



Multi-label News Classification

Tutorial: H19B



UNSW
SYDNEY



Introduction



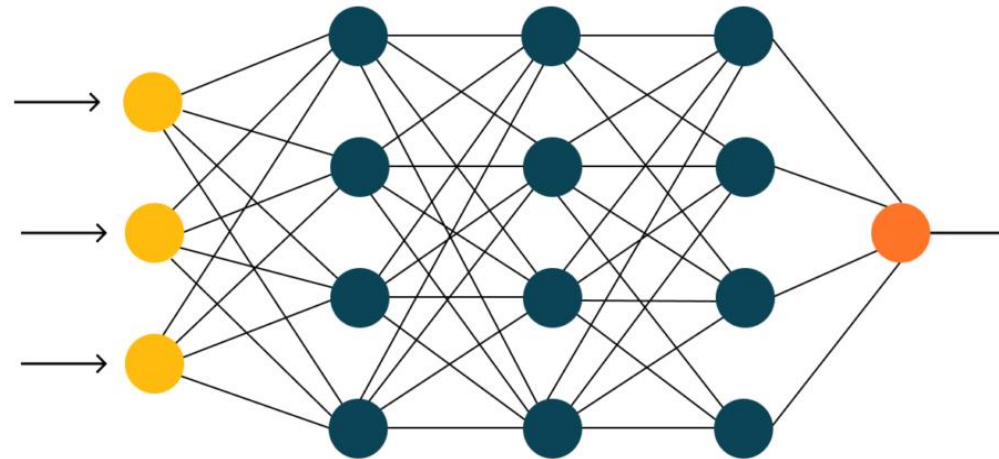
UNSW
SYDNEY



Problem Statement

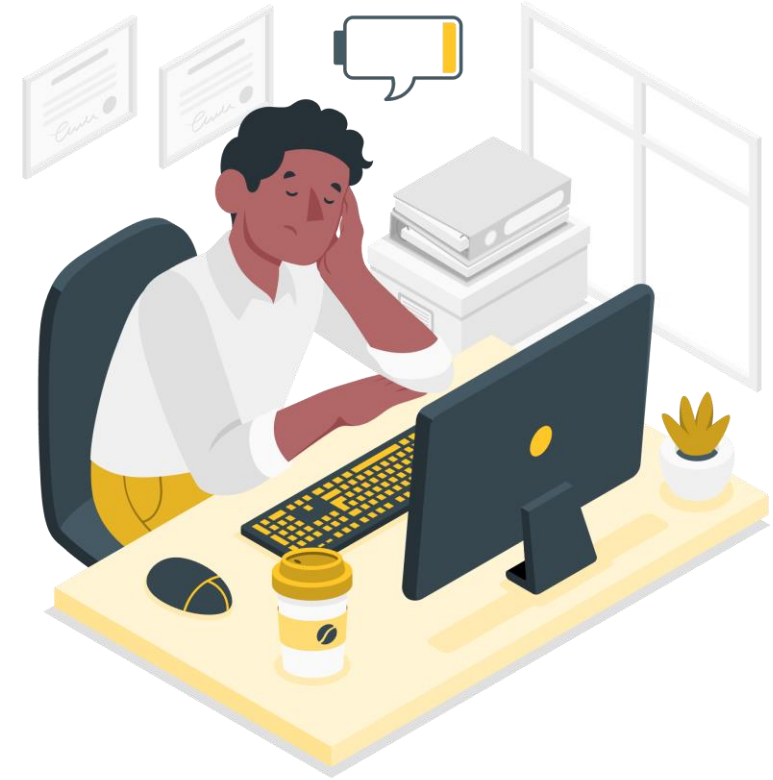
Accurately and efficiently labelling news categories to articles from the AG News dataset.

Develop Neural Network-based classification models



Motivation

- Categorisation is a fundamental process for many businesses.
- News organisations classify news to keep track of their audiences' interests and topic coverage. Quant traders and Lawyers to do quick and efficient research.
- As the volume of online news continues to grow exponentially, **manual categorisation has become impractical.**
- It requires large teams, making it time-consuming, expensive, and unsustainable at scale.
- To stay competitive, organisations must invest in automated, neural network-based systems to classify news efficiently and accurately.



