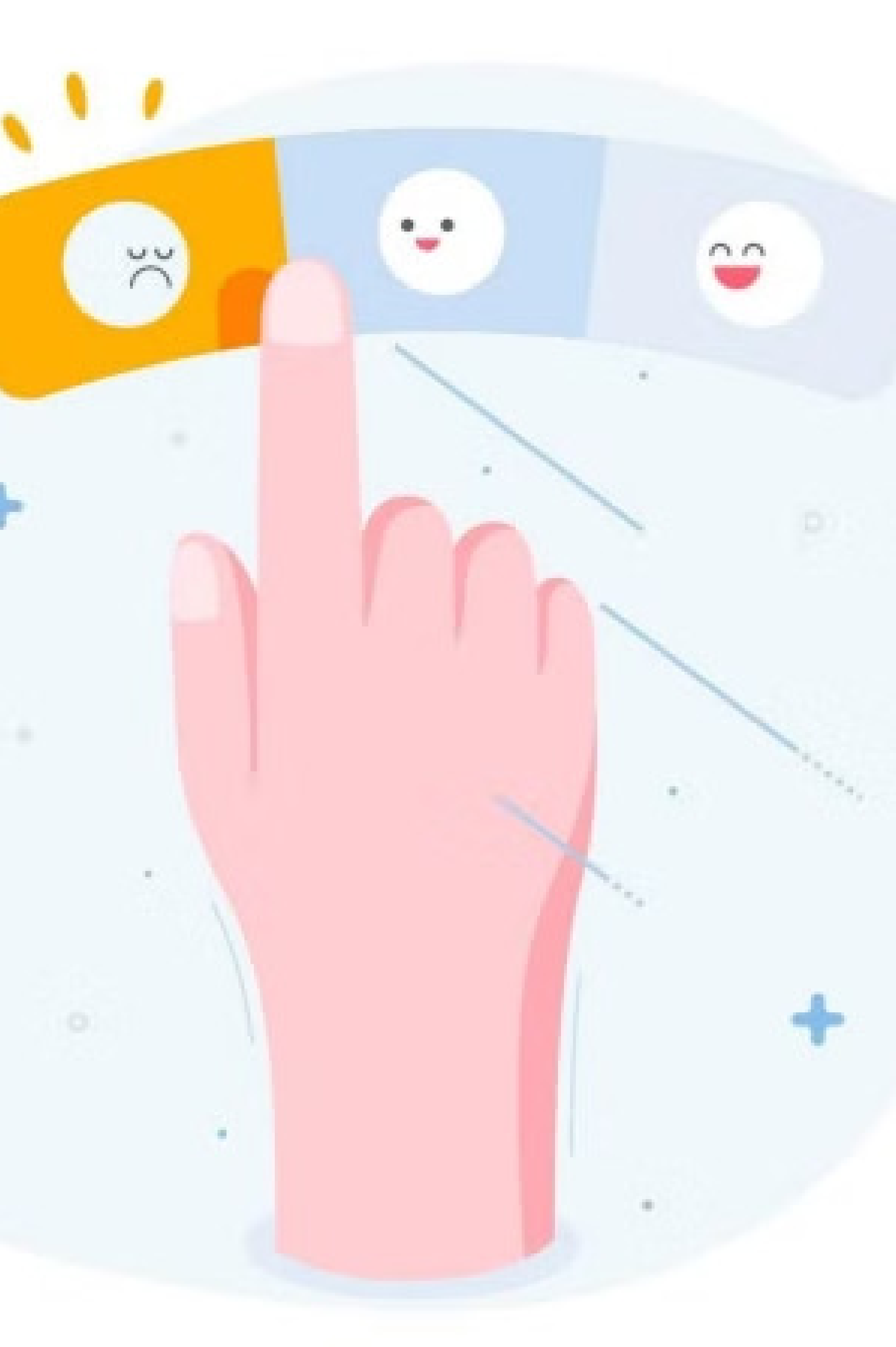


Real-Time Sentiment Analysis Pipeline

This DEPI graduation project explores sentiment analysis using various machine learning models, focusing on Natural Language Processing (NLP). The project compares traditional algorithms like Naive Bayes and Logistic Regression with modern embedding models like BERT, GloVe, Word2Vec, and Doc2Vec. The goal is to achieve high-performance sentiment classification for Amazon product reviews, classifying them as positive or negative.

