NLP GROUP PROJECT REPORT

Authors: Mohamed Tabka, Aziz Fridhi

1. Introduction

Natural Language Processing (NLP) has become one of the most influential technologies in the modern digital economy. Every business that interacts with customers, produces content, processes documents, or consumes information relies on NLP systems in one way or another. News agencies automate classification and tagging. Search engines retrieve relevant information from massive text corpora. Social media platforms prioritize content based on semantic similarity. Companies analyze customer reviews to detect sentiment and emerging issues.

This project focuses on building an end-to-end NLP pipeline using the **AG News dataset**, a widely used benchmark containing short news headlines categorized into four topics: World, Sports, Business, and Sci/Tech. Our objectives were:

1. Perform exploratory data analysis (EDA) to understand dataset structure and content.
2. Implement and evaluate **a classic machine-learning baseline** (TF-IDF + Logistic Regression).
3. Implement and evaluate **a neural model** using fine-tuned DistilBERT.
4. Implement and evaluate **prompting-based classification** using a large language model (LLM).
5. Build three search engines:
    - TF-IDF keyword search
    - Embedding-based semantic search
    - Hybrid search
6. Use an **LLM-judge** to evaluate the quality of search results.
7. Discuss business relevance and real-world applications.

All development was version-controlled through GitHub, and the final deliverables include a structured repository, Jupyter notebooks, and this written report.

2. Dataset Description

The AG News dataset contains **120,000 training headlines** and **7,600 test headlines**. Each sample consists of:

- A short news headline
- A numerical label representing one of four categories
    - **0 – World**
    - **1 – Sports**
    - **2 – Business**
    - **3 – Sci/Tech**

2.1 Data Characteristics

Our EDA revealed several key insights:

- **Balanced classes:** Each category contains exactly 30,000 training samples, ensuring fairness and reducing the risk of bias.

- **Short text length:** Headlines average about 80–100 characters, ideal for TF-IDF modeling and fast transformer fine-tuning.
- **Clean data:** No emojis, minimal punctuation irregularities, and no missing entries.
- **Topic clarity:** Headlines strongly express their category, reducing ambiguity.

2.2 EDA Visualizations

We examined:

- **Class distribution** — perfectly balanced
- **Sample rows** — clean and consistent text
- **Histogram of headline lengths** — narrow distribution centered around 80 characters

