**Model Summary : Abstract**

**Sentiment Analysis using NLTK & Transformers**

**1. Introduction**

The dataset used in this project is sourced from Kaggle, specifically the Amazon Fine Food Reviews dataset. This project employs various approaches for sentiment analysis:

* 1. **Traditional Approach using Python NLTK Toolkit-**
* Initially, we used the traditional approach with the Python NLTK toolkit to fit the model for sentiment analysis.

**1.2 VADER Model: Valence Aware Dictionary and Sentiment Reasoner-**

* We then used the VADER model, which utilizes a "bag of words" approach to perform sentiment analysis.

**1.3 ROBERTA Model with Hugging Face-**

* Next, we employed the RoBERTa model provided by Hugging Face, leveraging transformers and the method of transfer learning. Although initially trained on Twitter reviews, the RoBERTa model can be fine-tuned on the Amazon Fine Food Reviews dataset.

**1.4 Comparison between VADER and RoBERTa Model-**

* A comparison was made between the VADER model and the RoBERTa model to evaluate their effectiveness.

**1.5 Pretrained Pipelines for Sentiment Analysis Using Hugging Face**

* We explored pretrained pipelines for sentiment analysis using the Hugging Face library.
* **Libraries Used:** numpy, pandas, seaborn, matplotlib, nltk, tensorflow, keras, transformers

**2. Model Preprocessing**

**2.1)** **Data Cleaning**

**2.1.1 Quick EDA**

* Performed exploratory data analysis (EDA) using various visualizations such as bar plots, dist plots, and count plots.

**2.1.2 Basic NLTK Tokenization**

Applied NLTK's word\_tokenize method for tokenization, followed by part-of-speech (POS) tagging to identify the grammatical tags of words such as determiners, nouns, pronouns, and verbs.

**3. Model Building**

**3.1 VADER Sentiment Scoring:**

* Used NLTK's SentimentIntensityAnalyzer to obtain negative, neutral, and positive scores of the text. The VADER model follows a "bag of words" approach where:
* Stop words are removed as they do not contribute to sentiment.
* Each word is scored individually, and the total score is computed.
* Created an object of SentimentIntensityAnalyzer, fitted data, and conducted sentiment analysis using polarity scores. We considered 5-star ratings as positive reviews and vice versa.

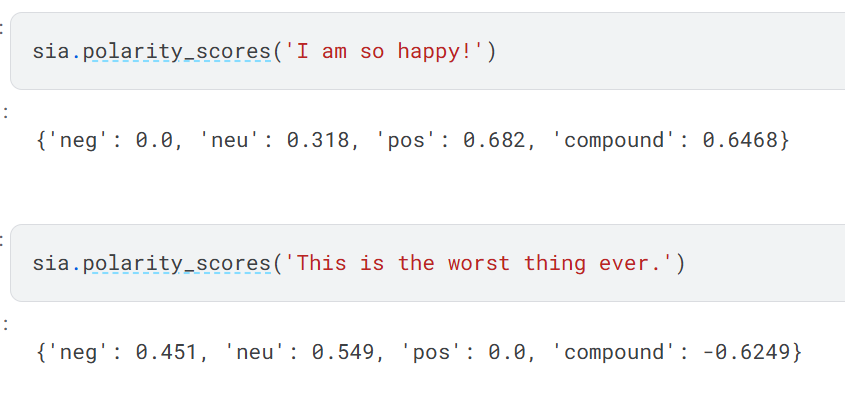


**3.2 Drawbacks of VADER**

* The VADER model scores each word individually but cannot comprehend the context and relationships among words, making it less effective for sentences with sarcasm, slang, idioms, etc.
* This limitation is addressed by the RoBERTa model, which Uses a model trained on a large corpus of data.Accounts for the context related to other words.
* We performed transfer learning with the RoBERTa model from Hugging Face on the Amazon Fine Food Reviews dataset.

**3.3 Combining Scores from VADER and RoBERTa**

* Combined the negative, neutral, and positive scores from both the VADER and RoBERTa models for each sentence in a table and concatenated these columns into the original dataframe.



**3.4 Using Hugging Face Pipelines**

* A quick and easy way to run sentiment predictions is using Hugging Face pipelines:
* from transformers import pipeline

sent\_pipeline = pipeline("sentiment-analysis")

* example = sent\_pipeline("Make sure to be happy and enjoy life")

# Output: [{'label': 'POSITIVE', 'score': 0.9923}]



**4. Conclusion**

* The implementation of sentiment analysis on the Amazon Fine Food Reviews dataset showcases the practical application of both traditional and advanced machine learning models. The traditional NLTK approach and VADER model provide a foundational understanding of sentiment analysis. However, due to the limitations of these models in understanding context and complex language constructs, advanced transformer models like RoBERTa are more effective. The transfer learning capabilities of RoBERTa, combined with the ease of use provided by Hugging Face pipelines, demonstrate significant improvements in sentiment analysis accuracy and contextual understanding. This project highlights the importance of choosing the right model and techniques based on the complexity and nature of the dataset.