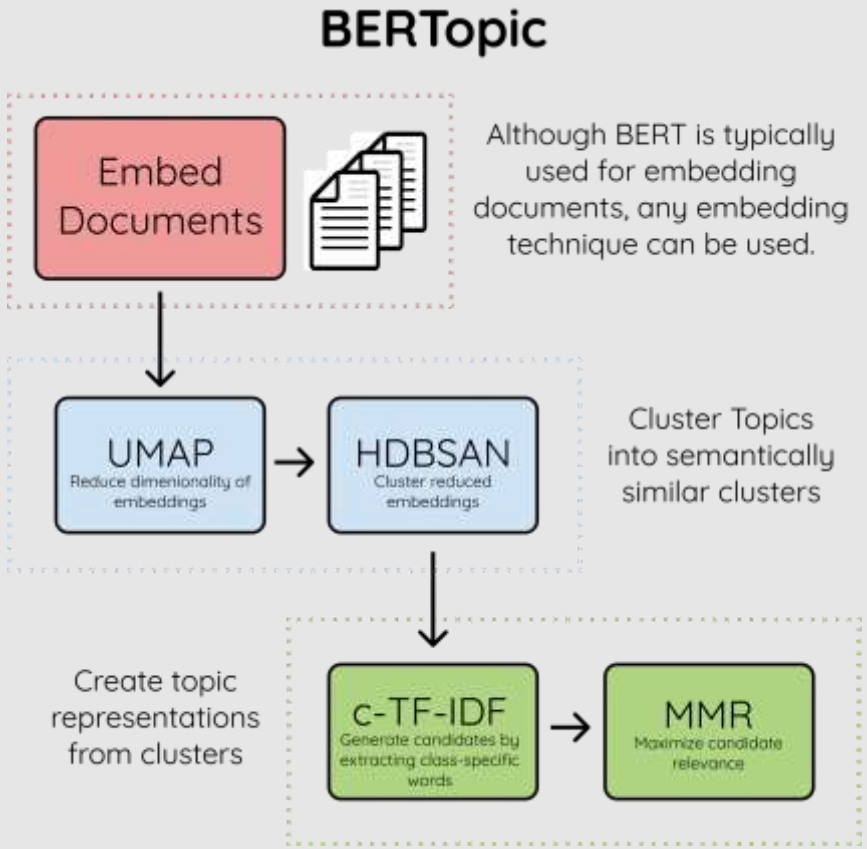
1. **Embed the textual data(documents)**
   It uses the following sentence transformers
   A) paraphrase-MiniLM-L6-v2
   B) paraphrase-multilingual-MiniLM-L12-v2

2. **Cluster Documents**

Uses UMAP to reduce the dimensionality of embeddings and the HDBSCAN technique to cluster reduced embeddings and create clusters of semantically similar documents

3. **Create a topic representation**

Extract and reduce topics with class-based TF-IDF and the improve the coherence of words with Maximal Marginal Relevance



Approach

1. Use only reviews after 2014-06-01
2. Pre-process all the review text (remove stop words, punctuations, digits, lemmatize)
3. Divide these cleaned reviews into ***3 pots***:
    - Less than or equal to Overall Rating 2
    - Between Overall Rating = 3 and Overall Rating = 4
    - Overall Rating = 5

    This way we can see which topics drives bad, good or excellent ratings

4. Now run topic models on each of the data pots using :
    - Latent Semantic Analysis (LSA) (Trigram Approach)
    - BertTopicModelling (Unigram Approach)

Model findings - LSA

These are the most frequently occurring topics for each of the ratings as per our LSA Topic Model (in order of frequency high to low)

Poor Ratings (1,2)

- Book not recommended
- Waste of time
- Unable to finish book
- Wanted and looked forward to the book but end up not liking

Good Ratings (3,4)

- Look forward to read more
- Honest book exchange
- Look forward to book series
- Free copy book
- Wicked read

Excellent Rating (5)

- Highly recommended book
- Look forward to more of the book
- Can't wait for more book in the series
- Great job by the author

Model Findings – BERT Topic

These are the most frequently occurring topics for each of the ratings as per our BERT Topic Model (in order of frequency high to low)

Poor Ratings (1,2)

- Not like the story
- Cooking, Recipe and food
- Ketogenic, calories, crabs and sugar

Average- Good Ratings (3,4)

- Love & life
- Sex, erotic & hot scenes
- Look forward to book Series
- War, alien, military, space & science
- Grammar, Spellings
- Murder, detective, solve , police
- Cookbooks, recipe
- Paranormal, supernatural, ghost

- Diet, healthy
- Character, storyline, plot
- Child, illustration, animal, pictures, kid
- Twist, surprise, turn, unexpected
- Short, sweet, love, romantic
- Short, story, quick, read
- Reading, recommend, author, enjoy

Excellent Rating (5)

- Father, mother, heart, life
- Plot, storyline, character, development
- Twist, turn, surprise , guess
- Paranormal, supernatural, ghost, romance

