Sentiment Analysis

GROUP 7

Chioma Susan Nwade
Rui Fang
Nadiia Sharova
Maha Tahir

Introduction

Target Company

Our project is focused on companies that provide streaming services

Business Problem

These companies are currently looking for ways to improve the customer service of their movie platform.

Aim

Our aim is build a sentiment analysis model to assist streaming services companies to better understand how the end users of their movie platform feel in regards to operating the website and online portals.

The Model

The model will have the capability to filter reviews based on the sentiment expressed within them. By analyzing the text content of reviews, the model can determine whether the sentiment conveyed is positive, negative, or neutral.

Objective of the model is to help the streaming companies to :

- Gain clarity on what is working well within their movie recommendation system
- Enhance their movie and platform features based on feedback for where improvement are been suggested
- Eliminate content which is receiving the negative emoticons

Preprocessing

About the dataset

We will utilize the International Movie Database (IMDb) review dataset for sentiment analysis. This dataset comprises 50,000 movie reviews from a global spectrum, and it's organized for binary classification, labeling each review as either positive or negative. The dataset is divided into 25,000 samples for training and an equal count of 25,000 samples for testing. It will be accessed through the torchText library's datasets module.

Steps carried out:

- We used the torch.text library which is great tool for nlp projects
- We used the field method of the data class to decide how data needs to be preprocessed
- We Set our seed for reproducibility to 42
- We Made use of Spacy to determines how the sentences are going to be broken down or tokenized in nlp standard
- We limited the number of words the model will learn to 25000 most used words.

Libraries used for this project:

- OpenAI: To access the ChatGPT API, for practice to establish an understanding sentiment analysis.
- Pandas: Utilized for analyzing extensive datasets and drawing conclusions based on statistical theory.
- NumPy: Enabling array operations for efficient data manipulation.
- Random: Used for generating consistent results by setting seeds, ensuring reproducibility.
- Torch: Applied for implementing NLP algorithms.
- spaCy: Employed to determine sentence splitting in NLP tasks.
- Matplotlib: Utilized for creating visualizations.

Modelling

Sentiment Analysis Model

A model that evaluates and determines the emotional tone or sentiment expressed in a piece of text, categorizing it as positive, negative or neutral.

We have used 3 sets of data to develop the model:

1. Training
2. Testing
3. Validation

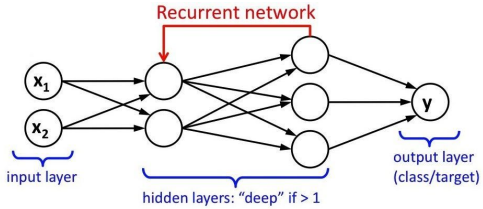
These three sets of data were fed to the model in the form of batches of 64 samples at a time.

Machine learning Algorithm

The Long short-term memory (LSTM) network (recurrent neural network (RNN)) is the structure we have applied, multi layer bidirectional rnn to be more specific.

This will allow us to build out layers on top of each other, it creates a copy of the output and loops it back through the network.

An advantage of this is that when customers are leaving feedback the following words can be suggested.

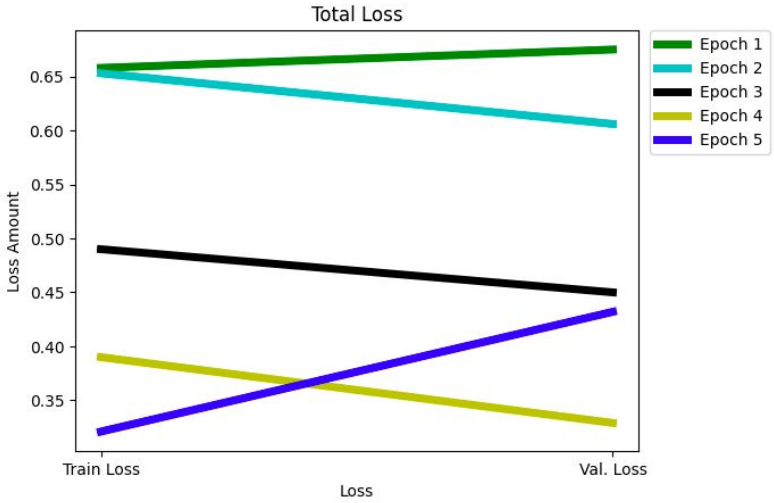


Running the Model

- A total of 5 epochs have been used for our model: the number of times the entire training dataset will be used to update the model. In this case, the loop will iterate over the dataset five times.
- The models are evaluated based on accuracy score and loss
- The ideal model is the one with the best validation loss

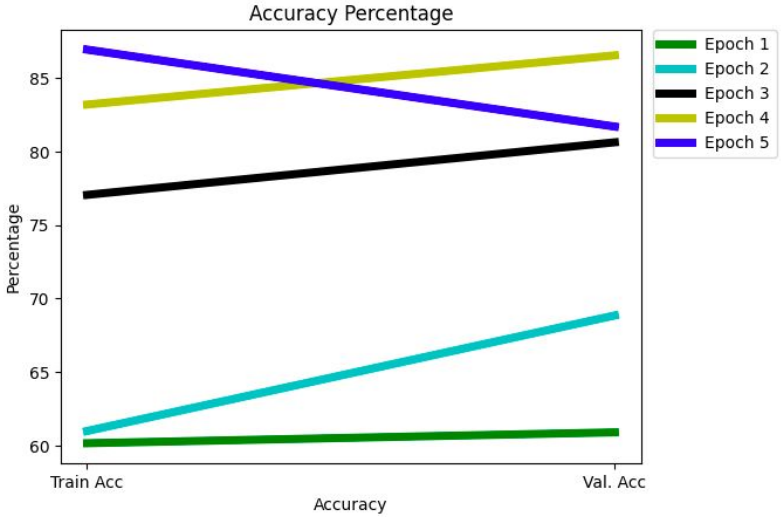
Results and Visualisation

On test Data: Our accuracy score was 85.83% while the test loss was 0.343



This image shows that ⅗ Epochs improved as their validation loss was lower than their train loss.

Epoch 4 is determined to be ideal model as its validation loss is the lowest.



their validation accuracy was higher than their train accuracy.

Epoch 5 is the only Epoch which dropped its validation accuracy.

Improvement

As there is always room for improvement, and staying up-to-date is equally important, we have assembled a few ideas:

1. Remove excess noise, as it is not recognized:

   - Punctuation

   - Stopwords

   - Emojis

2. Convert the data to lowercase or stem it.

3. Lemmatize your words, as slang is not recognized.

4. Add additional features, such as:

   - Part-of-speech tags

   - N-grams

   - Sentiment lexicons

5. It's important to note that feedback can introduce bias, which might lead to results that are not always 100% accurate.

6. Since data is perpetually evolving, this can lead to challenges with the current codebase. These challenges can manifest as sluggish performance or bugs in the system. Consequently, maintaining and continually updating the sentiment analysis model and its algorithm can be quite demanding.

Recommendation

We would recommend using this model the model be used to:

- Quickly identify and prioritise positive or negative feedback.
- Develop a stronger connection with customers which will in turn increase the company bottom line.
- Monitor your branding strategies.
- Incorporate changes and delete unwanted features.
- Develop a robust recommendation system.
- Customise service responses, Feedback techniques or improvement plans.
- Access real-time insights on customer preferences.

Conclusion

We have successfully developed a sentiment analysis model based on LSTM techniques, known for their high accuracy. Our model is easy to replicate, reuse, and modify. It can be trained on various types of data by making slight adjustments to the dataset, as demonstrated in our project.

Thank you

We hope you enjoyed our presentation