Task 21 - Capstone Project - NLP Applications

DESCRIPTION OF THE DATASET USED

Consumer Reviews of Amazon Products Dataset

The dataset is a list of 5,000 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database. The dataset includes basic product information, rating, review text, and more for each product. This data can be downloaded directly from here. In more specific terms the dataset has 5000 rows and 24 columns with the following data types – 1 Boolean, 1 float, 2 integer, and 20 object datatypes.

Of the 24 columns, 'reviews.dateAdded' column has 3948 missing entires, reviews.id has 49711 missing entries,' reviews.title' has 13 missing entries, and 'reviews.username' has 1 missing entry. The column of interest for this study is the 'reviews.text' column and fortunately has no missing entries.

The purpose of this research is to carryout sentiment analysis on the reviews provided of the product provided by Amazon to her customers and investigate the similarity score of some of the reviews to help amazon make informed decisions on their products based on consensus on products from the reviews.

DETAILS OF THE PREPROCESSING STEPS

Start:

Phase 1: Data Preparation

- Importing modules
  - Import NumPy
  - Import pandas
  - Import spacy
- Download the amazon dataset
- View and explore the first 5 rows of the dataset
- Explore the features and datatypes of the dataset
- Explore missing data of the 24 features in the dataset
- Narrow down feature extraction to the most relevant features for sentiment analysis
- Explore the dataset for missing data in the extracted features
- Clean the extracted features by dropping the entries with missing data
- Validate that the entries with missing data are dropped

Phase 2: Defining Functions and Preprocessing

- Define a function preprocess( ) to lemmatise, strip, remove words with less meaning and punctuations
- Create a new feature (processed_reviews), apply the function – preprocess( ), and assign it to processed_reviews

- View the cleaned and extracted dataset to have look at the new feature created with the processed strings or tokens
- Next, we download the spaCy language model, apply it to the preprocessed feature and tokenize in the following steps:
  - Download the spacy language model
  - Convert the processed reviews text into spaCy Doc objects and render to explore more
  - Iterate over each Doc object and print token information using a for loop
    - Render the processed docs with displacy
  - Performing Named Entity Recognition
    - Iterate over each Doc object and access named entities in further exploration

Phase 3: Sentiment Analysis

- Word Cloud: Positive and Negative Sentiments
  - Download and Install wordcloud
  - Import wordcloud
  - Import matplotlib
  - Import defaultdict – default dictionary
  - Declare two dictionary variables for positive and negative words
  - Install spacytextblob model
  - Install textblob.download_corpora and TextBlob to use .sentiment and .polarity
  - Create a graph image of positive and negative word clouds
- Define a function for Sentiment Analysis
  - Import DOC from spacy.tokens
  - Define a function to add custom extension attributes to each Doc
  - Create a TextBlob object for the text of the document
  - Set the polarity and sentiment attributes using the TextBlob object
  - Return a doc
- Register the custom extension attribute
- Because the Extension 'blob' already exists on Doc.
- To overwrite the existing extension, we set `force=True` on `Doc.set_extension
- Access the text data from the Pandas Series object and iterate over each text – for loop
  - Process the text and get the Doc object
  - Process the text and get the Doc object using the spaCy model – nlp( )
  - Calculate the sentiment polarity and subjectivity scores using TextBlob
  - Print the polarity and subjectivity scores for each text

Phase 4: Similarity Score

- Finding Similarity between a pair of reviews using – small – medium – large spaCy models
  - Load the respective spaCy model
  - Select two reviews for comparison
  - Process the text of each review
  - Compute the similarity score between a pair of reviews

End

EVALUATING SEMTIMENT SCORE RESULT

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment expressed in a piece of text. It involves analysing text data to identify and extract subjective or factual information, such as opinions, sentiments, and climate change by individuals or entities.

