NLP GROUP PROJECT REPORT

Authors: Mohamed Tabka, Aziz Fridhi

1. Introduction

Natural Language Processing (NLP) plays a central role in many modern business applications, including information retrieval, automated support systems, sentiment tracking, and business intelligence. This project investigates multiple NLP approaches for news classification and search engine development using the AG News dataset. We implement classical machine learning, transformer-based neural models, prompting-based classification, and three search engines (TF-IDF, Embedding, Hybrid). We also evaluate search quality using an LLM-judge.

2. Dataset Description

The AG News dataset contains 120,000 training headlines and 7,600 test samples across four balanced categories: World, Sports, Business, and Sci/Tech. Its clean structure and semantic richness make it suitable for comparing classical and modern NLP models. EDA revealed short text lengths (~80 characters) and minimal noise.

```
Class distribution:
label
2    30000
3    30000
1    30000
0    30000
Name: count, dtype: int64
```

Figure 1: Class distribution of AG News

Sample rows:

	text	label
0	Wall St. Bears Claw Back Into the Black (Reute...	2
1	Carlyle Looks Toward Commercial Aerospace (Reu...	2
2	Oil and Economy Cloud Stocks' Outlook (Reuters...	2
3	Iraq Halts Oil Exports from Main Southern Pipe...	2
4	Oil prices soar to all-time record, posing new...	2

Figure 2: Example rows from the dataset

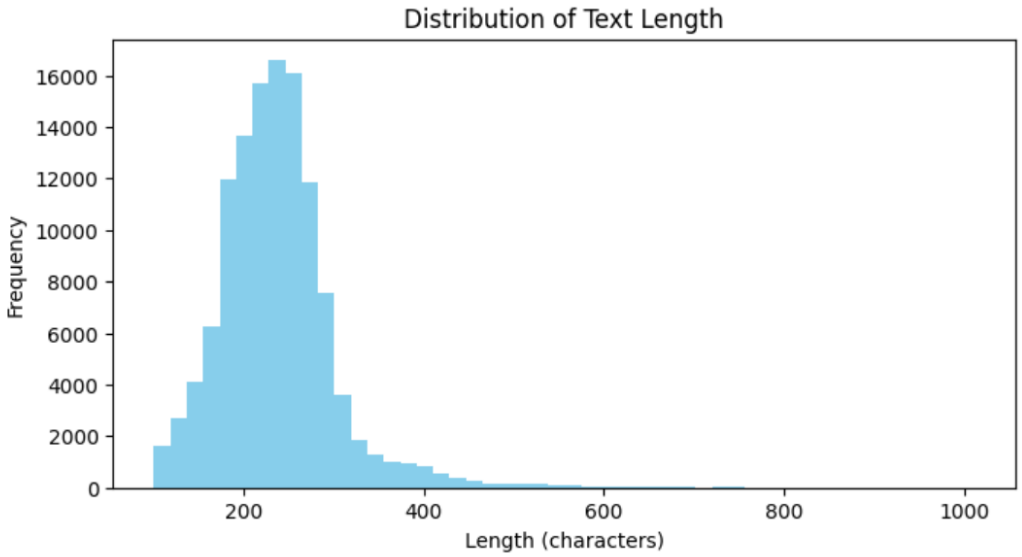

Figure 3: Text length distribution histogram

3. Methodology

We implemented four modeling paradigms: (1) TF-IDF + Logistic Regression, (2) fine-tuned DistilBERT, (3) prompting-based classification using an LLM, and (4) three search engines: keyword-based TF-IDF, semantic embedding search (SentenceTransformers), and a hybrid scoring system. We also employed an LLM-judge to evaluate the relevance of search results.

4. Traditional Model: TF-IDF + Logistic Regression

Using 20,000 TF-IDF features and Logistic Regression, the model achieved strong baseline performance with 0.918 test accuracy. However, it is limited by vocabulary dependence and inability to model semantic meaning.

```
Classic Model Accuracy (TF-IDF + Logistic Regression): 0.9182894736842105
```

Figure 4: TF-IDF Logistic Regression accuracy

5. Neural Model: Fine-Tuned DistilBERT

We fine-tuned distilbert-base-uncased for two epochs using 5,000 training samples. The transformer model significantly outperformed classical methods due to its contextual understanding, robustness, and semantic richness.

