Customer Feedback Analysis and Classification
Using NLP, Ensemble Techniques, and Model Deployment

Skills take away from this project

Python, Pandas

ML- Scikit Learn

DL-Tensorflow

Pre Trained Models/
Transformers using Hugging Face

Deployment using streamlit

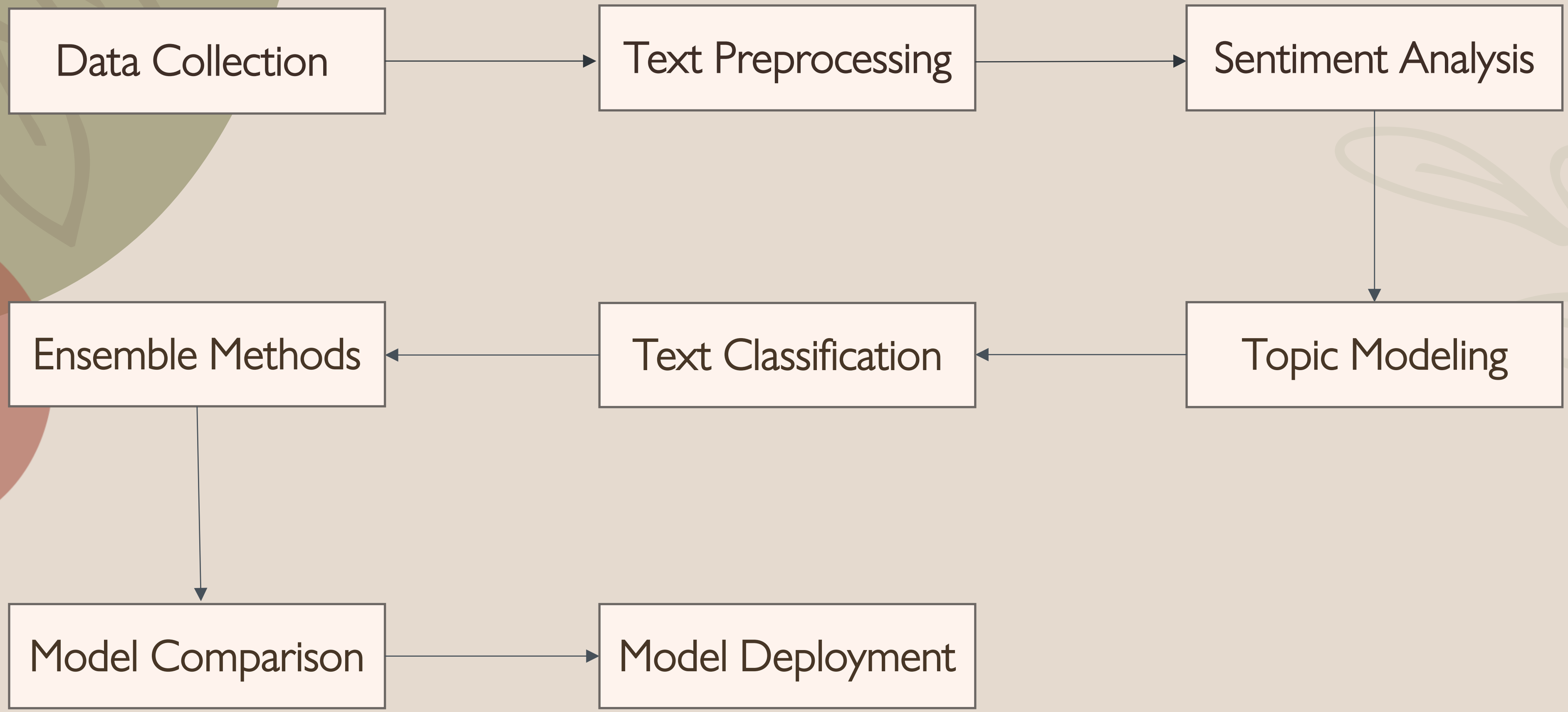
Problem Statement:

The project aims to analyze customer feedback to gain insights into customer sentiment and identify key topics and trends.

