# E-COMMERCE PRODUCT REVIEW SENTIMENT ANALYSER

\* "C:\Users\ADACHUKWU\OneDrive\Desktop\SentimentAnalysisImage.jpg"\*

## Table of Contents

- [Business Problem Overview] (#business-problem-overview)

- [Project Objective] (#project-objective)

- [Dataset Story] (#dataset-story)

- [Exploratory Data Analysis] (#exploratory-data-analysis)

- [Data Preprocessing] (#data-preprocessing)

- [Sentiment Analysis] (#sentiment-analysis)

- [Evaluation Metrics and Model Comparison] (#evaluation-metrics-and-model-comparison)

- [Model Deployment using Flask] (#model-deployment-using-flask)

- [Challenges and Limitations] (#challenges-and-limitations)

- [Future Work] (#future-work)

- [Conclusion and Recommendation] (#conclusion-and-recommendation)

## Business Problem Overview

Understanding customer feedback is vital for making informed business decisions about products and services, marketing strategies, and customer satisfaction. In the e-commerce sector, the large volume of customers reviews makes manual analysis of each entry impractical. To effectively and efficiently process this feedback from the customers, an automated sentiment analysis solution is needed to process and interpret customers reviews, enabling valuable insights that can aid businesses enhance their products offerings and services, ultimately leading to better customer experiences and satisfaction.

## Project Objective

The aim of this project is to perform sentiment analysis on the reviews to classify customer sentiments, which will assist the company in enhancing products features and customer experiences.

To achieve this objective, a detailed analysis of Ali Express reviews will be conducted to identify trends and preferences using pretrained models such as vADER and Supervised Machine Learning models to tag data extracted from the website.

## Dataset Story

### Origin

The dataset consists of Ali Express customer reviews for electronics from over a period of time and from different countries, encompassing a diverse range of customer insights and emotions over a period of time.

### Features:

* **Username:** Unique identifier chosen by each customer signed up to the company.
* **Location:** Geographical location of each customer that patronise the company.
* **Total Reviews:** Total number of reviews by each customer.
* **Date of Experience:** The date customer made the review on the website.
* **Content:** The full text of the customer’s review, providing detailed feedback.
* **Rating:** The rating out of 5 stars that the reviewer has given to each product.

## Exploratory Data Analysis

Prior to performing exploratory data analysis, the dataset was examined for shape, missing information, data types and duplicates. Missing information and duplicates were dropped from the data set, and the date was converted to Pandas date time to ensure uniformity in the dataset.

Subsequently, the dataset was examined for trends and patterns, and distribution of key variables. Furthermore, the frequency of the most used words was determined, and distribution and central tendencies of the dataset were ascertained using descriptive statistics.

## Data Preprocessing

To enhance the efficiency and effectiveness of the models, the text data are processed to ensure normalisation for future analysis. The steps performed to ensure this in the project are.

* **Lower case conversion:** Texts in the dataset are converted to lowercase to ensure uniformity, removing inconsistencies that could arise from differences in case sensitivity.
* **Removal of punctuation**: Punctuations are eliminated from the “Content” column to enhance efficiency of the models by making text data suitable for analysis.
* **Removal of Numerical Expressions:** Elimination of numbers from the text enables the models to focus more on the language processing and text analysis for more impactful data analysis.
* **Removal of stop-words:** Removing frequently used words which are often meaningless (“the”, “is”, “in”) from the text data will help focus our analysis on words with more and impactful meaning.
* **Removal of infrequent words:** The words that appear once in the entire dataset after the frequency of each word are calculated, are removed as they may be low semantic weighted. This allows the models to focus on more significant words.
* **Tokenization:** Using libraries in NLTK, the text in each “content” column was tokenised into words, causing them to be split into smaller units. Tokenization enables each word in content to be examined individually.
* **Lemmatization:** This step reduces the tokenized word into their root forms (basic forms) such as “running” to “run”.

## Sentiment Analysis.

To determine which model best suits our analysis, the project compared the efficiency of a pretrained model vADER ((Valence Aware Dictionary and Sentiment Reasoner) and supervised machine learning models (Naïve-Bayes, Logistic Regression and Support Vector Classifier models).

### vADER (Valence Aware Dictionary and Sentiment Reasoner)

Sentiment analysis was performed on the “Content” column of the dataset using the pretrained vADER model for analysis. The model assigns polarity scores to words based on the emotional tone of the texts. The aggregates of these polarity scores provided a compound sentiment score for each text, indicating positive, negative or neutral sentiments.

### Machine Learning Models

To train machine learning models to efficiently classify customer sentiments, we investigated **Naïve-Bayes**, **Support Vector Classifier**, and **Logistic Regression** models to determine the most successful sentiment classification model.

* **Train and Test Data set:**  Training and test data was split ratio 20:80.
* **Text Vectorization:** To ensure machine learning models are able to interpret the texts data, we convert the text to numerical values using the TF-IDF vectorizer.
* **Class Imbalances:** This was handled using the SMOOTHENING to handle the class imbalance in the dataset.

**####Model Selection.**

After the data set has been split into train and test data, vectorised and smoothened, the train dataset is fit into each model for exposure and test data set are used to check the evaluate the ability of the model to predict accurately.

## Evaluation Metrics

To assess the performance of the machine learning model, we utilised metrics such as accuracy, precision, recall, and F1-score.

Given that our dataset is imbalanced, F1-Score is used to determine the best performing model as it balances precision and recall scores.

### Model Comparison

Support Vector Classifier performed best of the three as it has highest F1-score and highest weighted average F1-score (90%). F1-Score and weighted average are used as the dataset exhibited class imbalance, with more positive reviews than negative or neutral reviews.

## Model Deployment using Flask

The best performing model i.e. Support Vector Classifier (SVC) is deployed using an open-source Python library, Streamlit to create web application that will allow users interact with our models for real-time prediction.

## Challenges and Limitations

* **Scalability of the System:** Limitations faced include the cost of handling large dataset, computational intensity required to test other advanced models, low processing time and balancing model complexity with performance.
* **Handling Multilingual Reviews:** Advanced models may be required to handle the complexity of processing multiple languages in the same dataset, potential translation errors, and the ability to detect multiple languages and emotional tone of these languages. This is beyond the scope of the current project.

## Future Work

* **Improving Sentiment Analysis Engine:** Improving the model for better accuracy, scalability, and adaptability to diverse dataset will be considered in future projects. Also, we will consider optimising feature extraction methods, integrating language processing techniques and continual improvement and updating of model to handle advancing linguistic expressions and user emotions.
* **Improving User interface:** Future project will involve intuitive design, responsive layout, accessibility features and iterative user feedback integration for continuous improvement.

## Conclusion and Recommendations

The project shows the organisations can improve on customer experience and gain competitive advantage in an already competitive market by leveraging the power of data science and visualisation to derive insights from customers feedback via effective analysis.

We will recommend the integration of advanced sentiment analysis tools such as deep learning and ensemble techniques to enhance the predictive capabilities of the application. Furthermore, integrating other e-commerce platforms and real-time feedback would improve the importance of the application in e-commerce sector.