# Conception Phase

The sentiment analysis application is designed to analyse customer reviews and classify them based on their emotional state/valence. The application will be able to handle effects such as negative, neutral and positive. The application consists of four main components: data collection, data pre-processing, sentiment analysis and validation. The components interact with each other to produce accurate and reliable results.

### 1. Data Collection:

The first component of the application is data collection. In this phase, customer reviews will be collected from a dataset found on Kaggle. It is about the reviews of digital video games, having following attributes: marketplace, customer_id, review_id, product_id, product_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, review_body and review_date. This dataset will be used for training the NLP models.

### 2. Data Pre-processing:

The second component of the application is data pre-processing. In this phase, the raw text data will be cleaned and pre-processed to make it ready for the sentiment analysis process. The pre-processing steps include removing stop words, removing punctuation marks, converting text to lowercase, lemmatizing and stemming.

### 3. Sentiment Analysis:

The third component of the application is sentiment analysis. In this phase, the pre-processed text data will be analysed to classify it into three different emotional states/valence, i.e., positive, neutral, or negative. The sentiment analysis process will use machine learning algorithm: Random Forest to classify the text data.

### 4. Validation:

The fourth component of the application is validation. In this phase, the accuracy of the sentiment analysis results will be validated. The validation process will compare the sentiment analysis results with the ground truth labels of the dataset to determine the accuracy of the sentiment analysis process. When the first validation is achieved, a hyperparameter optimization will be processed for further tuning and expect a higher accuracy.

## 5. UML Diagram:

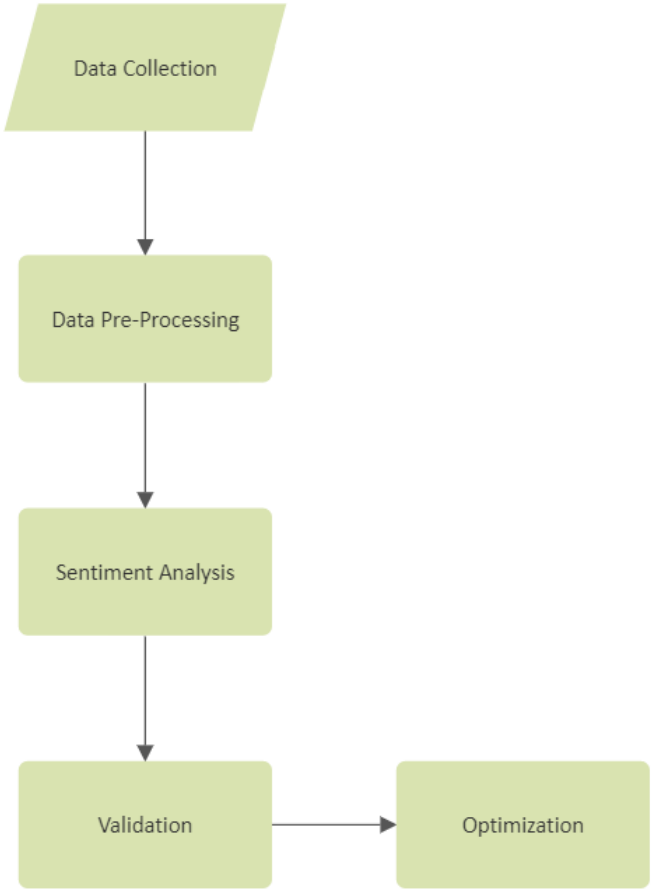


*Figure 1 – UML Diagram*

Overall, UML diagram provides a high-level overview of the different modules involved in sentiment analysis and how they work together to produce the final output.

## 6. Planned tools and modules:

- pandas for data analysis
- matplotlib for creating graphs
- nltk for analyzing text data and perform sub tasks
- wordcloud for visualizing text data to see patterns
- scikit-learn for machine learning processes
- Google Colab to create and run the project file