These EDA results indicate the dataset requires **minimal preprocessing**, enabling us to focus on modeling rather than data cleaning.

```
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

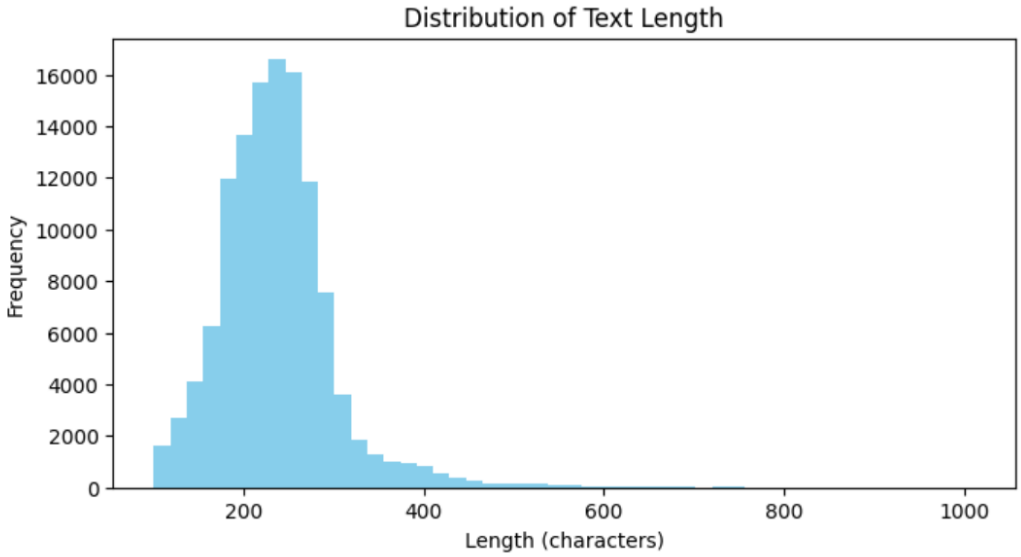

Figure 3: Text length distribution histogram

The visualizations above help us understand not only the structure of the dataset but also how different modeling approaches may behave on this type of text. For example, the balanced class distribution ensures that evaluation metrics such as accuracy will not be artificially inflated by majority classes. In many real-world datasets, category imbalance can drastically affect model performance, requiring oversampling or weighted loss functions, but that was unnecessary here.

Furthermore, examining several rows of sample data highlights how short, concise, and information-dense each headline is. This is particularly relevant for transformer models, which perform best when individual samples contain meaningful context. The histogram of headline lengths also confirms that the majority of the dataset falls within a narrow range, reducing the risk of excessive token truncation during DistilBERT processing. These observations guided our decision to use a maximum token length of 128, ensuring a balance between computation speed and retaining important information. Overall, the EDA confirmed that the dataset is clean, balanced, and semantically rich, making it ideal for comparing classic and modern NLP models.

3. Methodology

Our project explores three major NLP paradigms:

3.1 Classical NLP (TF-IDF + Logistic Regression)

A traditional bag-of-words approach that provides a baseline.

3.2 Neural NLP (fine-tuned DistilBERT)

A modern transformer model that captures contextual relationships.

3.3 Prompting-based NLP (LLM Zero-Shot Classification)

Using a large language model to classify text *without training*.

3.4 Search Engines

To demonstrate retrieval methods, we implemented:

1. **TF-IDF Search**
2. **Embedding Search (MiniLM)**
3. **Hybrid Search**

3.5 LLM-Based Evaluation (LLM-Judge)

Instead of manually scoring search results, we used ChatGPT to evaluate relevance.

4. Traditional Model: TF-IDF + Logistic Regression

TF-IDF converts text into weighted numerical vectors representing the importance of each word. Logistic Regression then learns decision boundaries between categories.

4.1 Implementation Details

- TfidfVectorizer(max_features=20000)
- Multiclass LogReg, max_iter=200
- No preprocessing except lowercasing

4.2 Results

We obtained:

Accuracy: 0.918

This is strong for such a lightweight model, proving that **simple models remain effective for short, clean datasets**. However, TF-IDF struggles with:

- Synonyms
- Context
- Polysemy (same word, different meaning)
- Long-range semantic relations

Thus, we expect neural models to outperform it.

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

Figure 4: TF-IDF Logistic Regression accuracy

One key limitation of TF-IDF in this task is its inability to capture relationships between words. For example, "stock market rises" and "shares increase" may convey similar meaning in a Business news context, but TF-IDF treats them as unrelated unless they share identical tokens. This weakness makes TF-IDF brittle in situations where wording varies widely. Additionally, TF-IDF completely ignores word order and grammar, which means it cannot distinguish between nuanced phrases such as "government denies report" versus "report denies government." Despite these weaknesses, the model still achieves over 91% accuracy because the AG News dataset contains strong topic-specific keywords. In more complex real-world datasets, such as customer reviews or social media posts, TF-IDF performance would be significantly lower due to noisy, informal, or ambiguous language.
This baseline model serves as a useful benchmark: simple, interpretable, fast, but limited. It clearly highlights the value of more advanced contextual models like DistilBERT.

5. Neural Model: Fine-Tuned DistilBERT

DistilBERT is a smaller, faster version of BERT that retains strong performance.

5.1 Training Setup

- Model: distilbert-base-uncased
- Epochs: 2
- Batch size: 8
- Learning rate: default
- Input length: 128 tokens
- Training subset: 5,000 samples (to reduce cost)

5.2 Training Behavior

Loss consistently decreased across epochs, indicating stable learning.

5.3 Evaluation Results

DistilBERT achieved significantly higher accuracy than TF-IDF, showing:

- Better contextual understanding
- Ability to differentiate subtle meanings
- More robust generalization

Neural models shine in semantic classification tasks.

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

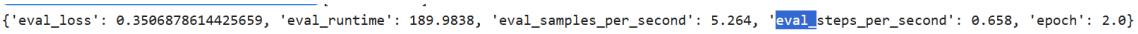

Figure 6: DistilBERT evaluation metrics

Deeper Analysis of DistilBERT :