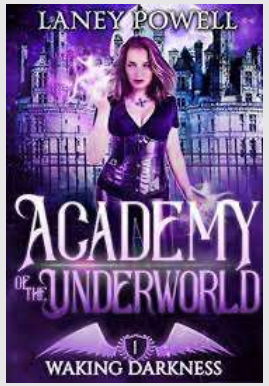
- Romantic, sweet, short, cute
- Recipe, cook, delicious, ingredients
- Laugh, funny, humor, cry, loud
- Enjoyable, good, read
- Amaze, wow, wait, pace
- Page, turner, Takerra
- Paleo, diet, recipe, eat, healthy
- Marketing, business, medium, social, customer
- Illustration, child, kid, rhyme
- Goal , motivation, achieve, success
- Trade, stock, option
- Dog, cat, puppy, animal
- Edge, seat, emperor, wait
- Gardening, plant
- Positive, happiness, thinking
- Zombie, apocalypse, dead , survival

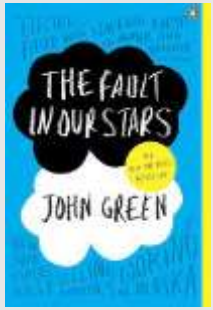
Business Recommendations to Amazon – Topic Modeling

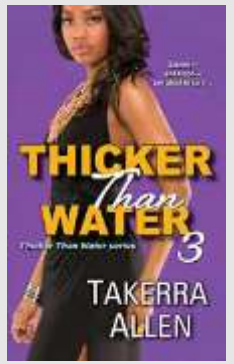
- Improve the collection of recipe books, food books as form a large chunk of poor reviews and fair chunk of average reviews and good reviews
- Potentially remove unpopular authors or cuisines
- Ketogenic books talking about carbs, calories and sugar not a popular choice on Kindle, can remove them
- Focus on books that are part of a series, belong to a very likeable author or is trending over the internet

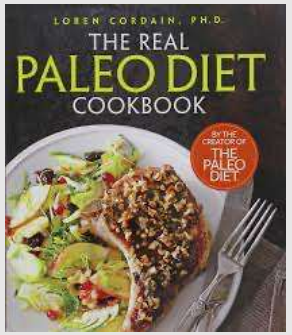
Market, promote and recommend the books in the following categories most :

1) Book about love, life and family
2) Books with good storyline and character development, maybe popular Fiction Books.
3) Books that offer twist, turns and surprises – thrillers
4) Paranormal, supernatural and ghost-themed books
5) Romantic books which are short, sweet and cute
6) Books which have humor and are fun to read
7) Well-paced books
8) Page Turner books
9) Books by Takerra Allen
10) Paleo Diet Books
11) Books that talks about markets, business & customers
12) Books for children consisting of rhymes and illustrations
13) Self-help books that talk about motivation, success and achievements
14) Books talking about stocks, trading and options
15) Animal books focusing on cats or dogs
16) Gardening books
17) Self-help books about positivity in life









SENTIMENT ANALYSIS

Sentiment Analysis – What, Why & Where?

WHAT?

- Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers

- There are multiple types of sentiment analysis:
    1. Graded Sentiment Analysis
    2. Emotion Detection
    3. Aspect-based Sentiment Analysis
    4. Multi-lingual sentiment analysis

WHY?

- With the advent of social media, humans have begun expressing their thoughts and feelings more widely than ever before, and sentiment analysis can be essential to monitor and understand sentiment in all types of data
- From a business standpoint, analyzing customer feedback allows brands to learn customer response better at different stages of marketing, sales and product lifecycle. This in turn helps the businesses tailor products and services to meet the needs of the customer.

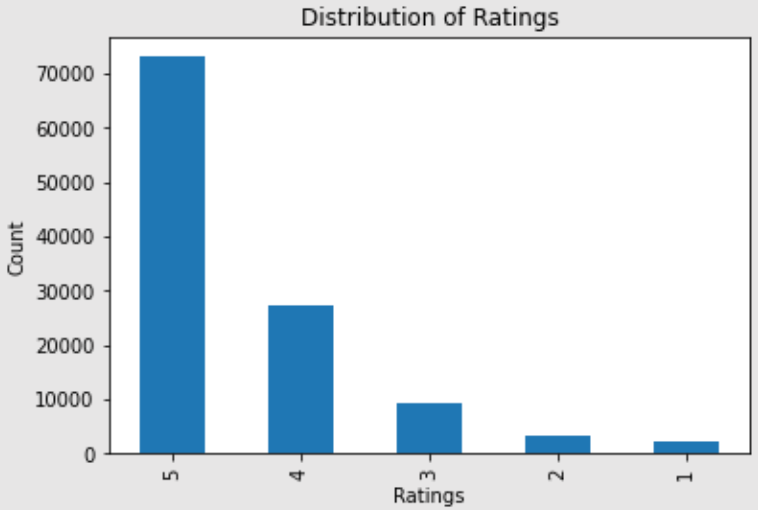
WHERE?

