Hate Speech analysis for DHL express

Prepared by

Naitik Shukla

Senior Data Scientist

For – Interview(Apr_2024)

Introduction and Key Points

Problem: The growing presence of online hate speech and abusive language on social media and online platforms is a significant issue that affects the quality of online discourse and interactions.

Impact: Online hate speech can lead to cyberbullying, harassment, and toxicity, creating an unwelcoming environment for users and hindering constructive conversations and civic participation.

NLP Community: With its focus on computational analysis of language, the NLP community has the potential to make a significant impact on addressing the problem of online hate speech and abusive language.

Problem and Impact

Growing Problem: Online hate speech is becoming more widespread and accepted, which can lead to a toxic online environment.

Negative Consequences: Exposure to hate speech can have detrimental effects on mental health, self-esteem, and social relationships.

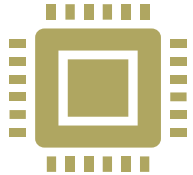
Marginalized Groups: Online hate speech can disproportionately affect marginalized groups, leading to further social exclusion and discrimination.

Real-World Impact: Online hate speech can contribute to an increase in offline violence and discrimination, perpetuating harmful stereotypes and prejudices.

NLP Techniques for Hate Speech Detection

NLP techniques can be employed to automatically detect and classify online hate speech, contributing to safer online environments.

Various approaches have been proposed, ranging from traditional machine learning methods to more recent deep learning and transformer-based models.



Some key NLP techniques for hate speech detection include:

Text classification using machine learning algorithms (e.g., SVM, Naïve Bayes, Random Forest)

Word embeddings for representing text data (e.g., Word2Vec, GloVe)

Contextual embeddings using transformer-based models (e.g., BERT, RoBERTa)

Adversarial techniques for data augmentation and model robustness (e.g., back-translation, adversarial training)

Neural models such as RNNs, CNNs, and **transformer-based models** (e.g., BERT, RoBERTa)

Applications

- NLP techniques for hate speech detection have been applied in various real-world scenarios, demonstrating their potential for addressing online hate speech.
- Some case studies and applications include:
  - **Moderating comments on social media platforms:** NLP models can be employed to automatically detect and flag potentially harmful content, enabling platforms to maintain a safer environment for their users.
  - **Identifying hate speech in code-mixed languages:** Techniques such as transfer learning and multilingual models can be used to detect hate speech in languages where resources are limited, or no annotated data is available.
  - **Analyzing the impact of counter-speech initiatives**: NLP can be used to study the effectiveness of counter-speech strategies and assess their potential for reducing the prevalence of online hate speech.

Let's Dive in more details about current technology landscape to handle hate speech

Techniques used in Market.

Text Classification using Machine Learning Algorithms

Machine learning algorithms, such as **SVM, Naïve Bayes, and Random Forest,** can be used for text classification tasks, including hate speech detection.

These algorithms analyze text data and assign labels based on patterns and features within the data.

Key steps in text classification include **data preprocessing, feature extraction, and model training and evaluation**.

Word Embeddings for Representing Text Data

Word embeddings are a way to represent text data by mapping words to continuous vector representations, capturing semantic meaning and relationships between words.

Techniques like **Word2Vec and GloVe** are popular methods for generating word embeddings.

Word embeddings can be used as input features for various NLP tasks, including hate speech detection.

Contextual Embeddings using Transformer-based Models

Transformer-based models, such as **BERT and RoBERTa**, can generate contextual embeddings that capture the meaning of words within a given context, improving hate speech detection performance.

These models are pre-trained on large amounts of text data, learning rich representations of language that can be fine-tuned for specific tasks.

Key steps in using contextual embeddings for hate speech detection include **model pre-training, fine-tuning, and model evaluation**.

Adversarial Techniques for Data Augmentation and Model Robustness

Data Augmentation: Adversarial techniques can be used to generate synthetic data, increasing the size and diversity of the training set and improving model performance.