Accurate and Efficient?

- Accuracy and efficiency are critical for media organisations:
 - **Accuracy:** How well the model correctly classifies articles into their respective categories.
 - **Efficiency:** The model's ability to perform quickly and with minimal computational resources.
- A highly accurate model minimises misclassifications and correctly identifies the article's topic.
- Fast training times, low memory usage, and real-time or near-real-time predictions.
- These factors are essential for scaling the model to large datasets and deploying in production environments where speed and cost-effectiveness are crucial.





Data Source



UNSW
SYDNEY

Where did the data come from?

- This project uses the **AG News dataset**, a benchmark dataset for text classification tasks. Collected from 2,000+ news sources.
- Created by Xiang Zhang, Junbo Zhao and Yann Lecun (*Character-level Convolutional Networks for Text Classification*, 2015)
- Enables consistent, reproducible training and evaluation of models.



Data Composition





Literature Review



UNSW
SYDNEY

Lit Review – Deeper Neural Networks – LSTM & CNN

- As shallow networks hit performance ceilings, deeper models like LSTMs and CNNs gained popularity.
- LSTMs (Hochreiter & Schmidhuber, 1997) addressed the vanishing gradient problem, effectively capturing long-range dependencies in text.
- LSTMs are ideal when word order and sequence are important.
- CNNs for text, adapted from image tasks, use filters to detect local n-gram patterns.
- Kim (2014) showed CNNs can perform well with fewer parameters.
- These models offer richer representations of language than simple feedforward nets, making them well-suited to news data.

Pro	Con
Capture word order (LSTMs) and key phrases (CNNs)	LSTMs are slow and hard to parallelise (not efficient)
Improved performance over traditional models	CNNs struggles with long-range dependencies

- Despite improvements, these models are limited in capturing full context—especially over long documents—motivating the shift to attention-based architectures.




Lit Review – Modern Architecture – Transformers

- Transformers revolutionised NLP with self-attention mechanisms (Vaswani et al., 2017), enabling models to capture global context.
- BERT (Devlin et al., 2019) introduced deep bidirectional context, significantly improving results on text classification benchmarks.
- Pretrained on large corpora and fine-tuned for specific tasks like news classification.
- RoBERTa (Liu et al., 2019) enhanced BERT by training longer and removing next-sentence prediction.
- Despite being computationally expensive, transformers are state-of-the-art for tasks like news categorisation.

Pro	Con
Strong understanding of context and semantics	High computational cost
Minimal feature engineering required	Large models need more data and memory

- While BERT pushes the performance ceiling, the focus now shifts to optimising transformers for scalability and efficiency in real-world applications.





