# Introduction

The fashion industry is one of the most dynamic and rapidly evolving industries globally, with women's fashion being a significant and influential part of it. With the rise of e-commerce, women's fashion has become more accessible than ever before, with online shopping platforms providing a convenient way for customers to purchase clothing items from the comfort of their own homes. However, with the vast number of products and brands available online, customers often rely on reviews to make informed purchase decisions. As such, sentiment analysis has become a critical tool for businesses in the e-commerce space to understand customer feedback better. Sentiment analysis is the process of determining the emotional tone behind written or spoken language. In this context, it involves analyzing customer reviews of women's clothing items on e-commerce platforms to determine the underlying sentiments expressed[1].

The goal of this report is to provide a comprehensive analysis of women's clothing e-commerce reviews using sentiment analysis. The report will detail the methodology used to collect and analyze the data, including the use of natural language processing and machine learning techniques. The findings of the analysis will be presented in a clear and concise manner, highlighting trends and patterns in customer sentiments towards different clothing items and brands. Furthermore, the report will explore the potential applications of sentiment analysis in the fashion industry, including how e-commerce platforms and fashion brands can use these insights to improve their marketing strategies and customer relationships. Additionally, we will discuss the ethical considerations involved in the use of sentiment analysis and how businesses can ensure that they are using this technology in a responsible and transparent manner.

In conclusion, this report is an essential resource for anyone interested in understanding customer sentiments towards women's clothing on e-commerce platforms. It offers valuable insights that can inform decision-making and help businesses provide better products and services to their customers. Through the analysis of customer feedback using sentiment analysis, we can gain a better understanding of customer needs and preferences, ultimately leading to more satisfied customers and increased sales for e-commerce platforms and fashion brands.

# Dataset

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

- **Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.
- **Age:** Positive Integer variable of the reviewers age.
- **Title:** String variable for the title of the review.
- **Review Text:** String variable for the review body.
- **Rating:** Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.

- **Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count:** Positive Integer documenting the number of other customers who found this review positive.
- **Division Name:** Categorical name of the product high level division.
- **Department Name:** Categorical name of the product department name.
- **Class Name:** Categorical name of the product class name.

## Contribution

This report has taken the reviews of women's clothing and processed them using text processing techniques, including NLP preprocessing, to obtain preprocessed data. Then, sentiment analysis has been performed on this data using the Text Blob library of NLP. In the end, we applied machine learning classifiers i.e., Navies bayes, KNN, Decision tree, Logistic regression, Random Forest and  to the data, including 5 classifiers for machine learning and two deep learning model are also evaluated. The novelty of this project is that we first counted the sentiments of the reviews, then trained the ML models and finally two deep learning model were also evaluated. Lastly, we compared all of these techniques, which had not been done before.

## Preprocessed Data

Text preprocessing is the process of cleaning and transforming unstructured text data to make it more suitable for analysis. It involves various techniques to clean, normalize, and transform the text data, including removing unwanted characters, lowercasing all text, and removing stop words, among others[2].

Here are some common steps used in text preprocessing for a sentiment analysis project:

- **Tokenization:** Tokenization is the process of breaking down text into individual words or phrases. This step is important because it allows the model to understand the structure of the text and identify the key elements that contribute to its meaning. For example, tokenizing the sentence "I love pizza" would break it down into three tokens: "I," "love," and "pizza."
- **Lowercasing:** Lowercasing involves converting all text to lowercase. This step is important because it ensures that the model does not treat different cases of the same word as separate entities. For example, "Dog" and "dog" would be considered two different words if they are not lowercased.
- **Stop word removal:** Stop words are common words that do not add much meaning to the text, such as "the," "a," and "and." Removing these words can help to reduce the size of the

vocabulary and improve the accuracy of sentiment analysis. Stop word removal can be performed using a pre-defined list of stop words, which can be customized for specific applications.