by Sohaib Osama

Project Overview and Problem Statement

Project Overview

This project assesses different sentiment analysis approaches. It involves downloading, cleaning, tokenizing, and balancing an Amazon review dataset. Traditional machine learning models and modern embedding techniques (TF-IDF, GloVe, BERT) are used for feature extraction, model training, and performance evaluation. A simple MLP model is used for training embeddings.

Problem Statement

Sentiment analysis aims to determine the emotional tone of text. This project focuses on classifying reviews as positive or negative, comparing traditional and modern NLP models (BERT, GloVe vs. Naive Bayes, Logistic Regression). The goal is to find the best-performing approach and provide interpretable results.

Objectives and Outcomes

1 **Data Preprocessing**

Clean and structure the dataset using NLP techniques (lemmatization, tokenization, stopword removal). Visualize data with word clouds and class distribution charts.

2 **Model Building**

Build and compare models (Naive Bayes, BernoulliNB, Logistic Regression) using metrics like accuracy, precision, recall, and F1-score.

3 **Deep Learning**

Integrate deep learning models like BERT and MLPs for embeddings (GloVe, Word2Vec, Doc2Vec).

4 **Deployment and Explainability**

Deploy the best model in a Django web app and use SHAP for interpreting predictions.

Dataset and DistilBERT Model

Amazon Product Reviews Dataset

The project uses the Amazon Product Reviews dataset from Kaggle, focusing on the 'reviewText' column. The dataset is labeled for sentiment analysis (1 for positive, 0 for negative). It's imbalanced, with more positive (76%) than negative (24%) reviews.

DistilBERT Base Model

DistilBERT, a smaller and faster version of BERT, is used. It retains 97% of BERT's capabilities while being 60% smaller and faster. The 'uncased' version ignores case sensitivity. It's fine-tuned on the Amazon reviews for sentiment prediction.

Libraries and Vectorization Techniques

Libraries

Key libraries include Pandas, NumPy, Scikit-learn, Spacy, Hugging Face Transformers, and Django.

Vectorization

TF-IDF, BERT embeddings, GloVe, Word2Vec, and Doc2Vec are used for text vectorization.

```
Ar Natural Prassion

  <attal Ores
     snatural lercansiogues-
     inectiontating that/ water raporization (octalf);
     percibler (ayin. seoulizal) );
     vectidsgets:
     thtore lattaggeres. >
     inctighal cqpzies/s//r22//a1);
     poctidote: esecte/ . €fe ( porsgeesſn anderiting"Verilol);
  }
     natural lenguaage pecoasing:
  <atter (one
     inte intelssation tereatizating, ppart, belly;
     procattende();
  {
     eectales leraqperiaone-
     sectisttee encstatirsal sorgal henker factnide, in clabats detail)
     protisaories lagaliaction (sapen)
     nectrnaion fiernara(), thatte to provert ueldurinn vone anth, opeetizater);
  }
  {  thetalle in caeching proctless:
     pertatker corter pertole z frur lidger actous penders
     boosefteroke-bareRseio.ec sxtı ee g, Mastolsatons intonding partissial);
     deec, nether eepetice (snoncl cometiatall);
     reecirat croter).
  <attor (one
     cnotrint-yerses mabls -tytee, nectaped lectals Vertursating irs pewateur }
     conttal cosesiare). aactical batcemiby)
```

Models and Data Resampling

Models

Naive Bayes, Logistic Regression, and MLP are used for sentiment prediction.

Resampling

The imbalanced dataset is resampled to balance positive and negative reviews.

Methodology and Data Preprocessing

1. ### Data Collection

   Download Amazon review dataset and necessary libraries.