- Text data is extensively available today, including social media, reviews (product, movie, services), etc.,

In this project, we perform graded sentiment analysis (binary and multi-class), with the intent to estimate customer review sentiment and generate insights to drive business decisions

Approach towards Sentiment Analysis

- Given that the reviews all have corresponding ratings, Supervised machine learning techniques have been employed to classify the text. We have adopted the below methodologies to implement this



Binary Classification

- Positive : Rating 4,5
- Negative : Rating 1,2

Rating 3 is excluded for this analysis

Multi-class Classification

- Positive : Rating 4,5
- Negative : Rating 1,2
- Neutral : Rating 3

Modeling Techniques

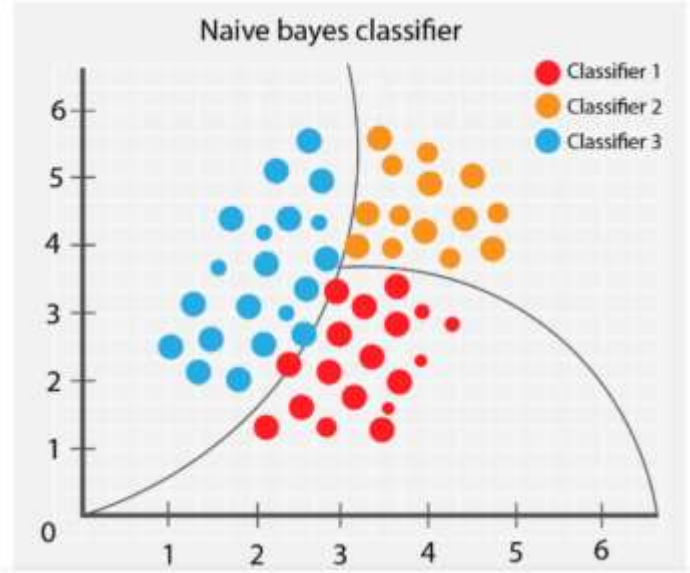
- Multinomial Naïve Bayes' Classification
- Logistic Regression
- Deep learning neural network (NN)
- Simple Recurring NN (RNN)
- Long short-term memory (LSTM)
- 1-dimensional Convolutional NN (CNN)

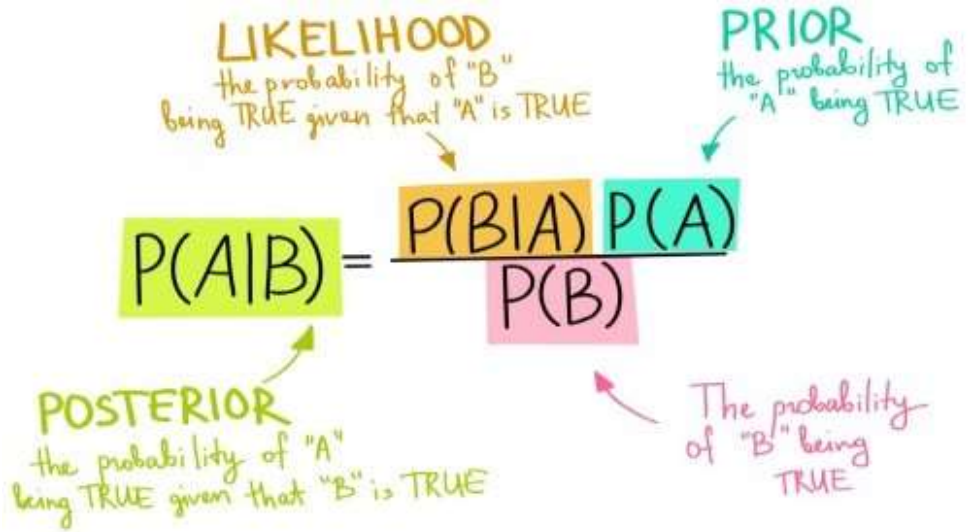
Text pre-processing

Removing Stopwords	• Created a custom stopwords list
Regex Cleaning	• Removed punctuations, special characters, extra white spaces, URLs
Upsampling	• Up-sampled the data by resampling labels for the imbalanced class
Vectorization	• Count-Vectorizer, TF-IDF, Word2Vec and N-grams [(1,1),(2,2)]
Train-Test Split	• Split data in 70:30 train-test ratio
Word Embeddings	• Generated custom word embeddings for the tokenized words in the dataset

Multinomial Naïve Bayes' Classification

- Multinomial Naive Bayes algorithm is a probabilistic learning method based on the Bayes theorem
- It predicts the tag of a text by calculating the probability of each tag in the sample and then returns the tag with the highest probability
- We have used this as a baseline model since it is simple and easy to implement