The primary goal of sentiment analysis is to classify the sentiment of a piece of text as positive, negative, or neutral. However, sentiment analysis can also involve more nuanced analysis, such as identifying specific emotions (e.g., happiness, anger, sadness) or polarity - assessing the intensity of sentiment.

Sentiment analysis can be applied to various types of text data, including product reviews, social media posts, news articles, customer feedback, and survey responses.

Sentiment analysis enables organizations to extract valuable insights from unstructured text data, allowing them to better understand customer sentiment, identify emerging trends, and make data-driven decisions to drive business success. These can be achieved through the use of metric scores.

In sentiment analysis, a metric score is a numerical value that quantifies the sentiment expressed in a piece of text. These scores are typically computed using various techniques and algorithms to analyse the textual content and determine the polarity or subjectivity of sentiment.

- A subjectivity score is a measure of the degree to which a piece of text expresses subjective or objective content. Subjectivity refers to the extent to which the text reflects personal opinions, feelings, or beliefs, rather than factual information.

- Subjectivity scores are often calculated in the context of sentiment analysis, which aims to determine the sentiment or attitude conveyed by a piece of text. While polarity measures the positivity or negativity of the sentiment, subjectivity measures how much of the text is opinionated or subjective in nature.

- Subjectivity scores typically range from 0 to 1, where a score closer to 0 indicates more objective content (i.e., factual information), while a score closer to 1 indicates more subjective content (i.e., personal opinions or feelings).

- In the context of sentiment analysis, subjectivity scores can help differentiate between texts that are purely factual and those that contain subjective expressions or opinions, providing additional insights into the overall sentiment conveyed by the text.

Summarily:

Positivity or Negativity of a sentiment - Polarity

Polarity of 1 = Positive Sentiment

Polarity of -1 = Negative Sentiment

Polarity of 0 = A Neutral Sentiment

Subjectivity Score ranges from 0 to 1:

- 0 indicates more objective content (i.e., factual information)
- 1 indicates more subjective content (i.e., personal opinion or feelings)

1. Negative Sentiment (Polarity) and more subjective content (subjectivity).

- Text_1: little confusing frustrating get verification code amazon wait 20 minute request code nother start set device 8.30 11.20 feed step away 5 hour go shop mall recieve 4 code amazon
- Sentiment: Sentiment(polarity=-0.29583333333333334, subjectivity=0.6)

- Text_2: small think get job
- Sentiment: Sentiment(polarity=-0.25, subjectivity=0.4)

- Text_3: pay playtime work want go long trip video download issue play time son 2 get bored quickly pro price protect externally
- Sentiment: Sentiment(polarity=-0.05416666666666668, subjectivity=0.5)

Discussion 1:

Looking at Texts in 1 above, we conclude that the classification of a negative sentiment and the sentiment content is the opinion of the reviewer and not necessarily a factual statement is accurate.

2. Positive Sentiment (Polarity) and more subjective content (subjectivity).

- Text: great product buy great deal
- Sentiment: Sentiment(polarity=0.8, subjectivity=0.75)

- Text: kindle perfect e reader easy use size perfect bring book
- Sentiment: Sentiment(polarity=0.8111111111111112, subjectivity=0.94444444444445)

- Text: excellent lightweight way lot read material travel
- Sentiment: Sentiment(polarity=1.0, subjectivity=1.0)

Discussion 2:

Looking at Texts in 2 above, we conclude that the classification of a Positive sentiment and the sentiment content is the opinion of the reviewer and not factual is accurate. The last text has both the polarity and subjectivity with a maximum score of 1. Looking at the text, excellent is an absolute positive sentiment and the opinion of the reviewer that the product is handy as a travel material is absolutely a strong opinion and correct as well.