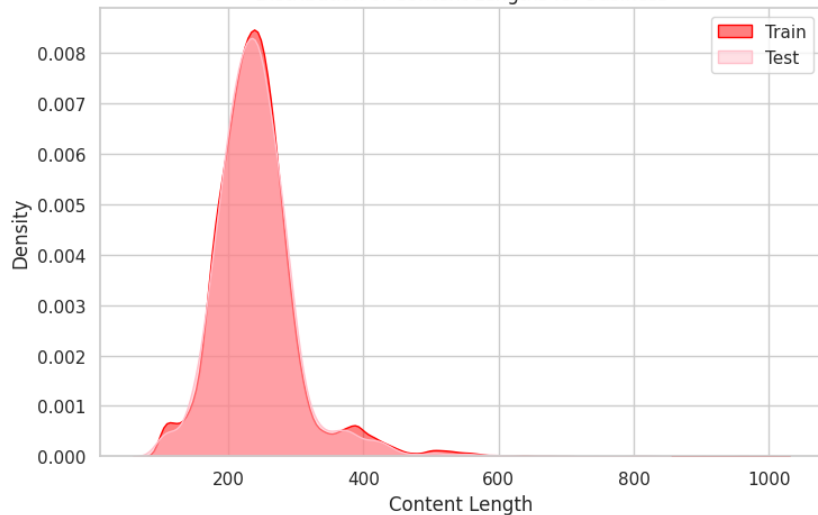
Exploratory Data Analysis



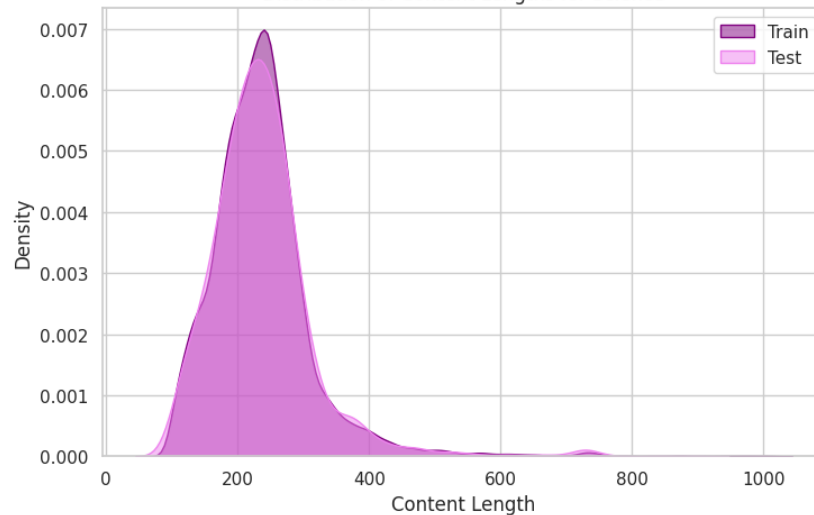
UNSW
SYDNEY

Data Length Distribution

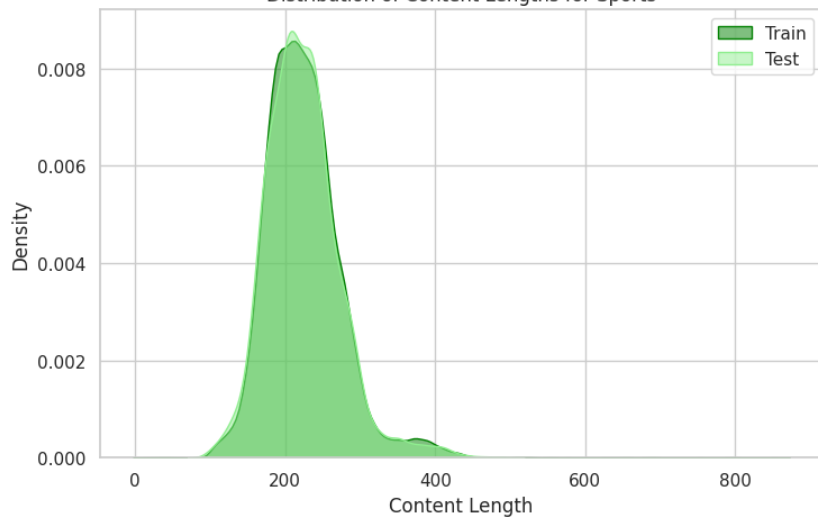
Distribution of Content Lengths for Business