Objective:

The objective is to compare different models and ensemble techniques to select the best model for prediction and deploy the model using Streamlit.

Approach


Data Collection
Text Preprocessing
Sentiment Analysis
Topic Modeling
Text Classification
Ensemble Methods
Model Comparison
Model Deployment

Text preprocessing

It is the process of cleaning and transforming the raw data into a structured format for Natural Language Processing(NLP).

Key steps included in NLP:

- Lowercase the text
- Punctuation removal
- Tokenization
- Lemmatization

Sentiment analysis

The process of analyzing digital text to determine if the emotional tone of the text is positive, negative or neutral.

- Machine Learning(Logistic Regression and Support vector machine(SVM))
- Deep Learning(LSTM)
- Transformers(Hugging face model-Roberta)

Logistic Regression

Logistic Regression is a statistical method used for **binary classification problems**, where the outcome is categorical. Despite its name, it is actually a classification algorithm and not a regression algorithm. It predicts the **probability** of a binary outcome based on input features.

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm primarily used for **classification** and **regression** tasks. It is particularly effective in high-dimensional spaces and is commonly used for **binary classification.** It is a supervised machine learning algorithm that classifies and analyzes data.

Long-Short Term Memory

LSTM (Long Short-Term Memory) is a type of **Recurrent Neural Network (RNN)** designed to overcome the limitations of standard RNNs, particularly the problem of **vanishing gradients**. It excels at learning and modeling **long-term dependencies** in sequential data, making it widely used for tasks like time series prediction, speech recognition, and text generation.

LSTM

- LSTMs utilize a set of **gates** and **memory cells** to selectively retain or discard information across long sequences, enabling them to effectively capture long-term dependencies.
- By addressing the **vanishing gradient problem**, LSTMs can maintain context over extended time steps, which is critical for sequential data tasks.
- Logistic Regression, being a binary classification model, predicts probabilities but was found to deliver unsatisfactory accuracy for the given task.
- SVM, while effective for binary classification, was observed to be computationally more expensive and time-consuming for prediction.
- In comparison, LSTM demonstrated **higher accuracy** and required **less computational time**, making it the most suitable choice for the task.

Accuracy of the Models

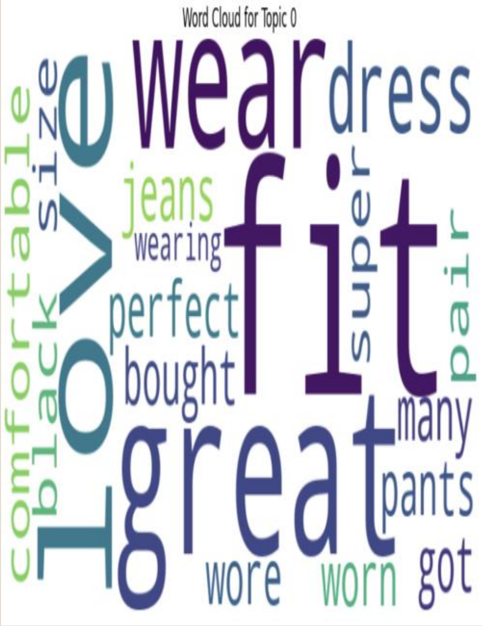
MODELS	PRECISION		RECALL		F1 SCORE	
	0	1	0	1	0	1
Logistic Regression	0.90	0.93	0.94	0.89	0.92	0.91
Support Vector Machine(SVM)	0.88	0.94	0.94	0.87	0.91	0.90
Long short-term memory(LSTM)	0.91	0.97	0.97	0.91	0.94	0.94