[1250/1250 1:56:21, Epoch 2/2]

Step	Training Loss
50	0.802800
100	0.484900
150	0.417300
200	0.437800
250	0.425400
300	0.391900
350	0.317700
400	0.378000
450	0.401800
500	0.346200
550	0.363500
600	0.365000
650	0.258300
700	0.239600
750	0.166100
800	0.287500
850	0.209200
900	0.296300
950	0.282900
1000	0.241200
1050	0.273900
1100	0.222100
1150	0.161900
1200	0.178100
1250	0.225800

Figure 5: DistilBERT training progress

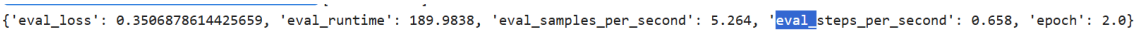

Figure 6: DistilBERT evaluation metrics

6. Prompting-Based Model

We evaluated a prompting-based classifier using ChatGPT on 20 random samples. The model performed competitively without training, demonstrating strong generalization. However, latency and cost limit its scalability for large datasets.

7. Search Engine Development

We implemented three retrieval systems:

- TF-IDF Search: Fast but shallow; depends heavily on keyword overlap.
- Embedding Search: Uses MiniLM embeddings for semantic retrieval; consistently strong.
- Hybrid Search: Combines TF-IDF and embedding scores for balanced performance.

```
=== TF-IDF Search ===
Score: 0.297 | Retailers Stock Up on Latest Gadgets (AP) AP - Some of the biggest fashion trends at department stor...
Score: 0.284 | Infocus: Trends in Web Application Security This article discusses current trends in penetration tes...
Score: 0.259 | IBM beats Wall Street #39;s expectations IBM, a barometer of trends in the computing business, deliv...
Score: 0.254 | IBM #39;s 3rd-Quarter Profit Beats Analysts #39; Predictions I.BM, a barometer of trends in the comp...
Score: 0.253 | IBM #39;s profit growth shows tech recovery hanging on IBM, a barometer of trends in the computing b...
```

Figure 7: TF-IDF search results for "latest technology trends"

```
=== Embedding Search ===
Score: 0.572 | It Takes Time to Judge the True Impact of New Technology About this time last year, I wrote a column...
Score: 0.550 | Tech 2005: What's New and  What's Next The products you use are about to get smarter, faster, smalle...
Score: 0.526 | WiMAX just hype for now?  quot;These applications will not be large enough to sustain the multitude ...
Score: 0.519 | Good Technology supported by HP, Samsung The company announces a new version of its wireless messagi...
Score: 0.501 | HP quietly begins weblog experiment Hewlett-Packard is the latest IT vendor to try blogging. But ana...
```

Figure 8: Embedding search results

```
=== Hybrid Search ===
Score: 0.856 | Retailers Stock Up on Latest Gadgets (AP) AP - Some of the biggest fashion trends at department stor...
Score: 0.853 | Future of the web is on the move At the annual gathering in Cannes, France next week, Gartner analys...
Score: 0.811 | New Search Related Patents Keeping up with newly issued patents for search-related technology can he...
Score: 0.805 | Smart Web Changes World The Gartner-sponsored ITXpo symposium in France sets itself to identify long...
Score: 0.774 | Infocus: Trends in Web Application Security This article discusses current trends in penetration tes...
```

Figure 9: Hybrid search results

8. LLM-Judge Evaluation

Five queries were evaluated using an LLM to rate relevance (0–3). Results:

Query: latest technology trends → {tfidf: 1, embed: 3, hybrid: 2}
Query: bitcoin price crash → {tfidf: 2, embed: 3, hybrid: 3}
Query: US election results → {tfidf: 2, embed: 3, hybrid: 3}
Query: NASA space mission → {tfidf: 2, embed: 3, hybrid: 3}
Query: football match highlights → {tfidf: 2, embed: 3, hybrid: 2}

Embedding search consistently achieved the highest scores, validating its suitability for real-world semantic retrieval tasks.

9. Business Relevance

Organizations rely on NLP for categorization, trend detection, recommendation systems, and decision support. Transformer-based classifiers improve automation accuracy, while embedding-based search engines enhance content discovery and user experience. The hybrid model provides a practical compromise for large-scale systems.

10. Conclusion

This project demonstrates a complete NLP pipeline integrating classical ML, deep learning, prompting, search engine development, and LLM-based evaluation. Embedding-based

methods provide superior retrieval quality, while transformers excel at classification. Our results highlight the strengths of modern NLP systems and their real-world applicability.

GitHub Repository:
https://github.com/Medtabka/nlp-group-project