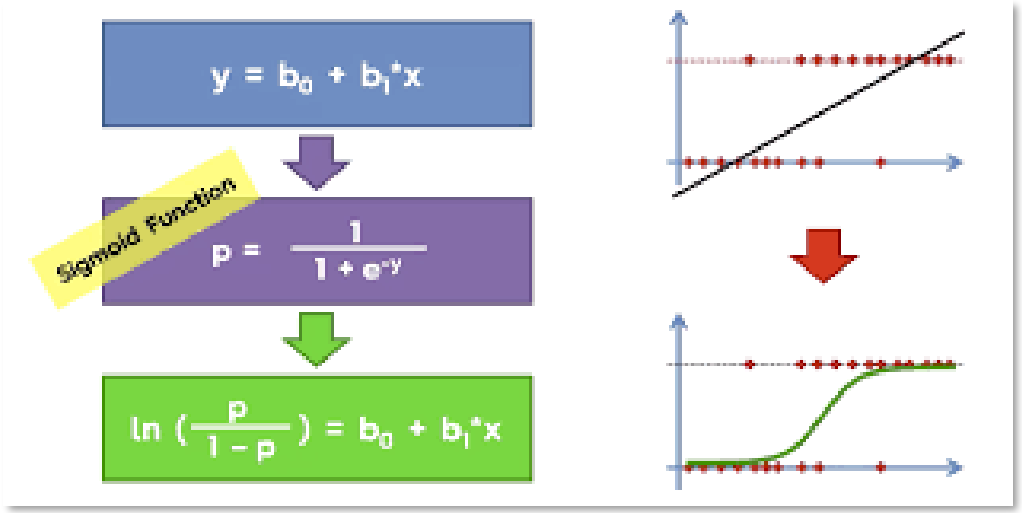




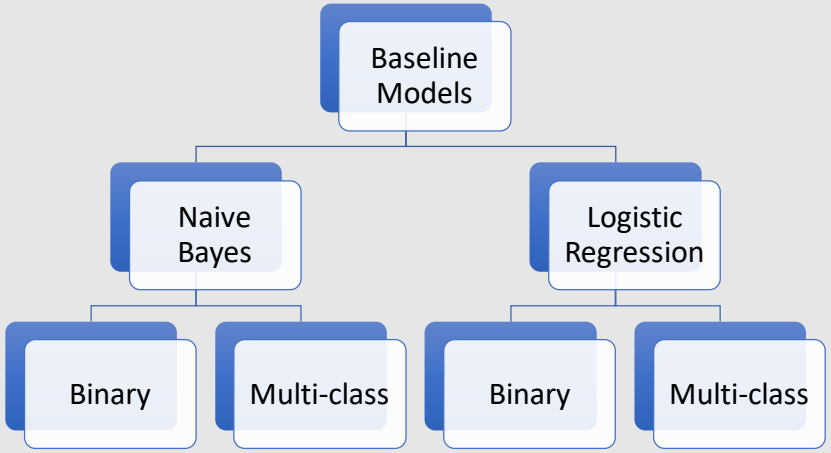
Logistic Regression

- This method is used for classification and models the probability of response variable given independent variables using a logistic function
- We use both binary and multiple logistic regression models in our analysis

- **Multinomial logistic regression (MLR)** is an extension of logistic regression that adds native support for multi-class classification problems.
- Although it is a binary classification model by default, extensions like one-vs-rest can allow logistic regression to be used for multi-class classification problems
- changing In MLR, the loss function is replaced by cross-entropy loss and multinomial probability distribution is used



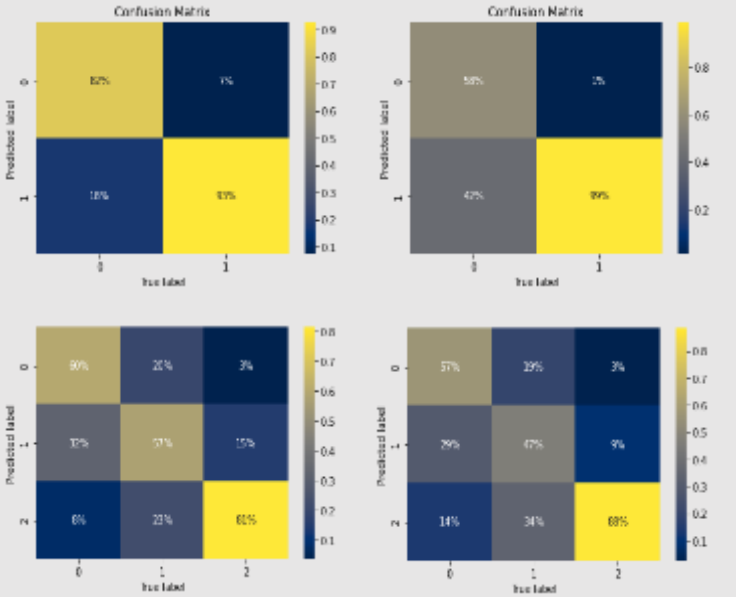
Baseline Model Results



Model	# Class	Vectorization	N-Grams	Max_features	Accuracy
Multinomial Naïve Bayes	Binary (0, 1)	TF-IDF Vectorizer	(1,1)	None	92.10%
				3000	91.26%
			(2,2)	None	95.03%
				3000	86.44%
Logistic Regression	Binary (0, 1)	Count Vectorizer	(1,1)	None	95.60%
				3000	93.79%
			(2,2)	None	96.75%
				3000	87.89%
Multinomial Naïve Bayes	Multi (0,1,2)	TF-IDF Vectorizer	(1,1)	None	78.50%
Logistic Regression	Multi (0,1,2)	Count Vectorizer	(1,1)	None	83.48%

