Tweet Classification Using DistilBERT

Objective

To classify tweets into binary categories (true/false or 0/1) using the **distilbert-base-uncased** transformer model. The goal is to analyze and predict the authenticity of tweets using NLP and deep learning techniques.

Dataset

- Source: Truth_Seeker_Model_Dataset.csv
- Features: **tweet, author, statement, manual_keywords, 5_label_majority_answer, 3_label_majority_answer**
- Target: BinaryNumTarget (0 = False, 1 = True)

Preprocessing

- Dropped unnecessary columns
- Encoded categorical columns using **LabelEncoder**
- Cleaned tweets by removing:
    - URLs
    - Mentions
    - Hashtags
    - Special characters
    - Converted text to lowercase

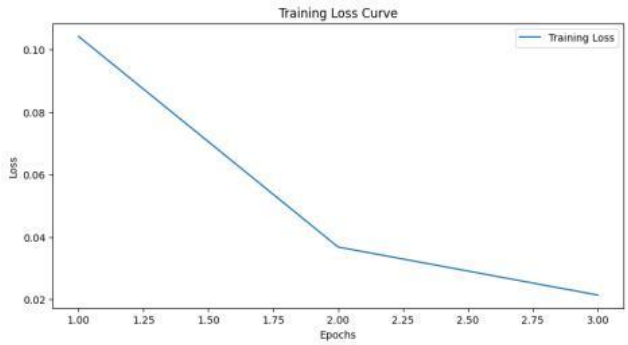
Model Used

- Pre-trained Model: distilbert-base-uncased
- Classification Head: Added for binary classification
- Tokenization: Max length = 128
- Optimizer: AdamW
- Loss Function: BCEWithLogitsLoss
- Epochs: 3
- Batch Size: 16
- **Performance**

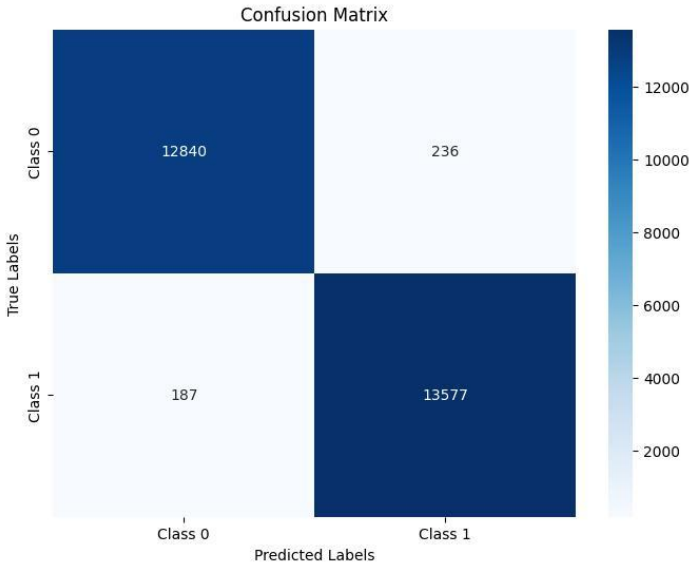
Performance

- **Metric Score**
- Accuracy 98%
- Precision 98%
- Recall 99%
- F1-Score 98%

Training Loss Curve



Evaluation



Conclusion

The model achieved **high accuracy** with a **low loss**.
DistilBERT effectively captured context and semantics of tweets.
Suitable for fact-checking and misinformation detection tasks.