Project Title: Sentiment analysis for marketing

Team Leader: GoppyKrishna M

Team Members: Gokul S, Hariharan T, Hari Prasath R B

Introduction

The project aims to perform sentiment analysis on customer feedback to gain insights into competitor products. By understanding customer sentiments, companies can identify strengths and weaknesses in competing products, thereby improving their own offerings. The project employs Natural Language Processing (NLP) techniques and data visualization to achieve these goals.

Problem Statement

The project's main problem is to analyse customer feedback and extract valuable insights. Specific challenges include identifying customer sentiments, understanding product feedback, assessing brand perception, conducting competitive analysis, and tracking market trends.

Design Thinking

1. **Data Collection:** Identify a dataset containing customer reviews and sentiments about competitor products.

2. **Data Preprocessing:** Clean and preprocess the textual data for analysis.

3. **Sentiment Analysis Techniques:** Employ different NLP techniques like Bag of Words, Word Embeddings, or Transformer models for sentiment analysis.

4. **Feature Extraction:** Extract features and sentiments from the text data.

5. **Visualization:** Create visualizations to depict the sentiment distribution and analyse trends.

6. **Insights Generation:** Extract meaningful insights from the sentiment analysis results to guide business decisions.

Project Outcomes

The Project helps companies Visualize and Understand:

- **Customer satisfaction**
- **Product feedback**
- **Brand perception**
- **Competitive Intelligence**
- **Market Trend Analysis**

This information helps the Companies to improve their products and services by observing the performance of Competitor's products among the Public. Thus, improving the Brand’s perception among the public leads to Profit and Growth of the Company.

Key Features

- **Context Based Stop Words Identification** - Performed by taking Top-25 frequently used words in all the Sentiment Categories.
- **Negation Handling** - Identifying and Reversing negation statements in the text.
- **Robust Model** - Used a RoBERTa Deep-Learning Model for higher prediction accuracy.
- **Interactive Dashboard** - Ability to target specific a Company in the Dashboard.

Fine tuning Pre-trained Models

- Fine-tuning pre-trained sentiment analysis models is a way to improve the performance of a sentiment analysis model on a specific task or domain.
- It involves training the model on a new dataset that is labelled with the desired sentiment labels.
- This allows the model to learn the specific patterns and nuances of the new dataset, which can lead to significant improvements in accuracy.
- There are a number of advanced techniques that can be used to fine-tune pre-trained sentiment analysis models, but they can be complex and time-consuming.
- However, fine-tuning can be a powerful way to improve the performance of your model on a specific task or domain.

We've trained 2 models by fine-tuning a BERT model and a Roberta model.

1. **BERT Model**

BERT, or Bidirectional Encoder Representations from Transformers, is a machine learning model for natural language processing (NLP). It was developed in 2018 by researchers at Google AI Language and has since become the state-of-the-art model for many NLP tasks, including question answering, sentiment analysis, and named entity recognition.

BERT is a bidirectional model, meaning that it can learn the context of a word from both the words that come before it and the words that come after it. This is in contrast to previous NLP models, which were typically unidirectional, meaning that they could only learn the context of a word from the words that come before it.

BERT is based on the Transformer architecture, which is a type of neural network that is particularly well-suited for NLP tasks. The Transformer architecture uses a self-attention mechanism, which allows the model to learn long-range dependencies in text.

BERT is trained on a massive dataset of text and code, which allows it to learn a deep understanding of language. Once trained, BERT can be fine-tuned for a variety of NLP tasks by adding a task-specific output layer.

2. **RoBERTa model**

RoBERTa (Robustly Optimized BERT Pre-training Approach) is a large language model (LLM) trained to perform many kinds of natural language processing (NLP) tasks, such as question answering, summarization, and translation.

RoBERTa was invented by a team of researchers at Facebook AI, led by Yinhan Liu.

It is a transformer-based model, which means that it uses a self-attention mechanism to learn long-range dependencies in text.

RoBERTa is based on the BERT model, but it makes a number of improvements, including:

- Training on a much larger dataset of text
- Using more effective training procedures
- Removing the next-sentence prediction objective

These improvements have resulted in RoBERTa achieving state-of-the-art results on a variety of NLP benchmarks.

Pre-Processing the Data

The data preprocessing phase is a critical step in preparing the textual data for sentiment analysis. The primary goal is to clean and transform the raw text data into a format that can be used effectively by machine learning models. This phase ensures that the data is in a consistent and usable state for subsequent analysis.

Steps in Data Preprocessing:

1. Handling Null Values:
    - Identify and handle null values in the dataset to prevent issues during analysis.

- Null values can be identified and either removed or imputed, depending on the context.