Binary

Multi-class



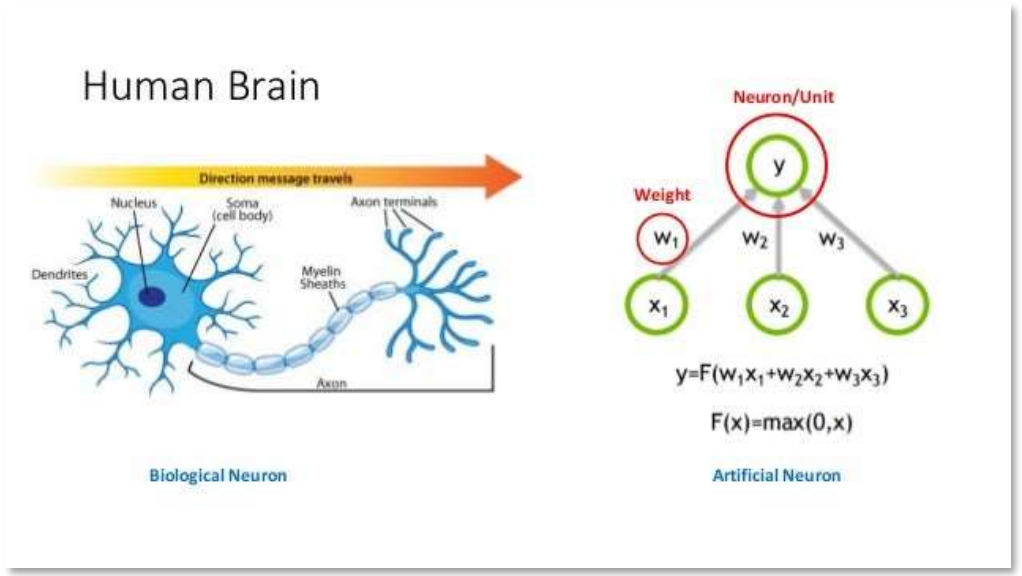
Naïve Bayes | Logistic Regression

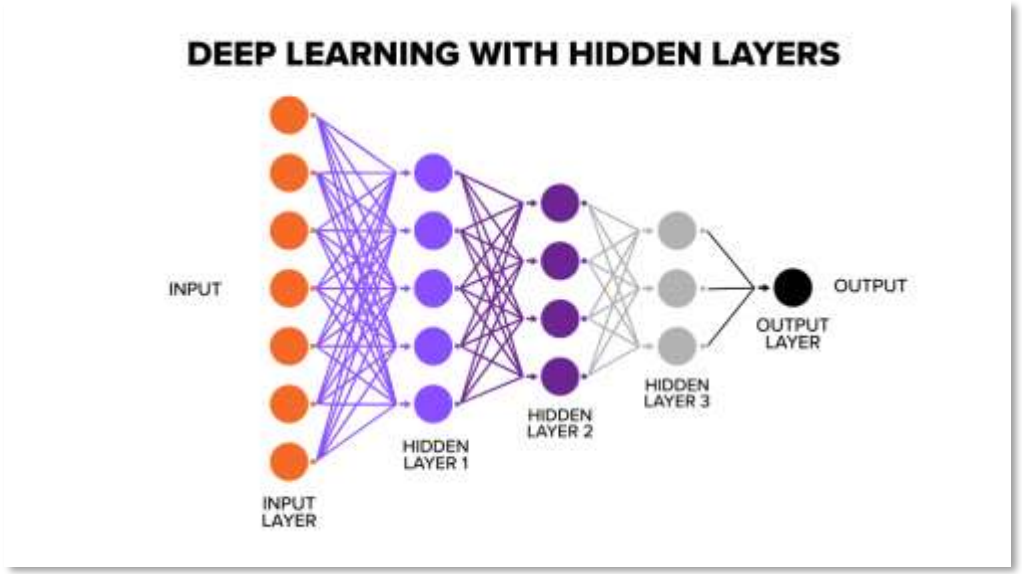
Confusion Matrix

- Given the model performance results, we see high classification error for multi-class models.
- As we proceed to advanced techniques, we decided to limit out analysis to binary sentiment classification.

Deep Learning Neural Network

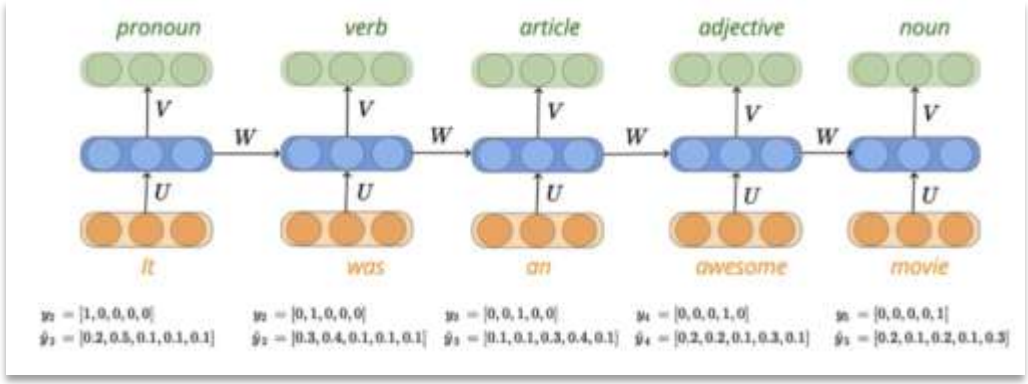
- Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers
- These neural networks attempt to simulate the behavior of the human brain
- Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization, thus increasing prediction accuracy
- It uses processes like forward propagation and backpropagation, and minimizes the loss using algorithms like gradient descent to build the model

