AMAZON PRODUCTS REVIEW SENTIMENT ANALYSIS

IBM Machine Learning Professional Certificate

Course 05: Deep Learning and Reinforcement Learning

By Moustafa Abada

Content:

- Dataset Description
- Main objectives of the analysis.
- Applying classification models.
- Model analysis and findings.
- Models flaws and advanced steps.

DATASET DESCRIPTION

Dataset Description

- The dataset used for this project on Amazon product sentiment analysis consists of more than 400,000 reviews of unlocked mobile phones sold on Amazon.com. The dataset includes information such as the product title, brand, price, rating, review text, and the number of people who found the review helpful. The analysis focuses on exploring insights related to reviews, ratings, price, and their relationships. The dataset covers a wide range of nearly 4,400 unlocked mobile phones within the 'unlocked phone' category on Amazon.com.

Dataset Description

- The dataset contains the following columns:

1. **Product Title**: The title or name of the unlocked mobile phone product.
2. **Brand:** The brand or manufacturer of the unlocked mobile phone.
3. **Price:** The price of the unlocked mobile phone.
4. **Rating:** The rating given by the reviewer on a scale of 1 to 5 for the product.
5. **Review text:** The written review or feedback provided by the reviewer.
6. **Number of people who found the review helpful:** The count of individuals who marked the review as helpful.

Amazon Unlocked Mobile Dataset

```
# Load csv file
df = pd.read_csv('Amazon_Unlocked_Mobile.csv')
df.head()
```

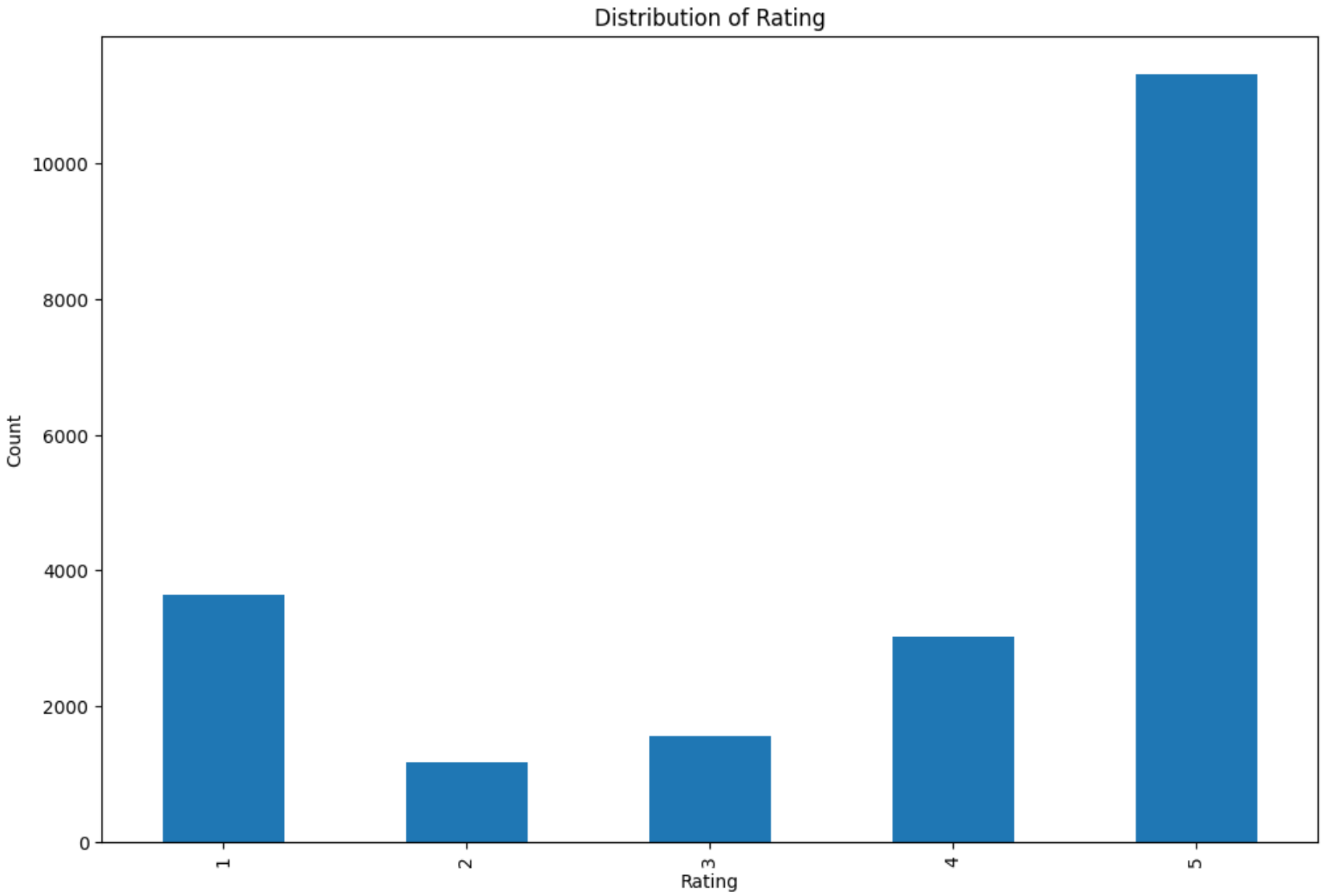
	Product Name	Brand Name	Price	Rating	Reviews	Review Votes
0	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	I feel so LUCKY to have found this used (phone...	1.0
1	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	nice phone, nice up grade from my pantach revu...	0.0
2	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	Very pleased	0.0
3	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	It works good but it goes slow sometimes but i...	0.0
4	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	Great phone to replace my lost phone. The only...	0.0

