SENTIMENT ANALYSIS

WITH SPACY AND NLP

Task 21 – Capstone Project – NLP Applications.

Developing a Python program that performs sentiment analysis on an Amazon dataset of product reviews.

Report:

1. **The dataset description:**

   The dataset lists over 34,000 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.

   Dataset link: Consumer Reviews of Amazon Products (kaggle.com) (97.2KB).

2. **Preprocessing steps:**

   Step 1: Declare a function to remove the stop words of the dataset.

   Step 2: Applying the *.dropna* method to remove all missing values.

   Step 3: Applying the data clean function to finish the preprocessing steps.

3. **Evaluation of results:**

   As a result of the evaluation, by applying the *sentiment_analysis* function, we return the overall sentiment as calculated by SpacyTextBlob class.

   The code then creates two datasets by filtering the original one for positive and negative analysis. These two datasets are counted, and it calculates the percentage of the positive and negative analysis.

   The final result demonstrates that Amazon customer's reviews of the analysed dataset, have a positive sentiment (88,8%) about the analysed products.

4. **Analysing the accuracy:**

   To analyse the accuracy, I selected five random reviews from the cleaned dataset, to manually determine if the associated sentiment was as expected.

   As a result the code output:

ID	Reviews_text	Score (sentiment_analysis)
100	I have one Alexa and three Echo dots and having Echo Show now is **awesome**!	1.0
4456	**Awesome**.. Everything I need and more. I'm **very happy** with the tablet. I'm **very happy**. Thank you, Kathleen Stroyek	0.8666666666666667
919	Got this for two days during the holidays. **Very good** service	0.7
223	The ease of use and comparability with Smart Home products make a this an **awesome** product.	0.6071428571428571
4323	My first tablet was a Kindle. I was curious about the updated version. **Very disappointed**	-0.425

As we can see from the results, the sentiment analysis system is accurate, returning the appropriate score to the reviews.

I've highlighted in bold words that were probably key in determining the sentiment of it review.

5. **Insights into the model's strengths and limitations:**

From what we could see, spacy analyses the text by detecting relevant n-grams [1]and calculating its polarity. The advantage of using n-grams is that we analyse words within a context instead of as separate words.

Still analysing words as n-grams may be susceptible to not correctly describing colloquial expressions in which the words have a different meaning.

6. **References:**

About sPacy: https://spacy.io/ https://spacy.io/universe/project/spacy-textblob

About Pandas: https://pandas.pydata.org/

[1] N-gram: An n-gram is a collection of n successive items in a text document that may include words, numbers, symbols, and punctuation. N-gram models are useful in many text analytics applications where sequences of words are relevant, such as in sentiment analysis, text classification, and text generation.

NLP vs Sentiment Analysis: https://www.ibm.com/topics/sentiment-analysis