Adversarial techniques, including **back-translation and adversarial training**, can be employed for data augmentation and to enhance the robustness of hate speech detection models.

These techniques aim to improve model generalization and performance by exposing models to perturbations and variations in the input data.

Neural Models for Hate Speech Detection

Neural models, such as **RNNs, CNNs, and transformer-based models,** have shown promising results in detecting and classifying hate speech in online text.

These models can capture complex patterns and relationships within text data, leading to improved performance in hate speech detection tasks.

Neural models can be combined with other techniques, such as **adversarial training** and **transfer learning,** to further enhance their performance in detecting and mitigating online hate speech.

Dig Deeper into our Implementation

Transformer Model - Distilbert-base for Hate speech classification
(Hate/No Hate)

Our Setup

Model:

- DistilBERT-base
- 66M parameters

Data:

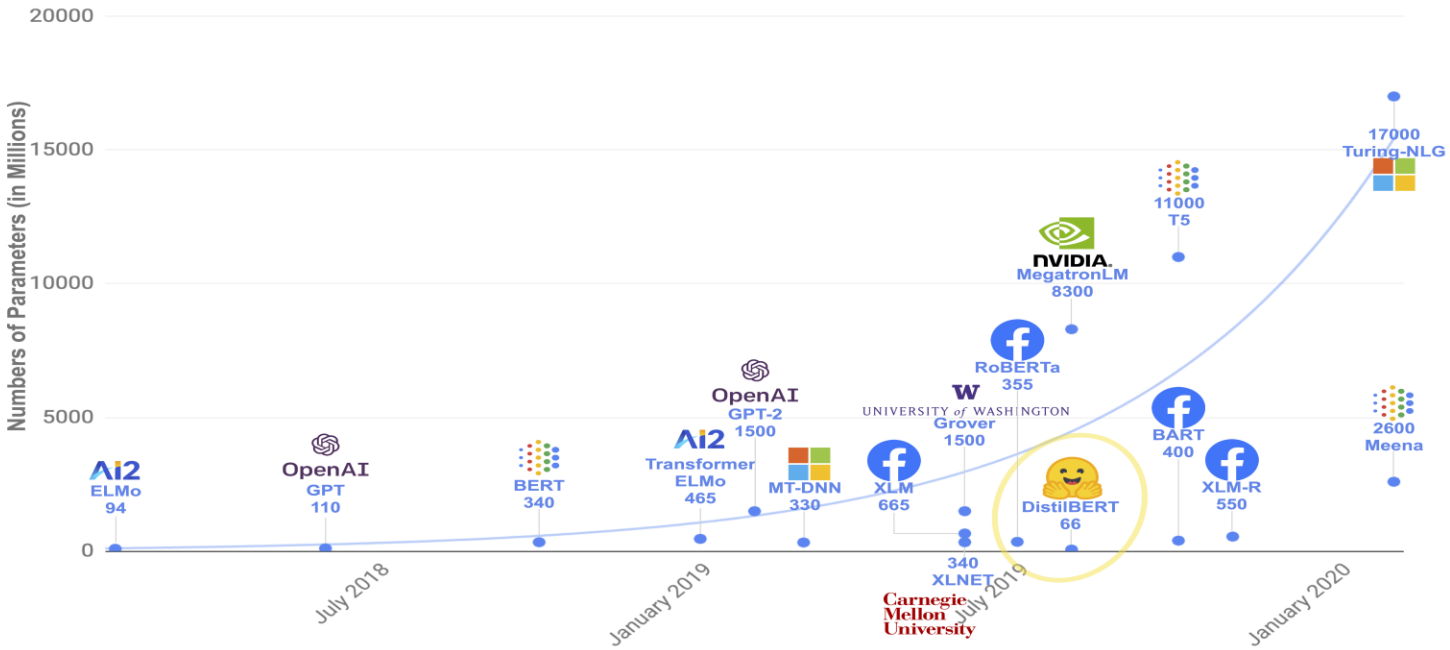
- Hate speech dataset from a white supremacist forum - Stormfront (Link)
- Train – 1913 labelled data
- Test – 300 labelled data
- Class – 2 (Hate/NoHate) Balanced