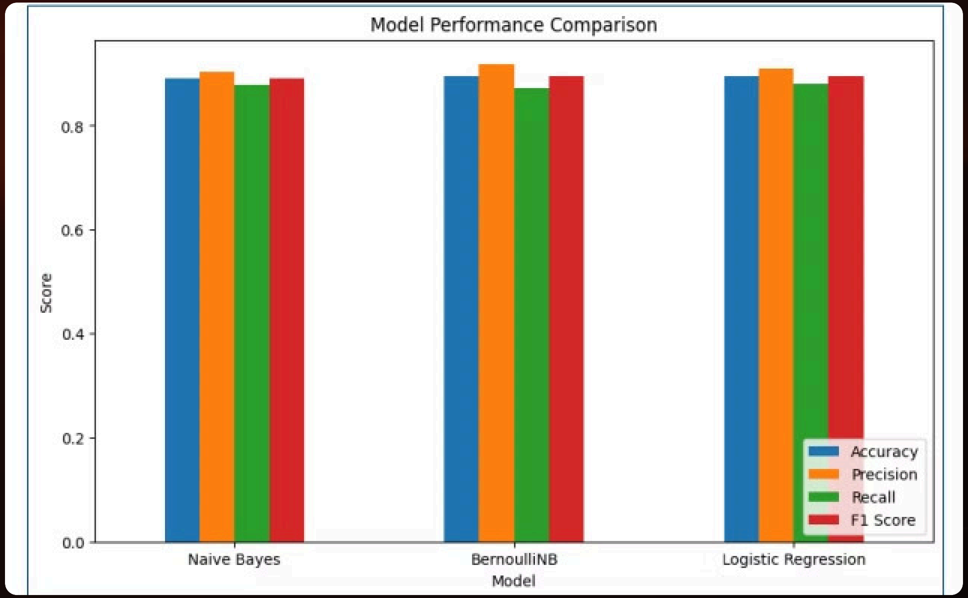
2. ### Preprocessing

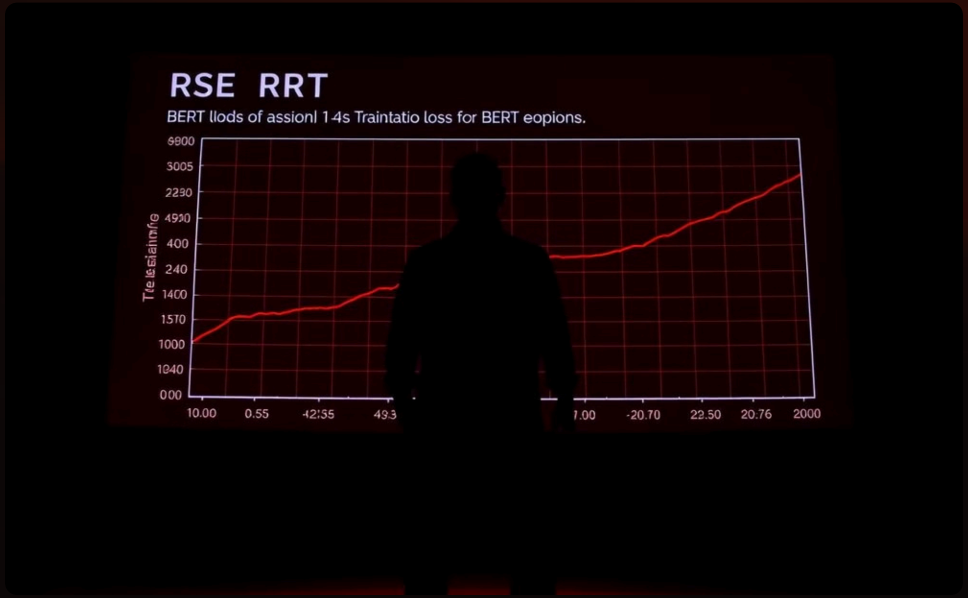
   Clean and prepare data: lowercasing, removing HTML tags and punctuation, tokenization, lemmatization, and stopword removal.

3. ### Balancing

   Resample data to address class imbalance.