3. Neutral Sentiment (Polarity) and more subjective content (Subjectivity).

- Text: kindle product fantastic kindle kindle fire use everyday naturally select kindle granddaughter
- Sentiment: Sentiment(polarity=0.10000000000000002, subjectivity=0.6333333333333)

- Text: love tablet kindle access email facebook play simple game shop online read book
- Sentiment: Sentiment(polarity=0.03333333333326, subjectivity=0.4523809523809524)

Discussion 3:

Looking at Texts in 3 above, one would be tempted to say that the review is a positive sentiment giving the use of fantastic. But the classification of the sentiment can be termed to be accurate as it is has more leaning towards neutral sentiment. My observation is that I was unable to see a neutral review that is factual in terms of subjectivity and most of the neutral reviews are near neutral but not actually neutral. I guess this could change with more data for the analysis. However, I have observed also that using the medium and large language model has not improved the results here.

4. Negative Sentiment (Polarity) and more factual content (subjectivity).

- Text: load book proper take dozen try erase dreregistere screen dark
- Sentiment: Sentiment(polarity=-0.075, subjectivity=0.25)

- Text: screen dark adjust brightness
- Sentiment: Sentiment(polarity=-0.15, subjectivity=0.4)

- Text: average alexa option thing screen limited
- Sentiment: Sentiment(polarity=-0.11071428571428571, subjectivity=0.271428571428)

- Text: qc bad product catch voice properly
- Sentiment: Sentiment(polarity=-0.3499999999999999, subjectivity=0.3833333333333)

- Text: download slow echo didn't worth have stand
- Sentiment: Sentiment(polarity=-2.7755575615628914e-17, subjectivity=0.25)

- Text: useless work amazon cloudcam purchase best buy allow monitor cloud cams kitchen tech savvy numerous alexa device unable work try instal enable amazon cloudcam skill sure new update automatically update plug final straw tech support rep tell can't use cloudcam,äîand unable connect actually solve problem useless help useless actually make work echo useless
- Sentiment: Sentiment(polarity=-0.10489510489510491, subjectivity=0.395648795648)

Discussion 4:

Looking at Texts in 4 above, the sentiments are often negative and leaning towards factual in terms of subjectivity. This can be seen to be accurate and worrisome as customers are not happy with the product and could affect the image of the brand that the company is selling. Especially the review that says "download slow echo didn't worth have stand". The characterisation of the review as being factual by 0.38 of subjectivity score is not good for the brand.

5. Positive Sentiment (Polarity) and more factual content (subjectivity).

- Text: alexa fun play family fill house music tell joke
- Sentiment: Sentiment(polarity=0.3, subjectivity=0.2)

- Text: wife love alexa early christmas present
- Sentiment: Sentiment(polarity=0.19999999999999998, subjectivity=0.3)

- Text: surprisingly useful ben 2 echo well gen 1 definitely recommend

- Sentiment: Sentiment(polarity=0.15, subjectivity=0.25)

- Text: fun item helpful thing home
- Sentiment: Sentiment(polarity=0.3, subjectivity=0.2)

Discussion 5:

Looking at Texts in 5 above, although the reviews have positive sentiments, the sentiment score of the reviews are mostly less than 0.4 which is around neutral and the subjectivity tends to be factual. This could imply that customers are not particularly impressed about the performance of the product and its actually a fair assessment.

6. Neutral Sentiment (Polarity) and more factual content (subjectivity).

- Text: fire tv prime membership amazon music subscription home automation device alexa app phone system functional
- Sentiment: Sentiment(polarity=0.0, subjectivity=0.0)

- Text: want kindle decide bb sale disappoint
- Sentiment: Sentiment(polarity=0.0, subjectivity=0.0)

- Text: gift parent like
- Sentiment: Sentiment(polarity=0.0, subjectivity=0.0)

- Text: item work
- Sentiment: Sentiment(polarity=0.0, subjectivity=0.0)

Discussion 6:

Looking at Texts in 1 above, the sentiment scores are both absolutely positive and factual and the reviews seem to corroborate the sentiment score and it is observed to be true.