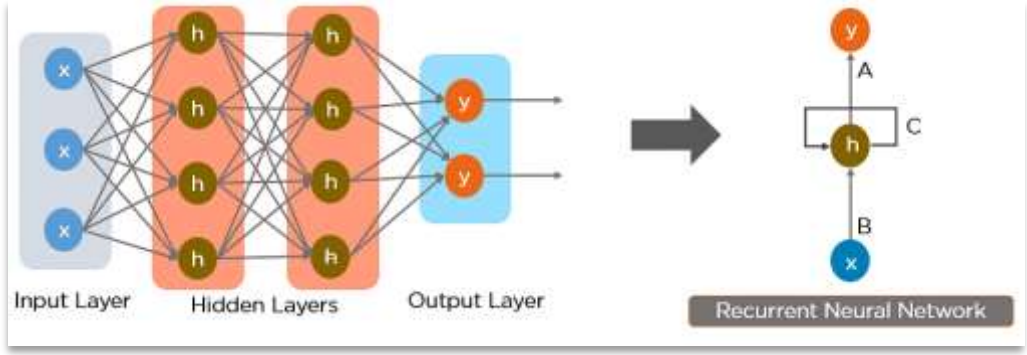




Simple Recurring Neural Network (RNN)

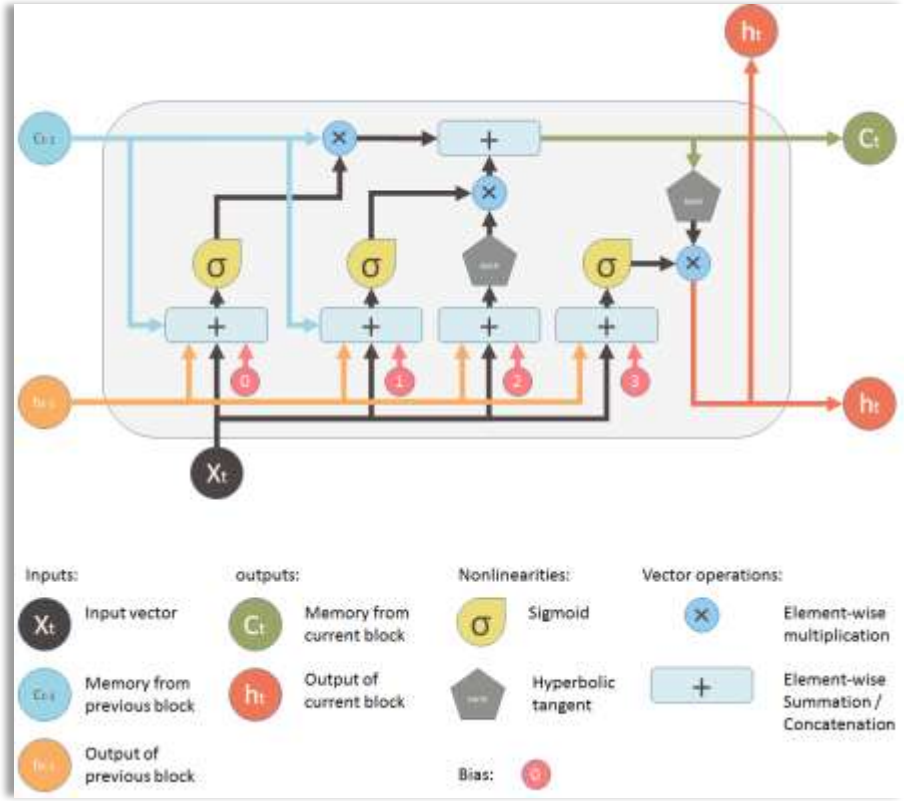
- A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data
- They are distinguished by their “memory” as they take information from prior inputs to influence the current input and output
- While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of RNNs depend on the prior elements within the sequence
- Another distinguishing characteristic of RNN is that they share parameters across layers



