Evaluating Traditional ML Models against Transformer Architectures for Hate Speech Severity and Target Detection

1 Introduction

Social media platforms such as Twitter and Facebook have made it easy to share opinions, but they have also increased the spread of hate speech, offensive language, and abusive comments. Manually filtering such content is impossible due to the large number of posts made every second. This project aims to design an *automatic hate speech detection system* using NLP that can identify hateful text, classify its severity, and determine its target category.

2 Problem Statement and Objectives

The study focuses on developing a two-stage hate speech classification pipeline. In the first stage, offensive language identification determines whether a text is offensive or not. The second stage first determines (if offensive), if it targets individual or is untargated and then classifies (if targeted) the target of the offensive or hateful content as directed toward an individual, a group, or others. The objective is to compare the performance of traditional machine learning models - Logistic Regression and Multinomial Naive Bayes with a Transformer-based model, DistilBERT, to assess how deep contextual embeddings improve classification accuracy and contextual understanding compared to conventional feature-based approaches.

3 Dataset Description

The experiments used the **OLID (Offensive Language Identification Dataset)** from Kaggle, consisting of labeled tweets across three subtasks that represent increasing complexity in offensive language detection.

Stage-1 (Subtask A): Offensive vs Non-Offensive tweets — **NOT: 8,840**, **OFF: 4,400**. The data shows moderate imbalance, with roughly twice as many non-offensive samples.

Stage-2A (Subtask B): Targeted (TIN) vs Untargeted (UNT) offensive tweets — **TIN: 3,876**, **UNT: 524**, indicating most offensive tweets are targeted.

Stage-2B (Subtask C): Target type classification — **IND: 2,407**, **GRP: 1,074**, **OTH: 395**, with about **9,364** tweets inapplicable from earlier stages. Most offensive content is directed at individuals.

Overall, the dataset is imbalanced across stages, and minority class upsampling was applied during preprocessing to ensure balanced training.

4 Solution Approach (Methodology)

4.1 Machine Learning Models

Two classical models were implemented:

- **Logistic Regression (LR):** A linear model trained on TF-IDF features.
- **Multinomial Naive Bayes (MNB):** Uses word-frequency-based probabilities for classification.

Classification Workflow:

Text → TF-IDF Vectorization → Classifier (LR/MNB) → Output

4.2 Transformer Model

The **DistilBERT** model was used as a deep learning approach. It is a smaller version of BERT that maintains similar accuracy while being faster and more efficient.

Classification Workflow:

Text → Tokenizer → DistilBERT Encoder → Classification Layer → Output

DistilBERT was fine-tuned for each subtask using local GPU resources.

4.3 Evaluation Metrics & Data Preprocessing

Evaluation metrics included **Accuracy**, **Weighted** and **Macro Precision**, **Recall**, and **F1-score**. Raw tweets were cleaned by removing URLs, mentions, hashtags, special characters, and extra spaces, then converted to lowercase. Data was split into training and testing sets (80:20), and minority upsampling addressed class imbalance.

5 Results and Analysis

Both ML and Transformer models were evaluated across three stages of the OLID dataset. The performance comparison is shown in Table 1.

Table 1: Performance comparison of models across stages

Stages	**Models**	**Macro Averaging**			**Weighted Averaging**			**Accuracy**
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	
Stage-1	LG	0.66	0.67	0.66	0.70	0.69	0.69	0.69
	MNB	0.66	0.66	0.66	0.70	0.70	0.70	0.70
	DistilBERT	**0.74**	**0.75**	**0.75**	**0.78**	**0.77**	**0.77**	**0.77**
Stage-2A	LG	0.56	**0.56**	**0.56**	**0.82**	0.82	0.82	0.82
	MNB	**0.60**	0.53	0.53	**0.82**	**0.87**	**0.83**	**0.87**
	DistilBERT	0.55	0.53	0.53	0.80	0.84	0.82	0.84
Stage-2B	LG	0.52	0.52	0.52	0.67	0.66	0.66	0.66
	MNB	0.45	0.47	0.45	0.62	0.66	0.64	0.66
	DistilBERT	**0.54**	**0.55**	**0.54**	**0.69**	**0.69**	**0.69**	**0.69**

5.1 Comparative Analysis

Across all stages, **DistilBERT** consistently outperformed traditional models in contextual understanding and overall accuracy.

In **Stage-1**, it achieved the highest **macro F1 (0.75)** and **accuracy (0.77)**, surpassing **LR (0.66)** and **MNB (0.70)**.

For **Stage-2A**, all models performed comparably, with **MNB** slightly leading (**accuracy = 0.87**) due to clearer lexical patterns, while DistilBERT remained robust (**0.84**).

In **Stage-2B**, all models showed reduced performance due to complex class overlaps, yet **DistilBERT** again led with **0.54 macro F1** and **0.69 accuracy**.

Overall, **DistilBERT** demonstrates superior contextual comprehension, while ML models remain efficient and interpretable for simpler tasks.

6 Conclusion

Across all stages, the DistilBERT model achieved the best balance between accuracy and precision. It captured implicit and context-based hate speech that traditional ML models often missed. Logistic Regression provided reliable and interpretable results, while Naive Bayes was suitable for lightweight keyword-based filtering. Overall, Transformer-based architectures such as DistilBERT offer superior contextual understanding and are ideal for modern hate speech detection systems.

End of Report