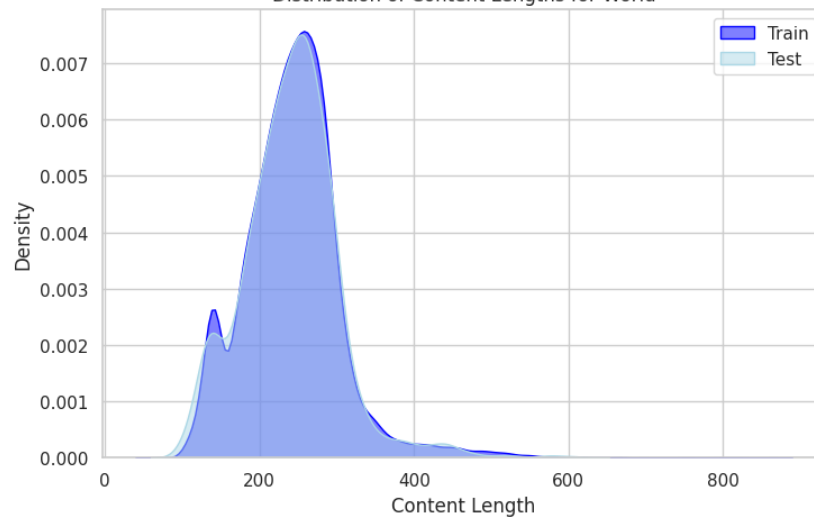
Distribution of Content Lengths for Science



Distribution of Content Lengths for Sports



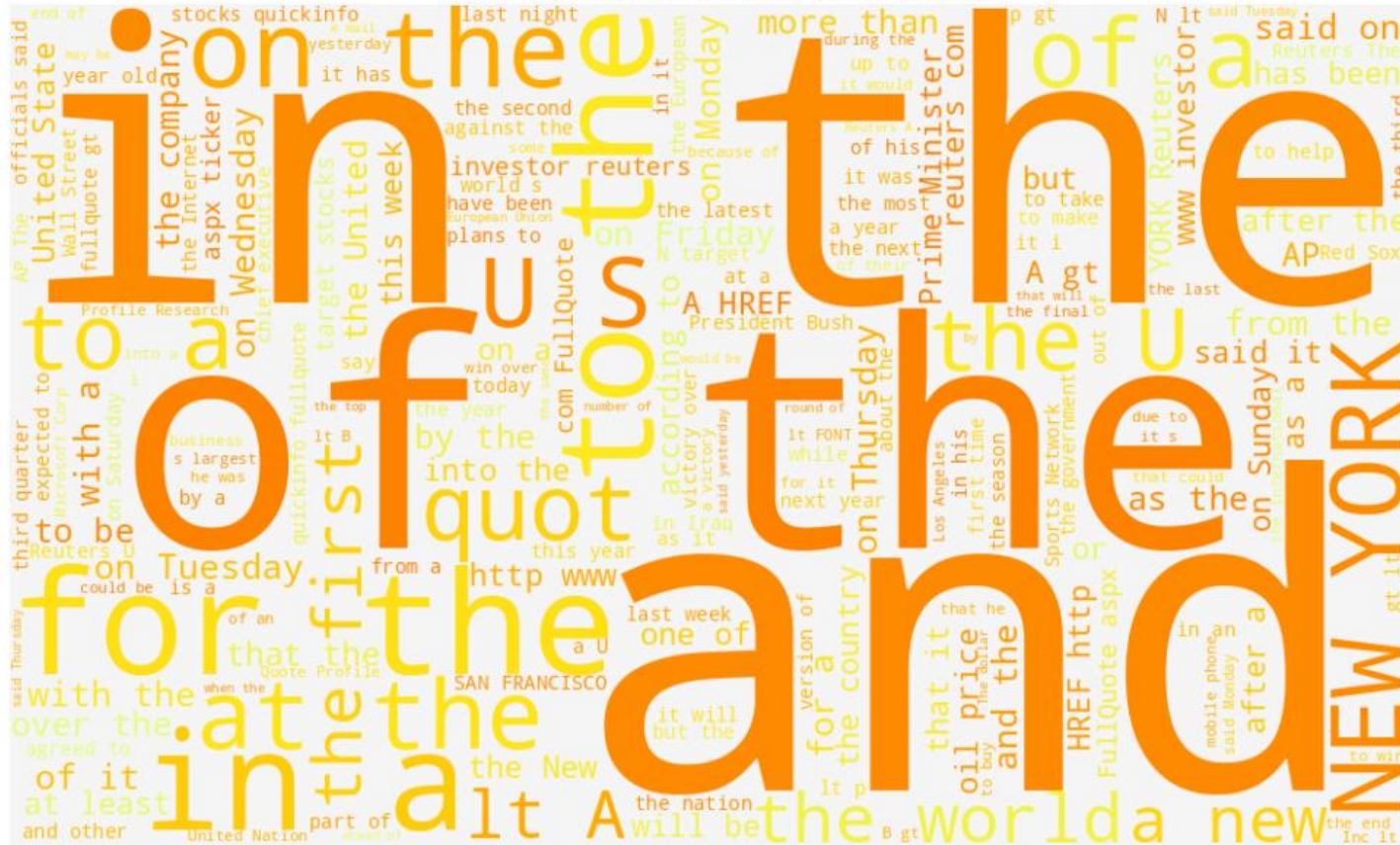
Distribution of Content Lengths for World