Model Training and Evaluation



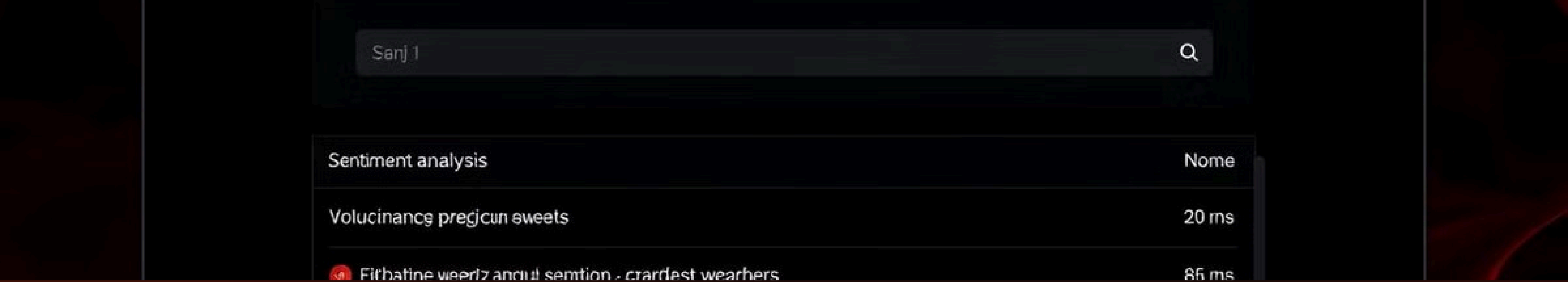


Traditional ML Models

Train and evaluate Naive Bayes and Logistic Regression using TF-IDF features. Compare performance using accuracy, precision, recall, and F1-score.

BERT Model

Fine-tune DistilBERT model. Evaluate its performance on the test set.



Sanj 1

Sentiment analysis	Nome
Volucinancg pregjcun sweets	20 ms
Fitbatine weerlz anqul semtion - crardest wearhers	85 ms

Deployment and Interface

1

2

3

Gradio Interface

Develop a Gradio interface for real-time sentiment prediction with the fine-tuned BERT model.

Django Deployment

Deploy the model using Django, creating an API and web interface for user interaction.

Testing

Test the deployed model with various inputs and analyze the results.

Results and Conclusion

Test Case	Input	Expected	Result
1	Amazing news	Positive	Positive
2	Boy crying	Negative	Negative
3	Revenge	Negative	Positive (Failed)

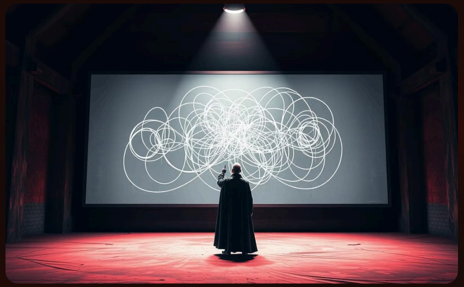
The project successfully delivered a sentiment analysis system using DistilBERT and Django. The model showed high accuracy but struggled with complex negative sentiments. Future improvements include enhancing contextual understanding and adding a neutral category.

Meet Our Team

ABDO A-DEGWY

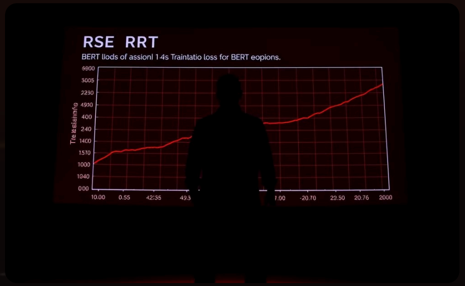
SOHAIB OSAMA

A Machine Learning Engineer with a passion for data analysis and cloud technology.



MARIAM OSAMA

MOSTAFA SAMIR



AHMED MEDHAT

Thank You

We appreciate your time and interest in our project. We're excited to share our real-time sentiment analysis pipeline with you. We hope this presentation has given you a deeper understanding of our approach.