- **Stemming or lemmatization:** Stemming and lemmatization are techniques used to reduce words to their base form. Stemming involves removing suffixes from words, so that "running" and "runs" would both be converted to "run." Lemmatization, on the other hand, involves converting words to their base form based on their dictionary meaning. For example, "was" and "were" would both be converted to "be." Stemming and lemmatization can help to reduce the size of the vocabulary and improve the accuracy of sentiment analysis.
- **Removing punctuation and special characters:** Punctuation and special characters, such as commas, periods, and question marks, do not add much meaning to the text and can be removed. This step helps to simplify the text and make it easier to analyze.
- **Removing URLs or email addresses:** URLs, email addresses, and other personal identifying information may not be relevant to sentiment analysis and can be removed. This step helps to protect the privacy of individuals and ensure that the model focuses on the relevant aspects of the text.
- **Spell checking:** Correcting spelling mistakes and errors in the text can improve the accuracy of sentiment analysis. This step involves identifying misspelled words and replacing them with the correct spelling. There are various spell-checking tools available that can be used to automate this process.

Overall, these preprocessing steps help to clean and transform unstructured text data into a more structured format that can be analyzed more effectively using sentiment analysis techniques. The given data is passed through these stated preprocessing steps and the data is cleaned by applying these one by one. The distribution of age column shows that most of the customers are between 30 to 50 ages. There are 2000 instances of customer which have age between 35 to 40 can be seen in figure 1

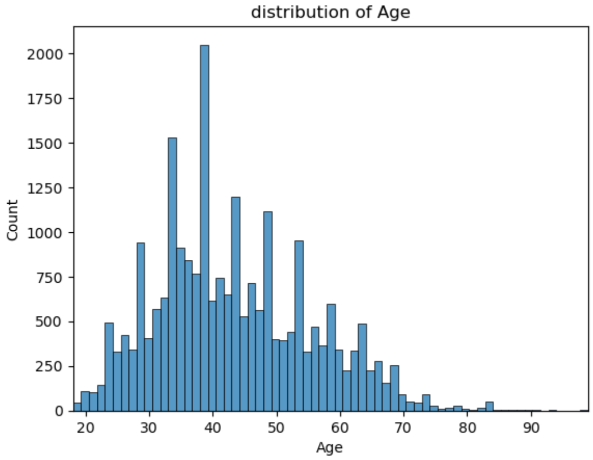

Figure 1 Distribution of Age

## Algorithm

As sentiment analysis prediction is a task of classifier to predict positive, negative and neutral sentiment and for this purpose we five machine learning models[3] and two deep learning models. For this purpose, different columns are label encode to convert data that will be used for training and also used one hot encoding for other categorical variables[4].

- **Naive Bayes:** Naive Bayes is a simple probabilistic classifier based on Bayes' theorem with the assumption of independence between the features. It is often used for text classification and spam filtering.
- **Logistic Regression:** Logistic regression is a linear classifier that models the probability of the input belonging to a certain class using a logistic function. It is commonly used in binary classification problems.
- **K-Nearest Neighbors (KNN):** KNN is a non-parametric algorithm that classifies new instances based on the similarity to the k nearest training instances. It is often used in pattern recognition and image classification.
- **Decision Tree:** A decision tree is a tree-like model where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label. It is commonly used in data mining and decision-making problems.
- **Random Forest:** Random Forest is an ensemble method that creates multiple decision trees and combines their predictions to improve the accuracy and avoid overfitting. It is often used in image classification, text mining, and other complex data analysis tasks.
- **Deep Learning Classifier:**

## Results

In this project, we have evaluated 5 machine learning algorithms including Naive Bayes, Logistic Regression, KNN, Random Forest, and Decision Tree. If we look at the accuracy, except for Naive Bayes, all other algorithms have an accuracy of more than 0.8, while Naive Bayes has an accuracy of 0.3 as shown in Figure 2.