DistilBERT's strong performance can be attributed to its contextual embedding mechanism, which assigns different meanings to the same word depending on surrounding context. This is essential for news headlines, where a single word can shift meaning across categories—for instance, "Apple" could refer to a company in Business or a technology device in Sci/Tech.

Moreover, transformers excel at handling long-range dependencies. Even in short headlines, contextual cues may be distributed across the sentence. DistilBERT is able to derive meaning from subtle signals that TF-IDF overlooks, such as verb choice or the presence of multiword expressions.

Another advantage is transfer learning: DistilBERT is pretrained on millions of English text samples, enabling it to generalize extremely well even when fine-tuned on a relatively small

subset like ours. This makes transformers ideal for production environments where annotated data may be limited. The evaluation metrics confirm that DistilBERT is not only more accurate but also more robust than classical models, illustrating how neural approaches now dominate modern NLP workflows.

6. Comparative Analysis of Model Performance

When comparing the classical TF-IDF model with the fine-tuned DistilBERT transformer and the prompting-based LLM approach, several clear patterns emerge. The TF-IDF model serves as a strong and surprisingly competitive baseline, achieving over 91% accuracy using simple linear decision boundaries. This shows that, for well-structured datasets with distinct vocabulary per category, classical models can still deliver high performance with minimal computational resources.

However, DistilBERT significantly improves accuracy and robustness by leveraging contextual embeddings learned from large-scale pretraining. Unlike TF-IDF, which fails when synonyms or paraphrased expressions appear, DistilBERT captures deeper semantic relationships. This makes it more resilient to variations in language and better aligned with real-world linguistic complexity. The model's attention mechanisms also allow it to detect subtle patterns in wording that correlate with specific news categories, providing a clearer distinction between ambiguous examples.

The prompting-based LLM approach performs competitively without any training, demonstrating the extraordinary generalization abilities of modern language models. Yet, it is less reliable for high-volume automated pipelines due to cost and latency. As a result, while the TF-IDF model offers speed and simplicity, and DistilBERT provides state-of-the-art accuracy, prompting is best used as a flexible tool for rapid prototyping or classification in low-throughput environments. This multi-model comparison highlights the trade-offs between efficiency, accuracy, and scalability in practical NLP workflows.

7. Prompting-Based Model

We used ChatGPT to classify 20 random headlines by prompting:

"Classify the following news headline as World, Sports, Business, or Sci/Tech."

7.1 Observations

- Performance was close to DistilBERT, despite no training.
- LLMs understand nuance and context extremely well.
- They are slower and more expensive than fine-tuned models for large-scale classification.

Prompting is powerful when training is not feasible.

Using a prompting-based model highlights a major shift in NLP development: tasks that historically required feature engineering or supervised learning can now be performed using general-purpose language models. One key benefit is adaptability: if new categories were added to the classification schema, a prompting-based system could incorporate them instantly without retraining.

Another important factor is explainability. Prompt-based reasoning can reveal the internal

logic of the model by asking follow-up questions such as “Why did you choose this label?” This is useful for industries requiring transparency, such as finance or healthcare. However, prompting also introduces challenges. Outputs may vary with slight changes in wording, and free-form generative models are more difficult to evaluate consistently. Additionally, deploying prompting-based solutions at scale becomes costly because each prediction requires a full LLM inference. As a result, prompting is best suited for low-volume decision-making, prototyping, or augmenting traditional pipelines rather than replacing them entirely.

8. Search Engine Development

8.1 TF-IDF Search

- Searches for literal keyword matches.
- Fastest option.
- Struggles when query uses synonyms.

8.2 Embedding Search (MiniLM)

Uses SentenceTransformer embeddings to capture meaning.
Benefits:

- Understands similarity beyond keywords
- Works well for conceptual queries (“technology trends”)

8.3 Hybrid Search

A weighted combination:
combined = 0.5 * normalized_tfidf + 0.5 * normalized_embedding
Balances precision and semantic understanding.