2. Eliminating Unnecessary Columns:
    - Identify and remove columns that do not contribute to the sentiment analysis task.
    - Reducing the dataset to essential columns can improve efficiency.
3. Text Cleaning:
    - Remove HTML entities: Convert HTML entities (e.g., <) to their respective symbols (e.g., <).
    - Decode Emoticons: Translate emoticons used in tweets to corresponding textual descriptions (e.g., 🙄 to "slightly smiling face").
    - Eliminate Mentions and Hashtags: Remove '@ mentions' and '# hashtags' related to airlines and external links, as they may not be relevant for sentiment analysis.
    - Removing Punctuation and Digits: Eliminate punctuation marks and numeric digits from the text data, as they might not carry sentiment information.
    - Tokenization and Lemmatization: Tokenize the text into words and then lemmatize them based on their Parts of Speech (POS) tags. For example, "loving" is lemmatized to "love," and "ate" to "eat."
    - Removal of Context-Based Stopwords: Remove stopwords that are context-based, such as "flight" or "fly," to ensure that they do not influence sentiment analysis.
4. Visualizing Processed Tweets:
    - Create word clouds to visualize the most frequent words in the processed tweets. In a word cloud, larger words indicate higher frequency in the dataset.

5. Saving Processed Data:

    - Save the cleaned and pre-processed data to a new file for further analysis and model training.

Model Training

1. **Models Used**

In this phase, several machine learning and deep learning models are employed to predict sentiment labels for customer feedback. The models used include:

- **Machine Learning Models**

    1. **Logistic Regression:** A widely used linear classification algorithm.
    2. **SVM (Support Vector Machine):** A powerful classification algorithm known for its ability to handle high-dimensional data.
    3. **Naive Bayes Model:** A probabilistic model based on Bayes' theorem, suitable for text classification.
    4. **Decision Tree Classifier:** A decision tree-based model for classification tasks.

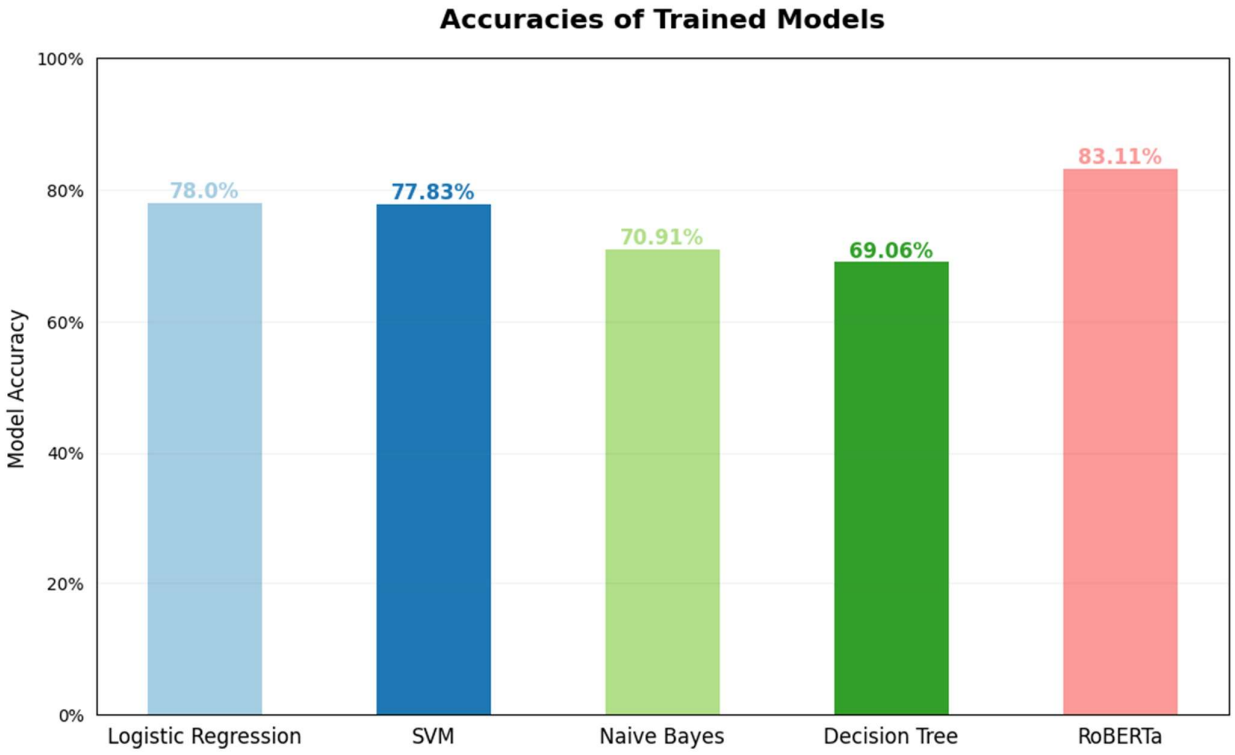
- **Deep Learning Models**

    RoBERTa Model: A pre-trained transformer-based model designed for natural language understanding tasks. It is used for improved sentiment analysis accuracy.

2. **Model Evaluation**

Once the models are trained, they are evaluated to assess their performance. Evaluation metrics are used to measure the model's accuracy, precision, recall, F1-score, and other relevant metrics. The evaluation process allows us to determine how well each model predicts sentiment labels for customer feedback.

3. **Model Comparison**



After evaluating all the models, a comparison is made to select the most accurate and effective model for the sentiment analysis task. The RoBERTa model is the one that performs the best with 83% accuracy and is chosen as the final model to be used in the project.