DATA ANALYSIS

Main objectives of the analysis:

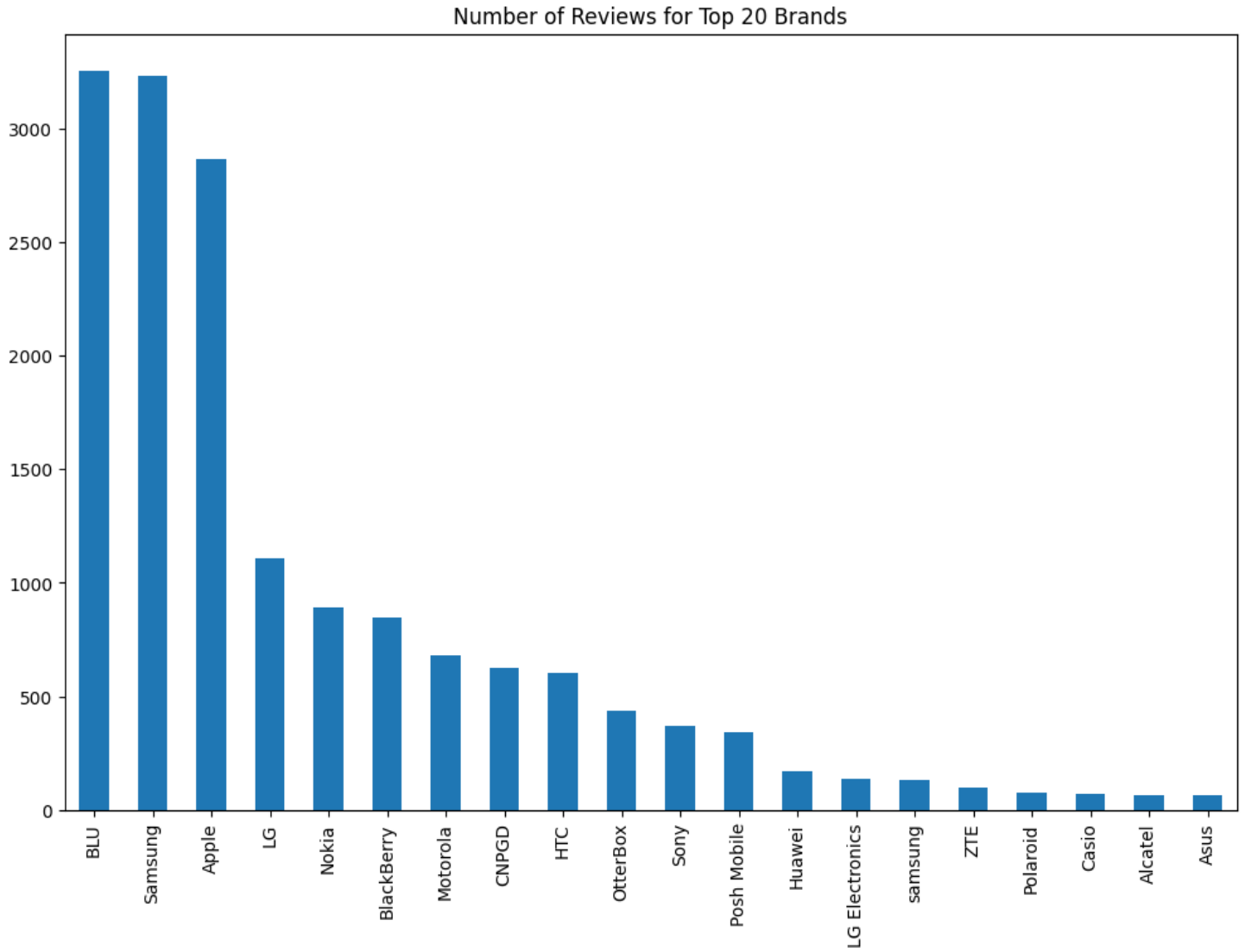
- This analysis has the following goals:

1. Perform exploratory analysis of ratings and reviews.
2. Find top brands and top products sold.
3. Word cloud of most-used words.
4. Sentiment analysis classification using LSTM.



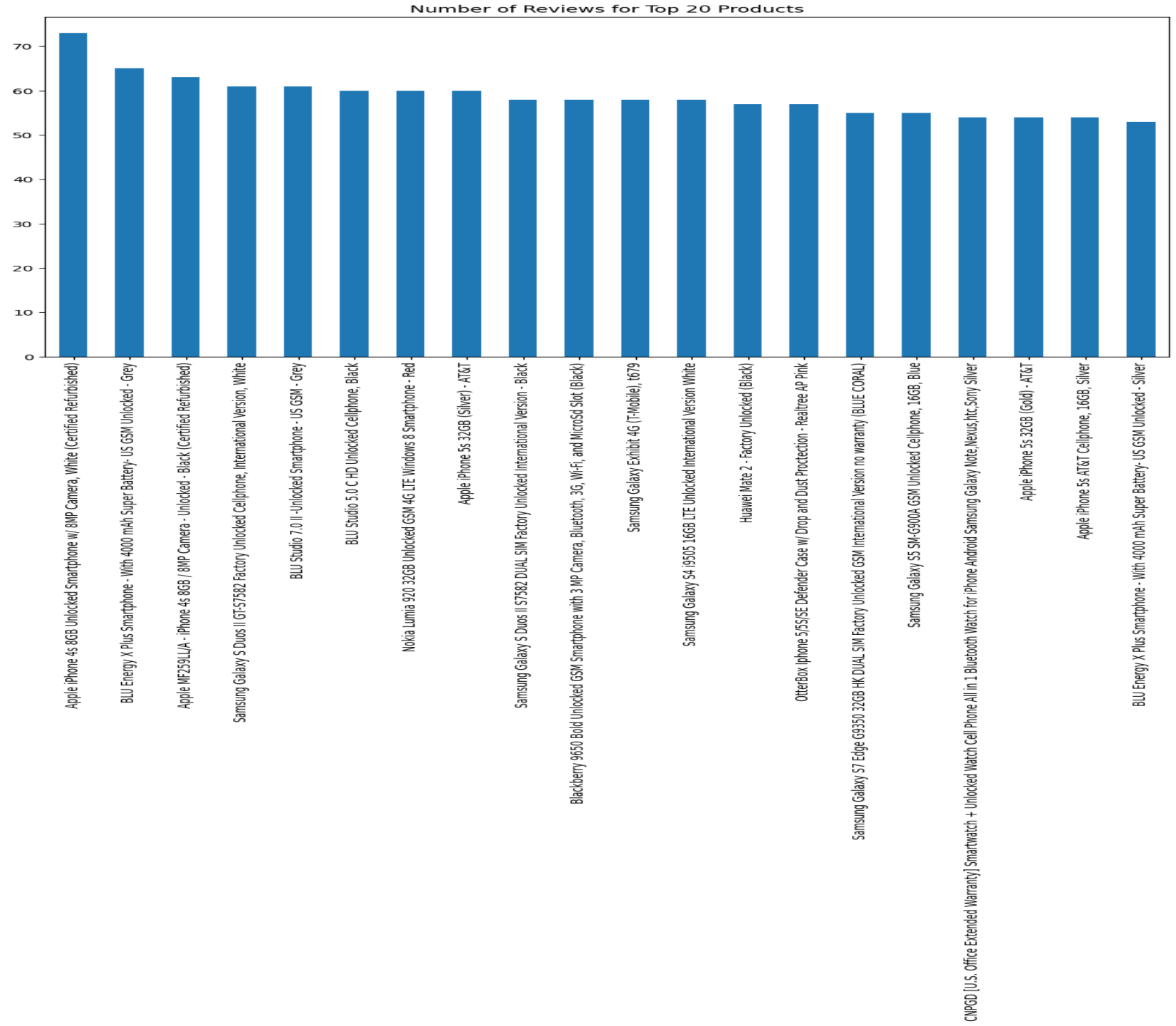
DATA ANALYSIS:

Bar Chart for the Reviews



DATA ANALYSIS:

Top 20 Brands Reviews



DATA ANALYSIS:

Top 20 Products Reviews

DATA ANALYSIS:

Word cloud of most-used words in reviews.

```
[ ] # Drop missing values
    df.dropna(inplace=True)

    # Remove any 'neutral' ratings equal to 3
    df = df[df['Rating'] != 3]

    # Encode 4s and 5s as 1 (positive sentiment) and 1s and 2s as 0 (negative sentiment)
    df['Sentiment'] = np.where(df['Rating'] > 3, 1, 0)
    df.head()
```

```
<ipython-input-54-d1b50b334b28>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Sentiment'] = np.where(df['Rating'] > 3, 1, 0)
```

	Product Name	Brand Name	Price	Rating	Reviews	Review Votes	Sentiment
5	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	1	I already had a phone with problems... I know ...	1.0	0
15	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	2	Had this phone before and loved it but was not...	0.0	0
36	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	Met all of my expectations. I can't complain a...	0.0	1
37	"Nokia Asha 302 Unlocked GSM Phone with 3.2MP ...	Nokia	299.00	5	Phone is working on, I was planning to use it ...	0.0	1
45	"Nokia Asha 302 Unlocked GSM Phone with 3.2MP ...	Nokia	299.00	5	I love it!	0.0	1

The following data processing techniques are performed:

1. Drop Rows with Null Values.
2. Drop unimportant features or columns such as Product Name, Brand Name, Price, Review Votes.
3. Add a new column called “Sentiment” that is 1 when the rating is 4 or 5 and 0 when the rating is 1 or 2. And, neutral columns with rating 3 are dropped.
4. Columns “Reviews” and “Sentiment” are used for classification with LSTM.
5. Only a small portion of the dataset is used because the dataset is so large!!

APPLYING CLASSIFICATION MODEL

```
[ ] # Split the dataset into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(df['Reviews'].values, df['Sentiment'].values, test_size=0.2, random_state=42)
```

```
[ ] # Tokenization
    tokenizer = Tokenizer(num_words=100000)  # Consider only the top 100,000 words
    tokenizer.fit_on_texts(X_train)
    X_train_seq = tokenizer.texts_to_sequences(X_train)
    X_test_seq = tokenizer.texts_to_sequences(X_test)

    # Padding
    max_seq_len = max(len(seq) for seq in X_train_seq)
    X_train_pad = pad_sequences(X_train_seq, maxlen=max_seq_len)
    X_test_pad = pad_sequences(X_test_seq, maxlen=max_seq_len)
```

```
[ ] # Word2Vec embedding
    word2vec_model = Word2Vec(sentences=X_train_seq, vector_size=128, window=5, min_count=1, workers=4)
    word2vec_weights = word2vec_model.wv.vectors
```

The following techniques are performed:

1. Train-Test Split.
2. Tokenizing the X-train data.
3. Padding the sequences.
4. Word2vec embeddings is performed to convert the data to embeddings to learns word representations from the text data and capture the meaningful relationships between words based on their co-occurrence patterns.

```
[ ] # Model creation
    model = Sequential()
    model.add(
        Embedding(
            input_dim=vocab_size,
            output_dim=word2vec_weights.shape[1],
            weights=[word2vec_weights],
            input_length=max_seq_len,
            trainable=False,
        )
    )
    model.add(LSTM(128))
    model.add(Dense(1, activation='sigmoid'))

    # Compiling the model
    model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
[ ] # Model training
    model.fit(X_train_pad, y_train, validation_data=(X_test_pad, y_test), epochs=5, batch_size=64)

    # Evaluating the model
    loss, accuracy = model.evaluate(X_test_pad, y_test)
    print(f'Test loss: {loss:.4f}')
    print(f'Test accuracy: {accuracy:.4f}')
```

```
Epoch 1/5
194/194 [==============================] - 853s 4s/step - loss: 0.4633 - accuracy: 0.7799 - val_loss: 0.4090 - val_accuracy: 0.8113
Epoch 2/5
194/194 [==============================] - 843s 4s/step - loss: 0.3839 - accuracy: 0.8266 - val_loss: 0.3753 - val_accuracy: 0.8258
Epoch 3/5
194/194 [==============================] - 844s 4s/step - loss: 0.3468 - accuracy: 0.8429 - val_loss: 0.3309 - val_accuracy: 0.8553
Epoch 4/5
194/194 [==============================] - 842s 4s/step - loss: 0.3063 - accuracy: 0.8666 - val_loss: 0.3195 - val_accuracy: 0.8675
Epoch 5/5
194/194 [==============================] - 845s 4s/step - loss: 0.2822 - accuracy: 0.8792 - val_loss: 0.3669 - val_accuracy: 0.8401
97/97 [==============================] - 95s 979ms/step - loss: 0.3669 - accuracy: 0.8401
Test loss: 0.3669
Test accuracy: 0.8401
```

The embeddings are fed into the LSTM Model and the training process is done. The model has a test accuracy 84% which is reasonable given the usage of small portion of the dataset to avoid large time of training.

Test Sample

As shown below, some reviews are shown with the predicted label and the original or true label.

1 represents **Positive** rating and **0** represents **Negative** rating.