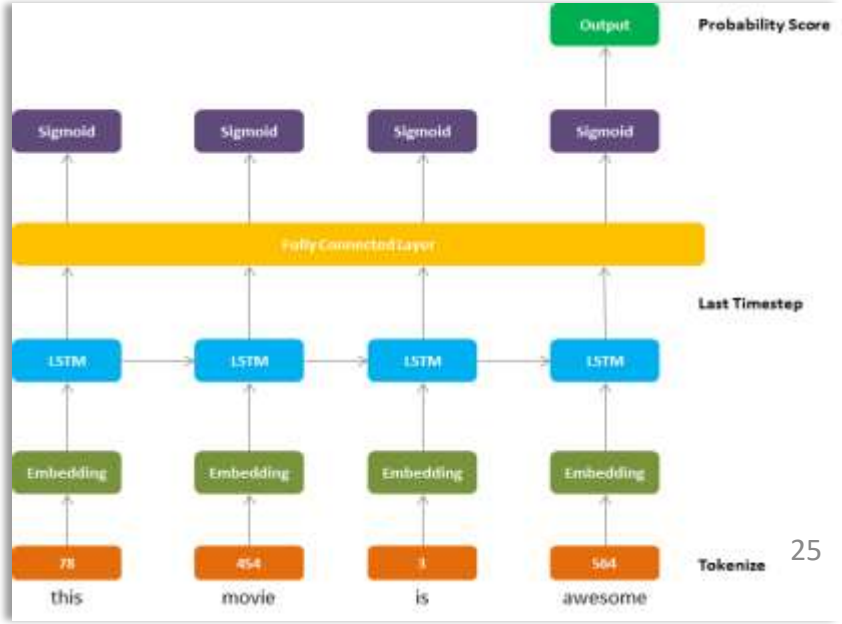


Long short-term memory (LSTM)

- Long Short-Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies
- A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate
- The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell

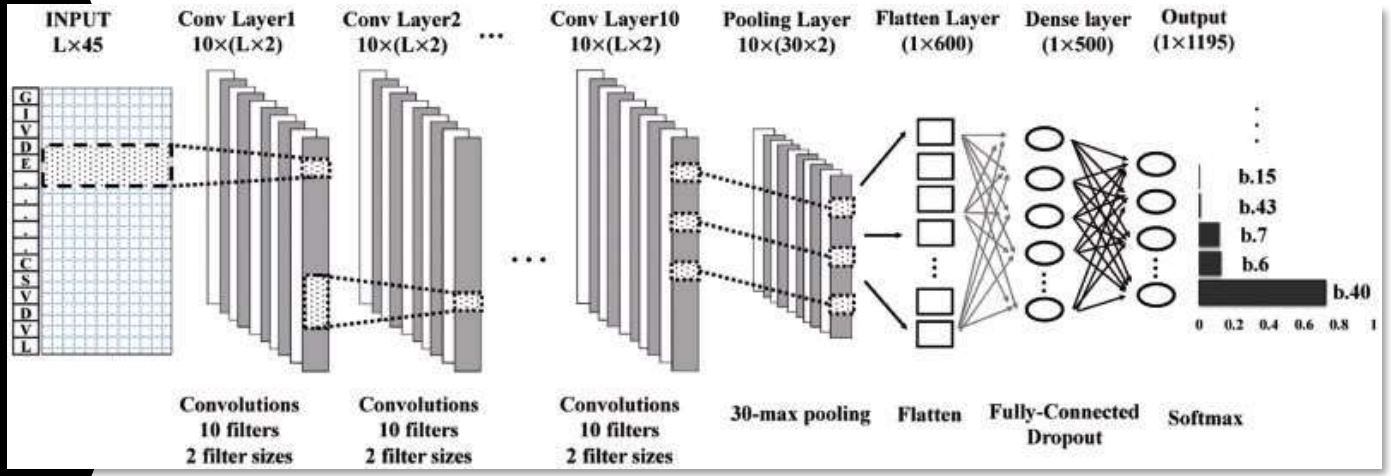


Model Architecture Illustration



1-Dimensional Convolutional Neural Network (CNN)

- CNN is a class of artificial neural network most applied to analyze visual imagery
- CNNs use shared-weight architecture of the convolution kernels that slide along input features and provide translation-equivariant responses known as feature maps
- CNNs are regularized versions of of multi-layer perceptrons (MLP)
- Although CNNs are mainly used for image data, they can also be applied to text data, as text also has adjacency information



Model Architecture Illustration

Model Results

Model	# Layers	Layer Type	Max Token Length	Vocab Size	Embedding size	Epochs	Accuracy
Deep Learning Neural Network	5	Embedding, Flatten, Dense, Dropout, Dense	1554	210440	50	2	97.09%
1-D CNN	5	Embedding, Conv1D, MaxPooling1D, Flatten, Dense	1554	210440	32	2	97.27%
LSTM	6	Embedding, LSTM(32), Flatten, Dense, Dropout, Dense	1554	210440	50	1	94.94%
RNN	5	Embedding, Masking, SimpleRNN, Dense, Dense	1554	210440	50	2	94.91%

- In comparison to the baseline models, all neural network models tend to perform better than the best baseline model on the test dataset, achieving high accuracy and precision
- Among the NN models, 1-D CNN performs the best on the test data achieving a 97.27% accuracy, closely followed by the deep leaning NN model (97.09% accuracy)

Model Output

“ I enjoy vintage books and movies so I enjoyed reading this book. The plot was unusual. Don't think killing someone in self-defense but leaving the scene and the body without notifying the police or hitting someone in the jaw to knock them out would wash today. Still it was a good read for me. ”

Original Classification - “Positive”

Prediction Score – 0.9987

Predicted Classification - 

“ The arabian fantasy, slavegirl trained to seduce and submit herself. Not to my taste, too much sex and not very erotic. Kind of mechanical..... “

Original Classification - “Negative”