EVALUATING SIMILARITY SCORE RESULT

Similarity Score of 1 = Two reviews are similar

Similarity Score of 0 = Two reviews are dissimilar

In natural language processing (NLP), the similarity score is a metric used to quantify the similarity between two pieces of text, such as sentences, documents, or tokens. It measures how closely related or similar the meaning of the texts is.

There are various methods to calculate similarity scores, with one common approach being to use word embeddings. Word embeddings represent words as dense vectors in a high-dimensional space, where words with similar meanings are closer to each other in the vector space.

One popular technique for calculating similarity scores using word embeddings is cosine similarity. Cosine similarity measures the cosine of the angle between two vectors, which indicates the similarity in direction between them. A higher cosine similarity score suggests greater similarity between the texts.

Review.Text at Index [0]	Review.Text at Index [1]	
I thought it would be as big as small p aper but turn out to be just like my pal m. I think it is too small to read on it... not very comfortable as regular Kindle . Would definitely recommend a pape rwhite instead..	This kindle is light and easy to use especially at the be ach!!!.	
SIMILARITY SCORE		
Small Language Model	**Medium Language Model**	**Large Language Model**
0.5644872632355882	0.757820996637799	0.7676570057405966
Review.Text at Index [2]	**Review.Text at Index [3]**	
Didn't know how much I'd use a kindle so went for the lower end. I'm happy with it, even if it's a little dark.	I am 100 happy with my purchase. I caught it on sale at a really good price. I am normally a real book person, but I have a 1 year old who loves ripping up pages. The Kindle prevents that, it's extremely portable (it fits better in my purse than a giant book), and I have it loaded with lots of books. I finish one and start another, without having to go store. It serves all my needs. I picked this one over the Paperwhite because the price was unbeatable and the only difference that I could see was this one wasn't backlit. A simple book light from the Dollar tree solves that issue. This is my second Kindle (the first being the old Keyboard model, which I put down because I fell out of love with the keyboard. Lol) and it most likely won't be my last..	
SIMILARITY SCORE		
Small Language Model	**Medium Language Model**	**Large Language Model**
0.5794955496962099	0.852867287341213	0.8529098273420891

Table 1: Similarity score on pairs of review text on different spaCy models.

Discuss the similarity score results:

In order to discuss the similarity score results, a pair of reviews were considered as shown in the table above. These pairs of reviews were tested for similarity scores on the small, medium, and large spaCy models. The results indicated that there are varied scores in each comparison that indicated each pair of review are similar to each other.

The similarity between the first pair of reviews the small spaCy model is 0.5644872632355882, which suggests that there is 50% similarity between the pair of reviews. The similarity score improved its performance to 0.757820996637799 when the medium spaCy model was used to compare the first pair of reviews. This improvement in performance is significant when compared to the increase in performance between the medium spaCy model and the large spaCy model. The variance in performance between medium spaCy and large spaCy model is 0.0098360091027976. This is a valuable piece of information when a user caught up between CPU performance and accuracy of the result.

Looking at the second pair of review it will be safe also to say that, like in the first pair, there is a significant increase in the performance of the medium spaCy model over the small spaCy model and a 50% similarity score between the second pair of reviews. Conversely, the variance between the medium and large spaCy model is very small.

In summary the inference that can be drawn from this experiment is that the bigger the training model the better the similarity score and better performance. Also, the availability of computational resources can influence the spaCy model to deploy. When the resources are limited one could get a fairly accurate similarity score by using medium over large spaCy model.

INSIGHT INTO THE MODEL'S STRENGTH AND LIMITATIONS

In summary the inference that can be drawn from this experiment is that the bigger the training model the better the similarity score and better performance. Also, the availability of computational resources can influence the spaCy model to deploy. When the resources are limited one could get a fairly accurate similarity score by using medium over large spaCy model.

In terms of the sentiment analysis, the performance is accurate and there is nothing that seem to suggest that the choice of spaCy model will affect the sentiment analysis.