Compute

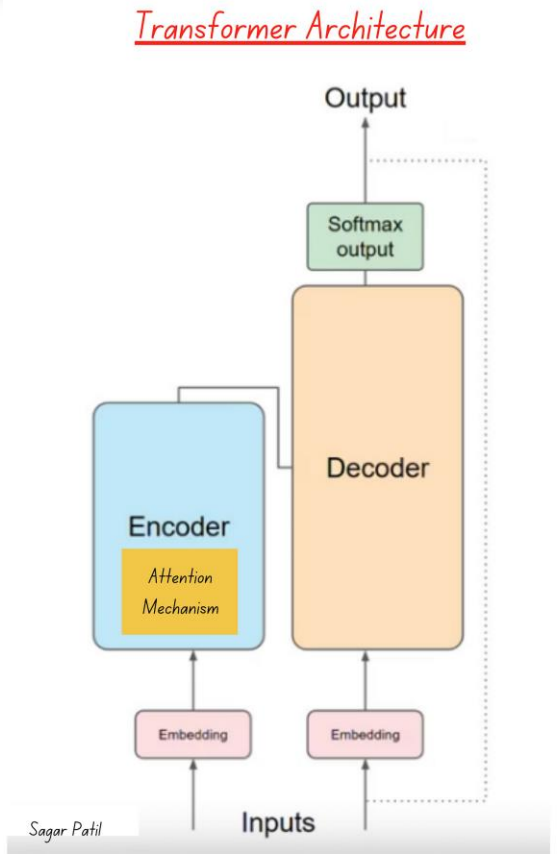
- PyTorch
- Python

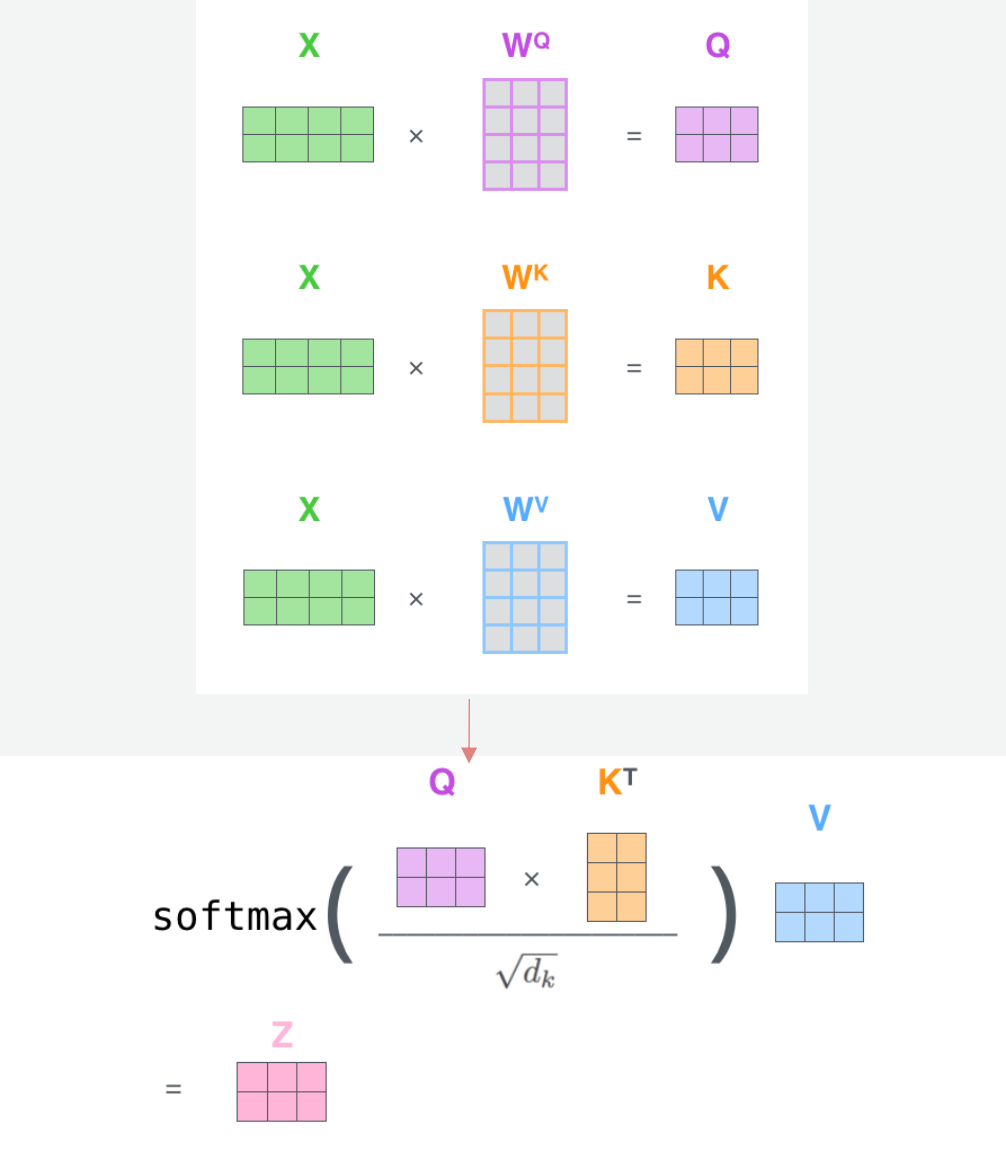
Link to GH repo

At Core of our Models lies… **Transformers**