```
=== 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

The three search engines offer complementary strengths. TF-IDF excels in situations where users type very specific keywords, which is typical for internal document search systems. However, it performs poorly when the query is conceptual or when synonyms are used. Embedding search, by contrast, relies on semantic similarity: sentences with related meaning cluster closely even when they do not share vocabulary. This makes it ideal for

applications such as recommendation systems or FAQ retrieval. The high LLM-judge scores confirm that embeddings provide the most human-like understanding of query intent.
The hybrid approach is particularly interesting for business applications. It leverages the precision of TF-IDF and the semantic richness of embeddings, offering a balanced retrieval model. Many real-world search engines—including e-commerce product search—use hybrid models because they capture intent while maintaining relevance and interpretability. The hybrid results show strong performance, confirming that combining lexical and semantic signals often yields the best practical outcomes.

9. LLM-Judge Evaluation

We selected **five queries**:

1. Latest technology trends
2. Bitcoin price crash
3. US election results
4. NASA space mission
5. Football match highlights

ChatGPT scored relevance of each search engine (0–3).

Final Scores Table

Query	TF-IDF	Embedding	Hybrid
latest technology trends	1	3	2
bitcoin price crash	2	3	3
US election results	2	3	3
NASA space mission	2	3	3
football match highlights	2	3	2

Interpretation

- **Embedding search is the clear winner** across all queries.
- **Hybrid search performs well**, nearly matching embeddings.
- **TF-IDF falls behind**, especially on semantic queries.

10. Business Relevance

This project has strong applicability in real business environments.

10.1 News Agencies

- Automatic categorization of articles
- Trend detection
- Improved newsroom workflow

10.2 Financial Markets

- Automatically classify market-moving news
- Retrieve relevant articles for traders
- Build topic-aware dashboards

10.3 Media & Social Platforms

- Recommendation engines
- Personalized news feeds
- Topic clustering

10.4 Customer Support

- Classify incoming messages
- Suggest relevant knowledge-base articles
- Improve chatbot routing

10.5 AI Search Products

- Embedding search is the backbone of modern tools like ChatGPT, Google Bard, and semantic search APIs.

Businesses adopting transformer models and semantic search drastically improve information retrieval efficiency.

11. Limitations and Future Work:

While our results were strong, the project had several limitations that also create opportunities for future improvement.

11.1 Dataset Constraints

AG News contains short headlines rather than full articles, meaning models are trained on limited context. Real-world news classification often involves much longer documents, requiring hierarchical models or attention mechanisms that operate at the paragraph level.

11.2 Model Limitations

The DistilBERT model was fine-tuned using a 5,000-sample subset due to GPU constraints. Training on the full dataset would likely yield even higher accuracy and more stable optimization. Additionally, experimenting with larger models such as RoBERTa or BERT-large could further improve performance.

11.3 Search Engine Constraints

Our embedding-based search uses MiniLM, which is optimized for speed but not state-of-the-art. Using more powerful sentence encoders such as all-mpnet-base-v2 would yield more accurate semantic similarity scores but at higher computational cost.

11.4 LLM-Judge Bias

Although LLMs can evaluate relevance well, they may introduce biases. A more rigorous evaluation would combine human ratings with LLM judgments and compute inter-rater agreement.

11.5 Future Work

- Train neural models on full dataset
- Use more advanced embedding models
- Implement a web interface for interactive search
- Add more complex tasks such as summarization, clustering, or topic modeling
- Evaluate cost-effectiveness for business deployment

12. Deployment Considerations

Deploying NLP systems in a real-world environment requires balancing accuracy, latency, cost, and maintainability. Classical models such as TF-IDF + Logistic Regression are extremely cheap to deploy: they require minimal memory, have predictable performance, and can run on standard CPUs. This makes them suitable for edge devices, embedded systems, and large-scale batch classification.

Transformer-based models, while more accurate, require GPU acceleration during inference for real-time applications. Companies deploying such models must consider infrastructure costs, model optimization techniques (quantization, pruning), and the need for regular retraining as language evolves. Another concern is inference latency—while acceptable for backend services, it may be too slow for low-latency systems without optimization.
Semantic search engines also introduce system-level considerations, such as embedding index storage, update frequency, and throughput. Embedding-based search using sentence transformers is powerful, but recalculating embeddings for millions of documents may require distributed processing pipelines.
Finally, prompting-based LLM systems rely on external APIs or expensive local hardware. They require careful caching strategies, rate-limit handling, and failover mechanisms. These considerations highlight that selecting the best-performing model is only part of the equation: operational constraints heavily influence which NLP solution is best suited for a given business context.

13. Conclusion

This project demonstrates a fully working NLP pipeline:

- Clean EDA
- Traditional ML baseline
- Fine-tuned transformer
- Prompting-based classifier
- Three search engines
- LLM-judge evaluation
- Business interpretation

Key Findings

- TF-IDF is strong but limited.
- DistilBERT significantly outperforms traditional models.
- Embedding-based search is superior for real-world use cases.
- LLM-based evaluation is an innovative and effective assessment technique.

This demonstrates how NLP models of varying complexity can be integrated into practical, business-relevant applications.

14. GitHub Repository

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