Label	KL Divergence
Business	0.001
Science	0.003
Sports	0.001
World	0.002

Word Map

Word Cloud Before Pre-Processing



- Raw data contains significant number of stop words
- Cloudes the networks in identifying unique patterns in sentences structures

Data Cleaning and Processing

1. Unique Character Analysis

- Exploring the data vocabulary set

2. Expand Common word to Full Form

- Replacing abbreviations with full words

3. Remove non-context adding sequences

- URLs, HTML, Stop Words and digits

4. Tokenise

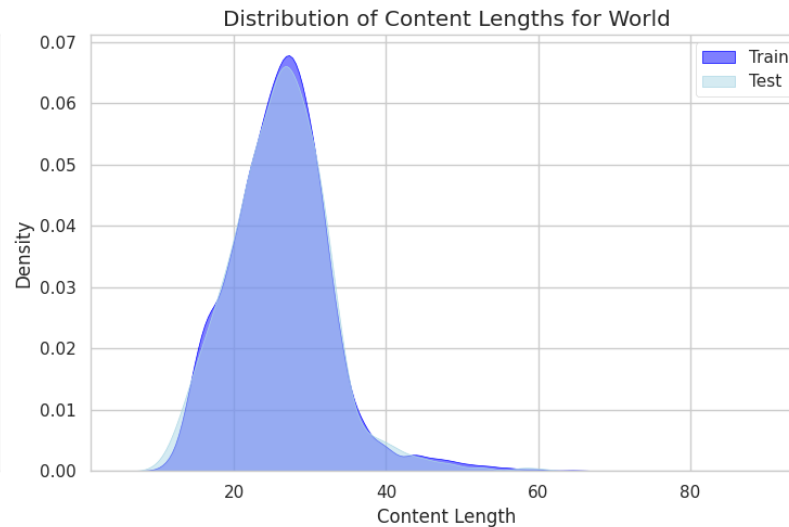
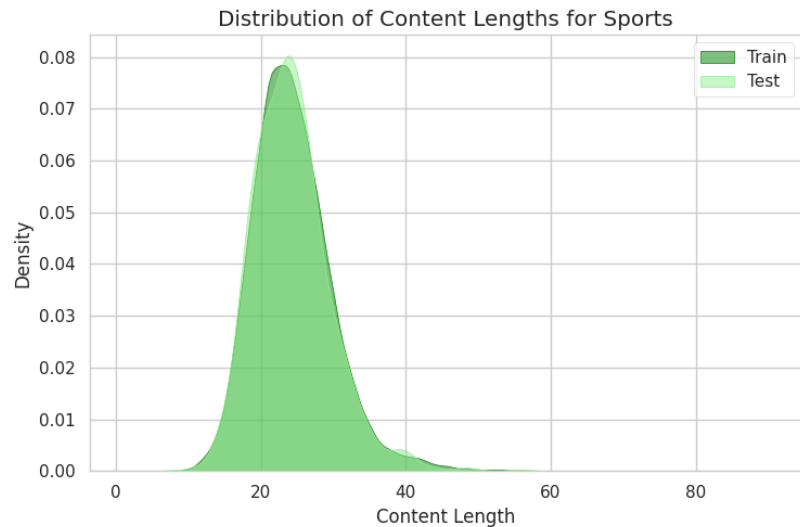
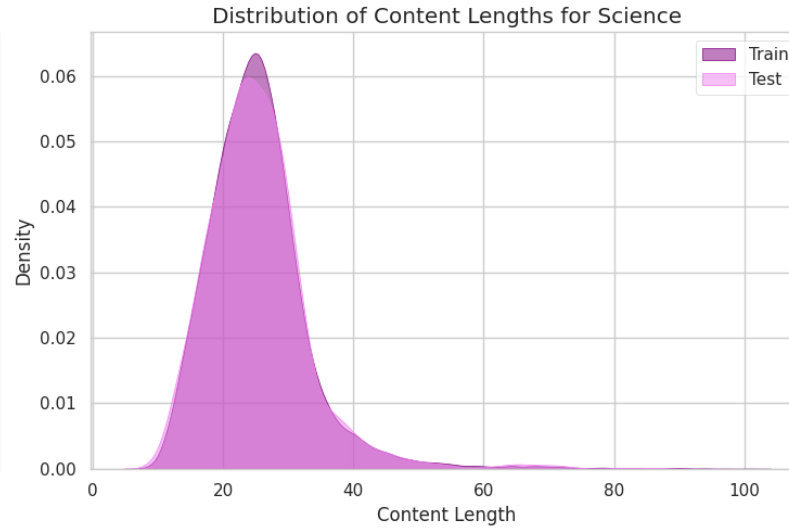
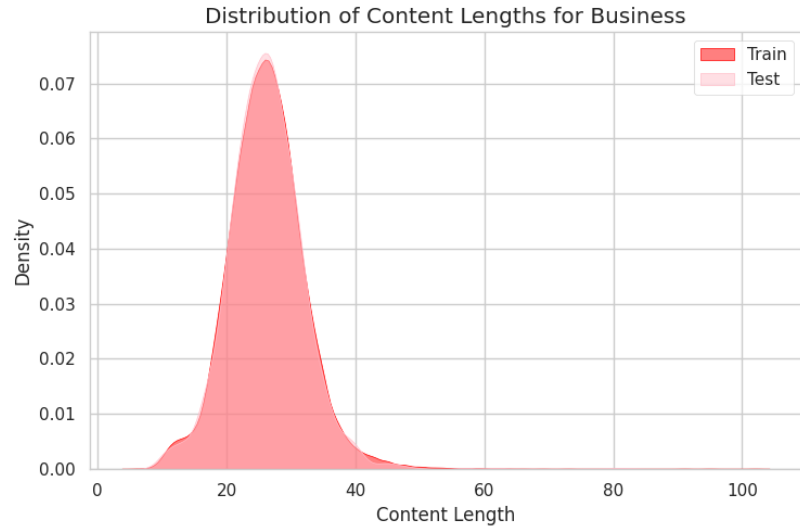
- Breaking down the strings into chunks

Word Cloud After Pre-Processing



- Most frequent words are **now more representative**
- With stop words removed and text cleaned, we observe **clearer distinctions** in word distributions
- Sharpens the feature space and improves the model's ability to learn meaningful patterns


Data Length Distribution (After Data Processing)



Label	KL Divergence
Business	0.024
Science	0.037
Sports	0.017
World	0.031

Challenging Aspects

Challenge	Description
Category Overlap	News articles often span multiple topics (e.g., tech + business), making it hard to assign a single label.
Limited Context	Titles and short descriptions can lack the depth needed for confident classification.
Language Ambiguity	Phrasing, synonyms, and sentence structure vary widely across articles.
Subtle Differences	Small changes in wording can shift an article from one category to another.
Noisy Input	Presence of stop words and non-informative boilerplate text without preprocessing.
Computational Constraints	Training deep learning models on large datasets requires significant memory, processing power, and time—posing challenges for tuning, experimentation, and deployment.