Timeline of popular Transformer model releases





Self Attention in Matrix Form

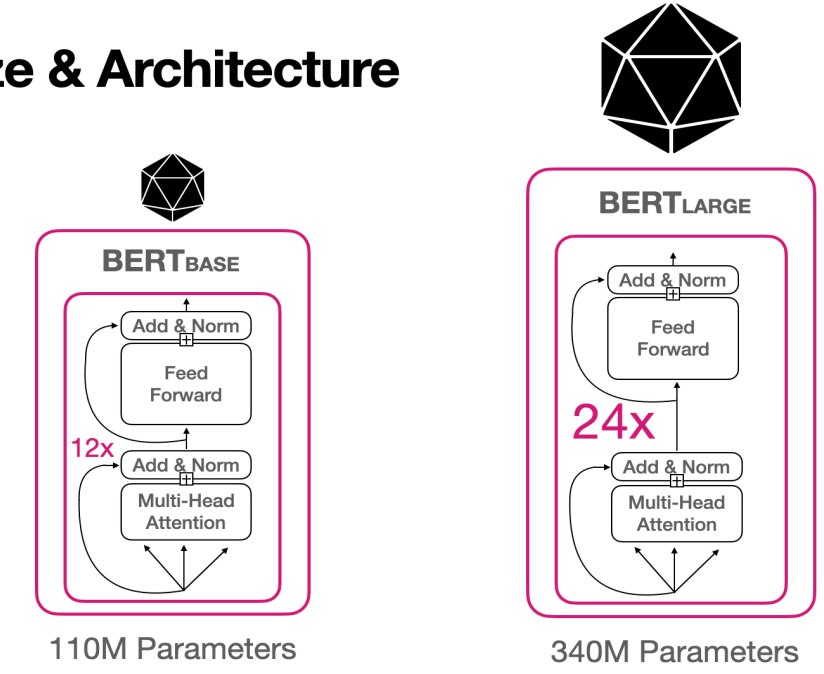
Attention Is All You Need!