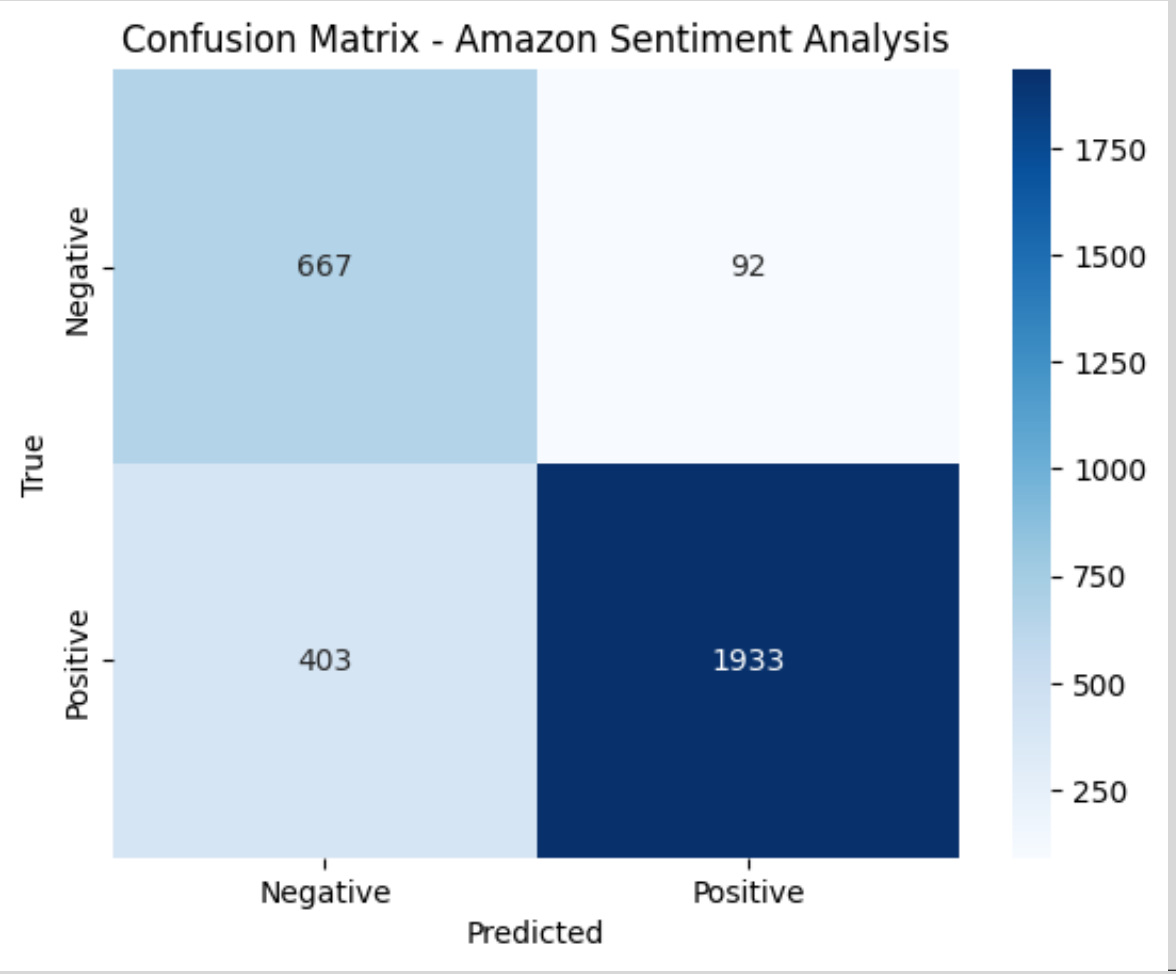
```
1/1 [==============================] - 0s 346ms/step
Test Review: Excellent Recommended works well in VENEZUELA
Original Label: 1
Predicted Label: Positive

1/1 [==============================] - 0s 280ms/step
Test Review: Excellent
Original Label: 1
Predicted Label: Positive

1/1 [==============================] - 0s 194ms/step
Test Review: After few weeks of use phone broke down, battery wasn't charging all the way,
Original Label: 0
Predicted Label: Negative
```

MACHINE LEARNING ANALYSIS AND FINDINGS

Model Analysis and Findings:

- In the model analysis, several techniques were applied to analyze the data. First, a train-test split was performed to divide the dataset into training and testing subsets. Next, the X-train data was tokenized, which involved breaking down the text into individual words or tokens. To ensure consistent input shapes, padding was applied to the sequences. Additionally, Word2Vec embeddings were used to convert the text data into vector representations, capturing meaningful relationships between words based on co-occurrence patterns. These embeddings were then fed into the LSTM model for training. Despite using a small portion of the dataset to minimize training time, the model achieved a reasonable test accuracy of 84%. This suggests that the model was able to learn and generalize patterns effectively, showcasing its potential for further analysis and predictions.

Confusion Matrix of the Model



MODELS FLAWS AND ADVANCED STEPS

Model Flaws and Advanced Steps:

- One potential flaw in the model that the model only used small portion of the dataset as the dataset is so large and it will take so much time to train. Also, this dataset is only for amazon products, maybe in the future, I will consider using a diverse dataset and add movies reviews dataset. To address these flaws and improve the model, several advanced steps can be taken:

1. Incorporating a more sophisticated sentiment analysis technique, such as using a pre-trained language model like BERT or using advanced techniques like aspect-based sentiment analysis, can provide more accurate sentiment predictions.
2. Gathering a larger and more diverse dataset, perhaps including reviews from multiple sources or domains, can help improve the model's generalization capabilities.
3. Implementing techniques like cross-validation, hyperparameter tuning, and model ensembling can enhance the model's robustness and performance.

THANK YOU!

IBM Machine Learning Professional Certificate

Course 05: Deep Learning and Reinforcement Learning

By Moustafa Abada