The accuracy and performance of the selected model contribute to the quality of the insights and recommendations derived from customer feedback analysis, aiding businesses in making data-driven decisions.

Deployment

1. **Import Necessary Libraries:** This step involves importing the required libraries and dependencies to facilitate the deployment process.
2. **Load the Trained Model:** The trained RoBERTa model is loaded into memory, allowing it to be used for sentiment analysis.
3. **Pre-process the Testing Data:** In this phase, the testing data is pre-processed. The dataset is shuffled, and a subset of 200 records is chosen due to computational constraints. This number can be increased if there is sufficient computational power.
4. **Predicting the Sentiment:** The cleaned text data is tokenized and used to predict sentiment labels (positive, negative, or neutral) using the trained RoBERTa model. A few of the sentiment predictions are displayed as a sample.
5. **Saving the Input Data for the Dashboard:** The pre-processed data, including cleaned text and sentiment predictions, is saved and prepared for visualization in the dashboard.
6. **Creating the Dashboard:** The includes several key components:
    - Radio button group for selecting the company of interest.
    - Horizontal stacked bar chart representing the count of positive, negative, and neutral tweets about the selected company.
    - Pie chart illustrating the reasons for negative tweets about the chosen company.
    - Two-word clouds displaying the most frequent words in positive and negative tweets, offering insights for market trend analysis.

This deployment phase is essential for providing users with an intuitive interface to access sentiment analysis insights. Users can input data, receive sentiment predictions, and gain a deeper understanding of customer feedback through visual representations.

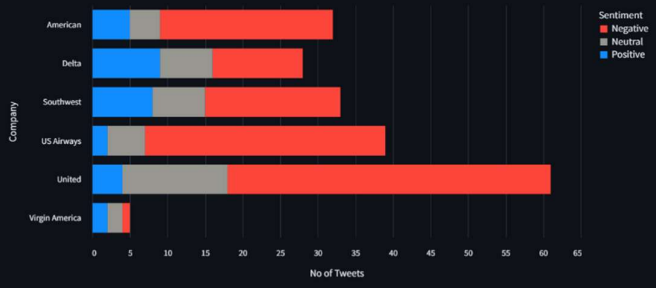
Deploy

Twitter Sentiment Analysis Dashboard

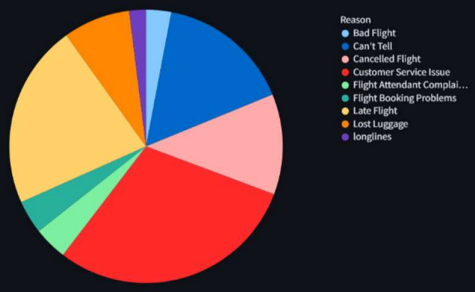
Select Company

- (•) All
- ( ) American
- ( ) Delta
- ( ) Southwest
- ( ) US Airways
- ( ) United
- ( ) Virgin America

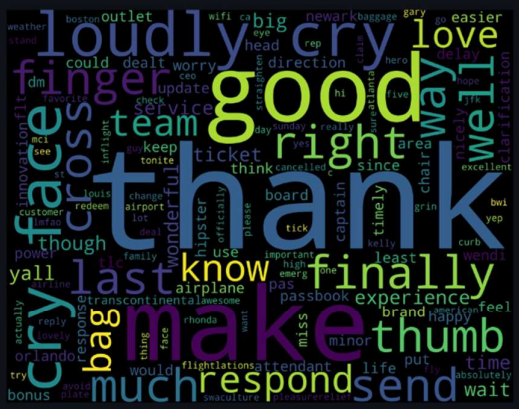
All Negative sentiment reasons



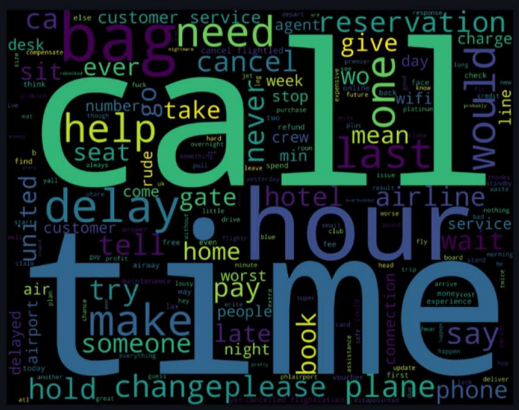
All Negative sentiment reasons



Word Cloud of Positive words by All customers



Word Cloud of Negative words by All customers



Conclusion

Achievements

The project empowers companies to improve their products and services by understanding customer feedback about competitor products. It provides insights into customer satisfaction, product feedback, brand perception, competitive intelligence, and market trend analysis.

Project Impact

The project's impact is reflected in the ability of businesses to make informed decisions based on customer sentiments. This leads to improved products, better brand perception, and increased profitability and growth.

Future Enhancements

Possible future enhancements include extending the dataset sources, improving deep learning model performance, and expanding the dashboard's functionality to provide more detailed insights.