- Attention is a mechanism used in neural networks to selectively focus on certain parts of the input when processing it.
- In natural language processing, attention is used to weigh the importance of different words or sub-phrases in a sentence when generating a prediction or output.
- Attention mechanisms have been shown to improve the performance of neural models on a variety of NLP tasks, such as machine translation, text classification, and sentiment analysis.

What is BERT?

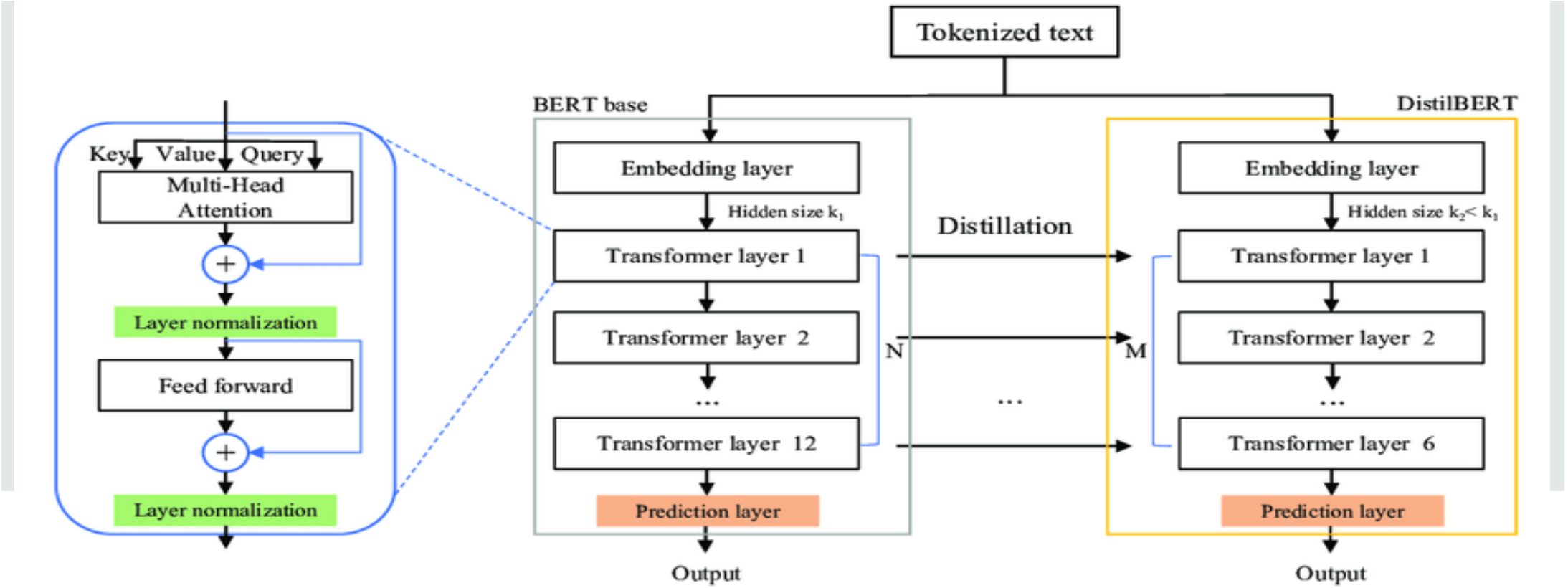
Bidirectional **E**ncoder **R**epresentations from **T**ransformers !!