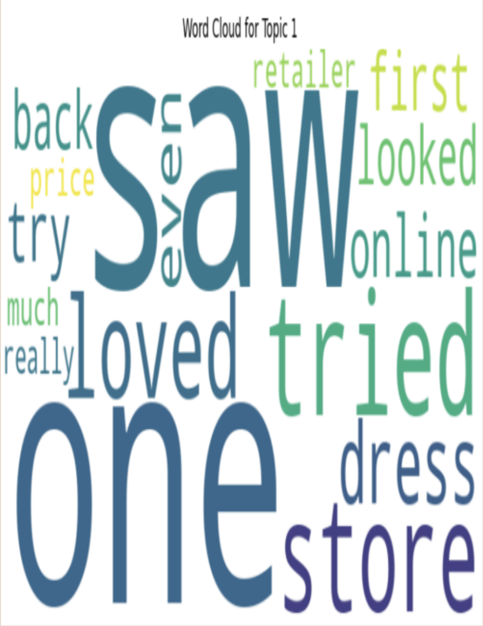
Transformer(Roberta) accuracy - 0.88

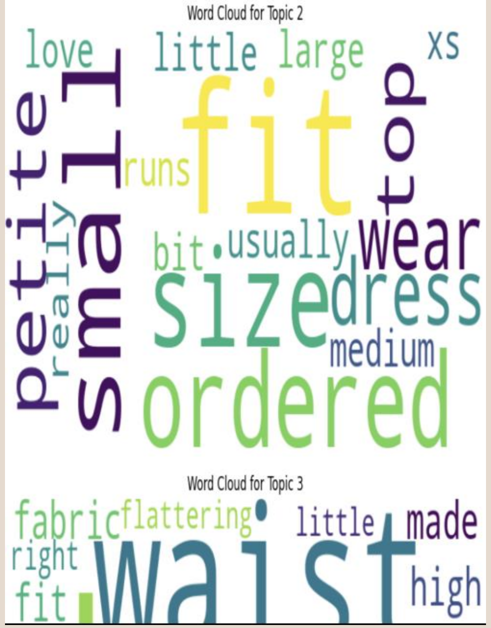
0 - Not Recommended
1 - Recommended

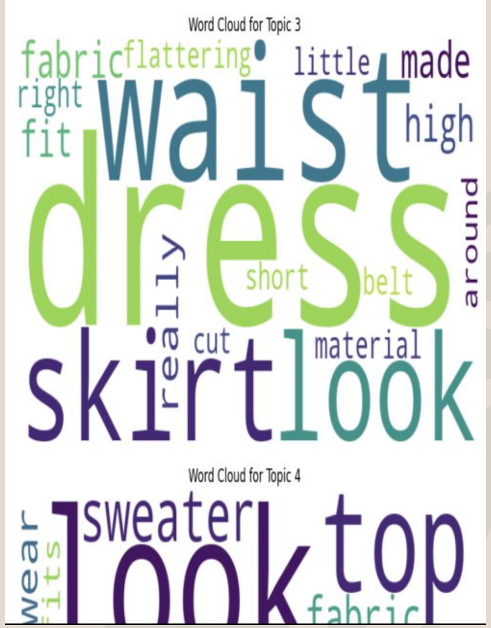
Topic modelling

Topic modelling is the unsupervised machine learning technique used in NLP to identify the main topic or themes present in a collection of text document it helps in discovering hidden structures in text data by grouping similar words into clusters each representing a topic.









Ensemble methods

Ensemble methods are widely used in machine learning and have become a cornerstone of modern predictive modeling. They improve accuracy, reduce overfitting, and increase robustness. There are several types of ensemble methods, and they can be broadly classified into **bagging**, **boosting**, and **stacking**.

Bagging:

Bagging is a technique where multiple instances of the same model are trained on different subsets of the data. These subsets are obtained by **random sampling with replacement**, which means some data points may appear multiple times in a subset, and others may not appear at all.

Summary of bagging accuracy:

Models	Accuracy
Logistic Regression	0.56
SVM	0.66
LSTM	0.94

Business use case

Common issues

- Fitting was not correct due to inappropriate size chart.
- Quality mismatch due to product was not same as in image.

Suggestion

- Providing detailed description and correct size chart
- Attachments of reviews by previous customers including images, videos.

Marketing strategy

- Assigning skilled tailors to specific areas for minor fitting adjustments.
- Providing offers for purchasing according to season times.
- Doing more collaboration with celebrities and influencers for the reach of products.
- Tie-up with retail sellers.
- Allow customers to order online and pick up from the nearest retail seller, reducing delivery time and cost

Thank you