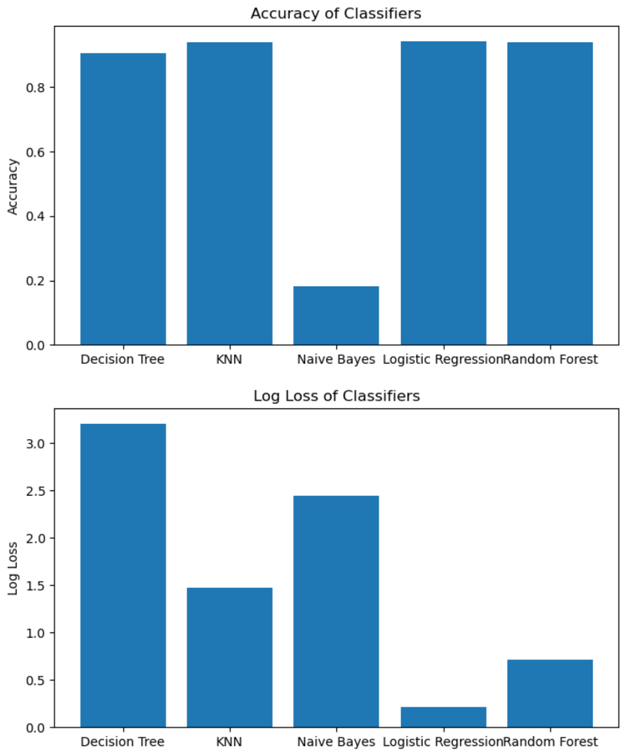

Figure 2 Accuracy and loss of Machine learning Classifier

Figure 2 also shows displays the log of loss recorded for different algorithms, indicating that Logistic Regression has the least loss among all the algorithms, whereas Decision Tree has the highest loss. This suggests that Logistic Regression is the best algorithm. However, deep learning classifier show more accurate result as shown in figure 3 that model 2 has accuracy reaching to 0.9645 while second model have accuracy 0.9625 where has both models have loss low than 0.13. Note one thing, we evaluated this result on 50 epochs and with batch size of 32 and 62 for model one and two respectively. Overall, results shows that deep learning classifier is the best classifier to work with this kind of data and other shows good results with given deep architectures.

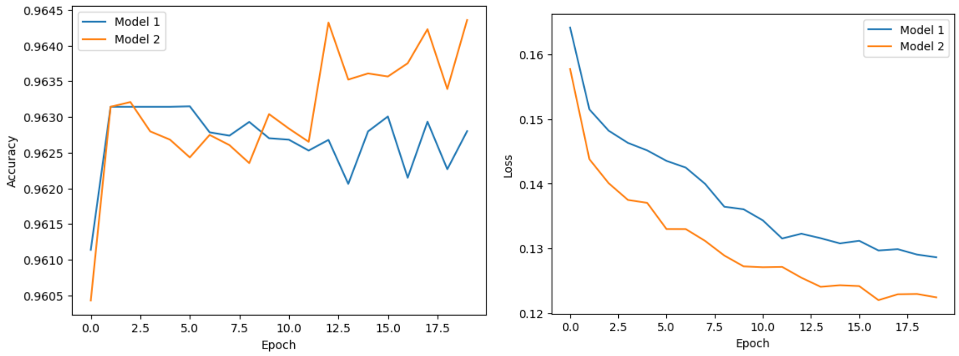

Figure 3 Shows the Accuracy and loss of Deep learning classifier

## References

[1] “Women’s E-Commerce Clothing Reviews.” https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews (accessed Mar. 26, 2023).

[2] “AI Machine Learning Bootcamp By Caltech CTME | Top AI Machine Learning Certification.” https://pg-p.ctme.caltech.edu/ai-machine-learning-bootcamp-online-certification-course?utm_source=google&utm_medium=cpc&utm_term=natural%20language%20processing&utm_content=19131908313-147723748961-651768498223&utm_device=c&utm_campaign=Caltechdomain+-+Search-US-25_54-en+-+NB+-+DataCluster-AIML-Bootcamp-CAIMLB-Caltech-adgroup-Natural+Language+Processing+-+Subcategory:Natural+Language+Processing-MT:Phrase&gclid=Cj0KCQjw2v-gBhC1ARIsAOQdKY2umkE7jNMxBsTKE8f7tCxWIFVpMaOyJpCdMDUOvk3bVTf4qlXy7L0aAp_NEALw_wcB (accessed Mar. 26, 2023).

[3] “Top 10 Machine Learning Algorithms You Need to Know in 2023 | Simplilearn.” https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article (accessed Mar. 26, 2023).

[4] “Categorical Encoding | One Hot Encoding vs Label Encoding.” https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/ (accessed Mar. 26, 2023).