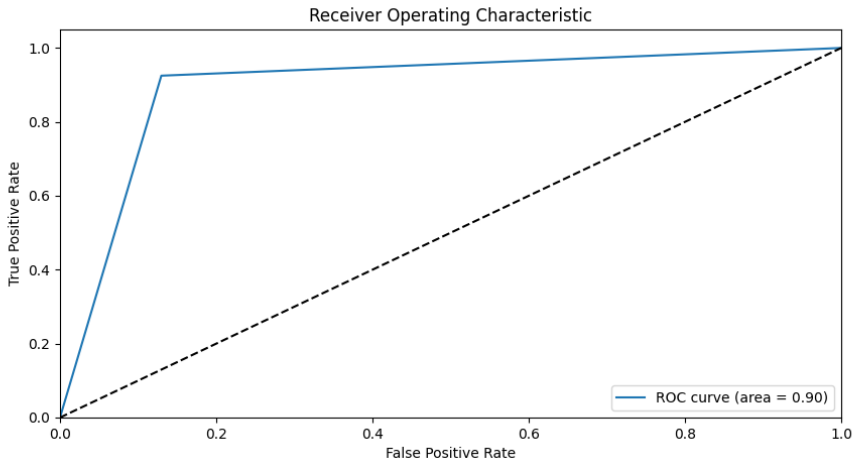
BERT Size & Architecture

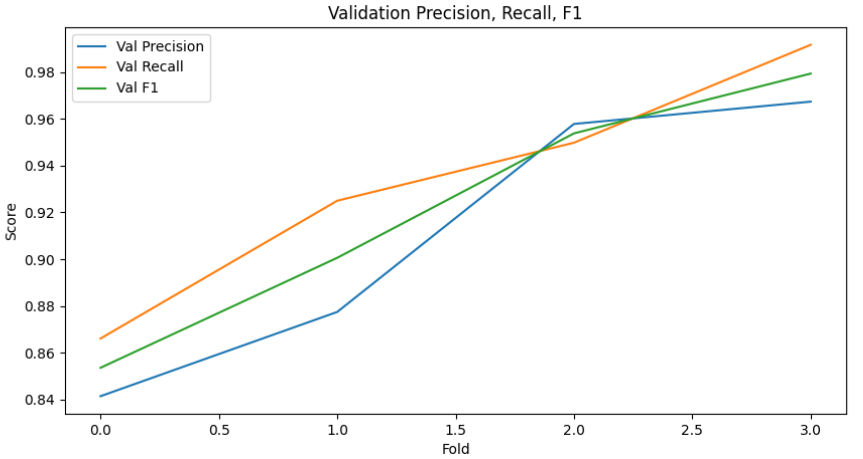


Available Model for BERT

What is DistilBERT: A Compact Transformer Model







Experimentation Result

Best Model Fold: 3

Epoch 1/2,

Train Loss: 0.26, Train Acc: 0.014, Val Loss: 0.16, Val Acc: 0.004

Epoch 2/2,

Train Loss: 0.12, Train Acc: 0.011, Val Loss: 0.15, Val Acc: 0.004

Confusion Matrix	Actual Positive	Actual Negative
Predicted Positive	229(TP)	10(FP)
Predicted Negative	12(FN)	227(TN)
ROC-AUC score	**0.95**	

Future pending work(backlog)..

- Fully Customized Classification head after Sequence Encoding of Transformer.
- Modify Code for extracting Best Parameter from Kfold and train model with full data.
- API endpoint for Model finetuning, with train file upload functionality.
- Custom/different loss for different use cases.
- Handle for unbalance dataset -
  - Up sample minority class.
  - Down sample majority class
- ....

Possible Usage in Business logic

Personalization:

DHL Express can use machine learning and AI to personalize customer experiences based on their feedback and interactions with the company. By combining hate speech classification with other personalization models, such as recommendation engines or customer segmentation, DHL Express can create more tailored and relevant experiences for each customer, while also ensuring that all interactions are respectful and appropriate.

Predictive Analytics:

DHL Express can use predictive analytics to identify customers who are at risk of becoming dissatisfied or churned based on their feedback and interactions with the company. By combining hate speech classification with other predictive models, such as customer lifetime value or churn prediction, DHL Express can proactively reach out to at-risk customers and take steps to address their concerns before they become major issues.

Thankyou