BUILDING SMARTER TEXT CLASSIFIERS WITH NLP & TRANSFER LEARNING

AYSE KOCYIGIT
JITHIN KUMAR

OVERVIEW

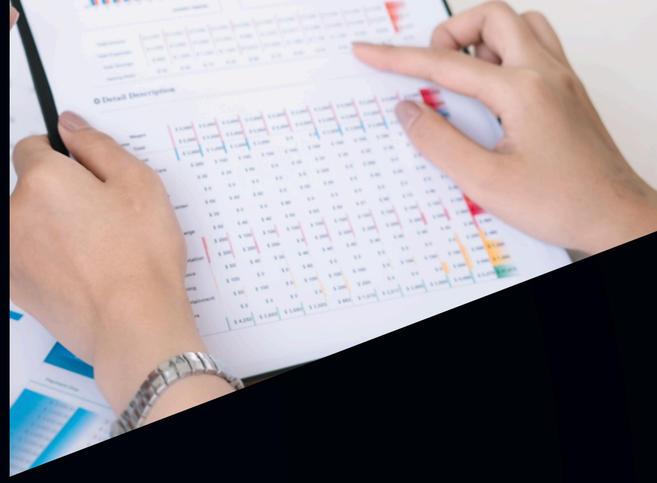
NATURAL LANGUAGE PROCESSING (NLP):

NLP is a field of AI that enables machines to understand, interpret, and generate human language. It involves steps like data collection, cleaning, feature extraction, model training, and evaluation. In this project, NLP is applied to classify fake vs. real news articles.

NLP TRANSFER LEARNING:

Instead of training models from scratch, transfer learning uses pre-trained language models (like DistilBERT) that already understand language patterns. These models are fine-tuned on the specific dataset, giving higher accuracy and faster training compared to traditional methods.

MULTINOMIAL NAIVE BAYES



Multinomial Naïve Bayes is a probabilistic machine learning algorithm widely used for text classification problems.

It works well with word frequency features, making it suitable for Natural Language Processing (NLP) tasks like sentiment analysis and spam filtering.

DATA CLEANING USED:

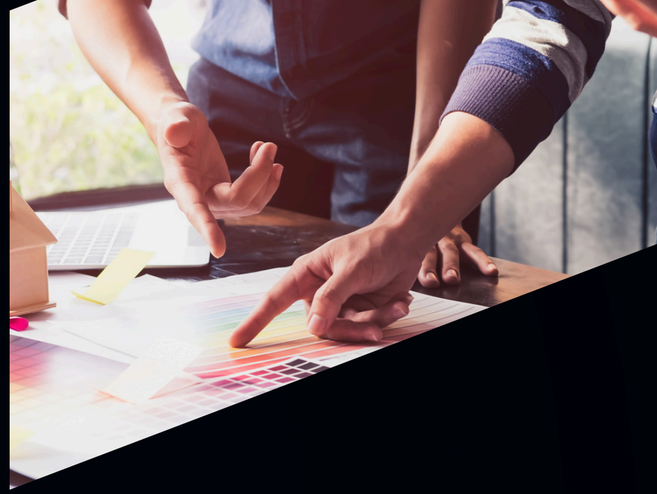
- Converted text to lowercase
- Removed URLs, punctuation, numbers, and special characters
- Stripped extra spaces

PERFORMANCE:

After applying the above preprocessing steps, the Multinomial NB model achieved an accuracy of 94%.

This cleaned data was then transformed using TF-IDF vectorization before training the Multinomial NB model.

WHY DISTILBERT INSTEAD OF BERT?



BERT:

- Large pre-trained language model from Google.
- Highly accurate but computationally heavy.
- Slower training and inference, requires more resources.

DISILBERT:

- A lighter, faster version of BERT created by distillation.
- 40% fewer parameters yet retains ~97% of BERT’s performance.
- 60% faster training and more memory-efficient.
- Ideal for projects with limited hardware but still aiming for high accuracy.

PERFORMANCE RESULTS & MODEL EVALUATION

98%	98%	98%	98%
OVERALL ACCURACY	PRECISION SCORE	RECALL SCORE	F1 SCORE

Classification Performance

- Strong performance across all evaluation metrics demonstrates the effectiveness of DistilBERT for text classification tasks.
- Consistent accuracy across different text categories.
- Minimal overfitting observed during validation phase.
- Efficient inference time suitable for production use.

Model Efficiency

- DistilBERT achieves near BERT-level performance while maintaining significant computational advantages for deployment scenarios.
- 60% reduction in model size compared to BERT.
- 6x faster inference speed for real-time applications.
- Lower memory footprint enables mobile deployment possibilities.

Generalization Ability

- Cross-validation results confirm robust performance across different data splits and demonstrate strong model generalization capabilities.
- Consistent performance across train–validation–test splits completely.
- Effective handling of class imbalance through sampling.
- Robust performance on previously unseen text patterns.

TECHNICAL CHALLENGES & SOLUTIONS IMPLEMENTED

CUDA COMPATIBILITY

Resolved GPU memory allocation errors by implementing batch size optimization and memory management strategies, ensuring stable training on available hardware resources.

LABEL MAPPING

Addressed inconsistent label formatting by implementing a robust preprocessing pipeline that standardizes categorical labels and handles edge cases in the classification schema.

CONFIGURATION MANAGEMENT

Fixed model configuration JSON formatting issues through systematic validation and implemented proper serialization methods for reproducible experimental setups and model deployment.

CONCLUSION

- The project successfully demonstrated the effectiveness of NLP techniques for text classification.
- Multinomial Naive Bayes provided a strong baseline with reliable performance (94% accuracy).
- DistilBERT transfer learning significantly outperformed the baseline, achieving nearly 99% accuracy, precision, recall, and F1 score, proving its robustness and efficiency.
- Technical challenges such as CUDA compatibility, label mapping, and configuration issues were effectively resolved, ensuring stable training and deployment.
- The results confirm that DistilBERT is a lightweight yet powerful alternative to BERT, offering high accuracy with reduced computational cost, making it suitable for real-world applications.


THANK YOU