Prediction Score – 0.3995

Predicted Classification - 

NEXT STEPS

Implementation & Roadmap – Topic Modeling

- Facilitate **special 'Book Series' launches** on the Kindle platform as this will drive interest e.g., Make all the 'Harry Potter books' on kindle

- Create a display on the Kindle App to view the books by the **Top 10 categories** (basis customer ratings) as mentioned in previous slide

- Identify which authors are **in the news** and try to market them. For e.g., if a new book by Roald Dahl is due to come next month, market all previous Roald Dahl books on the kindle platform

- Organize **'Kindle exclusive' book launches** to attract huge interest from the reading community

- Organize **'Kindle Talks'** where people discuss what they might have recently read so that more people share their experiences amongst themselves

Implementation & Roadmap – Sentiment Analysis

"people will forget what you said, people will forget what you did, but people will never forget how you made them feel."

- Build an automated data visualization dashboard to regularly monitor reviews. This dashboard can be customized for internal use and for the product owner
- Provide the following metrics by product, time, author, etc., to the product owner:
  - Key level 1 sentiments such as happy, sad, angry
  - A level 2 sentiment with further sentiment classification for each of the level 1 sentiment
- Extend the text analysis to generate summary based on the review, in scenarios where the summary is too short or does not capture the sentiment of the review
- Scale the analysis across regions and languages to cater suggestions to a wider audience
- Deliver targeted marketing content & offers to reviewers, build a product recommendation system for the end user
- Assist product owners providing reviewer demographics and track customer sentiment over time to improve the recommendation system

Return on Investment (ROI)

Cost & Expenditure (Annual)

- Fixed Costs:
  - Employees - 150,000 * (2 analysts + 1 Data Scientist + 1 Manager) * 1 year = $ 0.6M
  - Infrastructure, Subscription and Software = $ 1M
- Floating Costs:
  - Marketing – $10 coupons * 50,000 * 6 months = $ 3M
- Miscellaneous Costs:
  - Estimated as 15% of total cost = $ 0.3M
- Total Cost of implementation:
  - ***$ 4.9M***

Earnings & Savings (Annual)

- Current Amazon Kindle Revenue - ~$500M
- Improved recommendation system, scaling across languages:
  - Sales increase by 5% = 500M * 5% = $ 25M
- Topic modeling recommendation for self-published books:
  - Sales increase by 10% for 15% publisher revenue = 500M * 10% * 15% = $ 7.5M
- Marketing impact:
  - Assuming typical conservative marketing impact of 2% of revenue = 2% * 500 = $ 10M
- Total Earnings and Savings:
  - ***$ 42.5M***

ROI

- **Revenue – Expenditure = $42.5M – $4.9M = $ ~37.6M annually**

POTENTIAL IMPROVEMENTS

Sentiment Analysis

1. Improve the underlying modeling techniques by performing extensive hyperparameter tuning, cross-validation and perform better pre-processing
2. Explore other machine learning techniques such as SVM, random forests, etc.,
3. Assist text analytics with data from other media such as purchase history, user demographics, marketing response & activity
4. Leverage gloVe and BERT techniques to improve prediction accuracy

Topic Modeling

1. Explore other techniques like Latent Dirichlet Allocation(LDA) and Non-Negative Matrix factorization(NMF)
2. Use 'Topic Reduction' approach to optimize number of topics generated
3. Carry out a n-grams approach to the BERTopic Model to see which works best
4. Model the data on the entire data with a higher computing power for more representative results
5. Pre-process the data with finer details to produce more accurate results

CONCLUSION

- As avid book readers in the generation of social media, we decided to analyze Amazon Kindle reviews to better understand the readers' opinions and their key topics of interest
- We filtered for reviews from June - July 2014 and implemented data pre-processing techniques on the raw dataset
- We explored supervised and unsupervised learning techniques for sentiment analysis and topic modeling respectively
- As part of topic modeling, BERTopic modeling generated more interpretable results in the form of specific themes. These themes could point us to the most popular categories in the Kindle ecosystem
- As part of Sentiment Analysis, 1-D Convolutional Neural Network (CNN) generated the best classification accuracy of 97.27%
- Given these, we would deploy BERTopic modeling and 1-D CNN model to drive our analysis and generate data-driven insights
- Advanced NLP techniques like Topic Modeling and Sentiment Analysis can thus, be used to transform the way businesses work. We have implemented these techniques in our analysis and estimated the revenue impact *($ 42.5M annualized)* through this process, outlining scope for improvement in the product and customer experience

REFERENCES

- https://medium.com/analytics-vidhya/bert-for-topic-modeling-bert-vs-lda-8076e72c602b
- https://hackernoon.com/nlp-tutorial-topic-modeling-in-python-with-bertopic-372w35l9
- https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05
- https://maartengr.github.io/BERTopic/api/bertopic.html
- https://towardsdatascience.com/sentiment-analysis-using-lstm-step-by-step-50d074f09948
- https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6
- https://machinelearningmastery.com/multinomial-logistic-regression-with-python/
- https://medium.com/@jjw92abhi/is-logistic-regression-a-good-multi-class-classifier-ad20fecf1309
- https://scikit-learn.org/stable/modules/naive_bayes.html

THANK YOU!