Models & Methods



UNSW
SYDNEY

Practical Benchmarks

To establish a practical foundation, we began with two benchmark models:

1. Naïve Bayes Classifier

- A simple, fast, and highly interpretable probabilistic model
- Treats words as independent features and applies Bayes' theorem for classification
- Useful as a baseline to compare against more complex neural models

2. Zero-shot classification

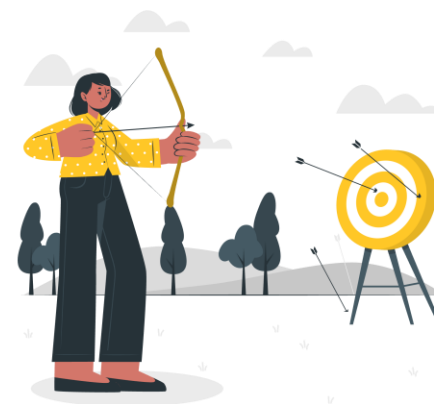
- Leveraged large pretrained language models:
 - Gemma 3
 - Gemini 2.5 Flash
 - Deepseek R1



Learnings

Base Model	Extension	Accuracy	Precision	F1-Score
Naïve Bayes	-	90.0%	90.0%	90.0%
Gemma 3: 12B	Zero-Shot	83.7%	85.2%	83.3%
	One-Shot	84.5%	85.2%	84.4%
Gemini 2.5 Flash	Zero-Shot	88.1%	88.1%	88.1%
	One-Shot	88.6%	88.5%	88.1%
Deepseek R1	Zero-shot	81.2%	82.6%	80.7%

- Naive Bayes performed surprisingly well, offering a strong traditional baseline.
- LLMs (Gemma, Gemini, DeepSeek) achieved solid results, but come with cost, control, and consistency tradeoffs.
- These findings, supported by our literature review, guided our move toward custom, trainable models with scalable architectures.
- Chosen Models – Iteration 1:
 - Simple Linear
 - Linear + Embedding
 - CNN + Embedding
 - LSTM + Embedding
 - Bi-LSTM + Embedding



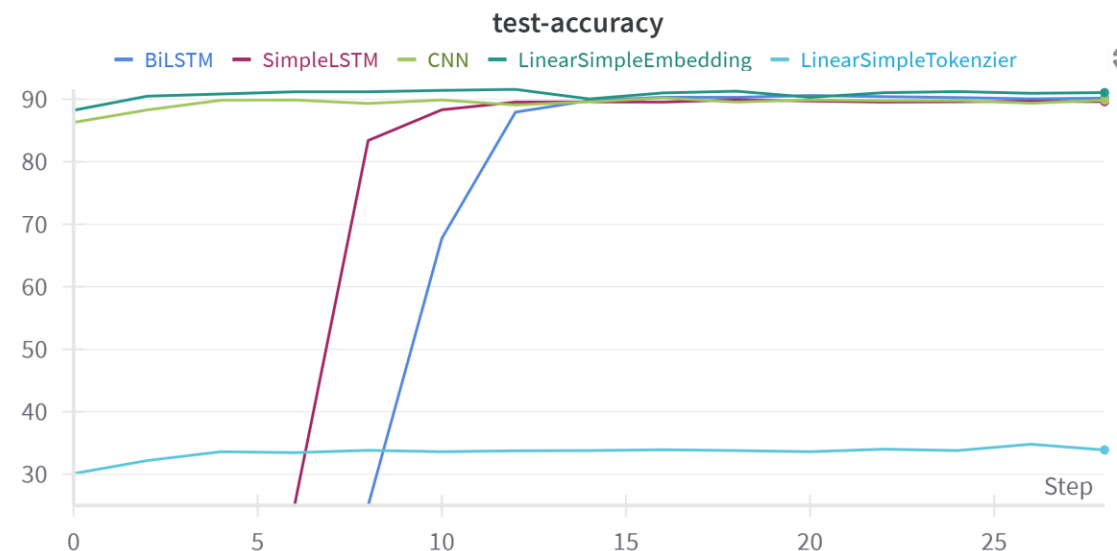
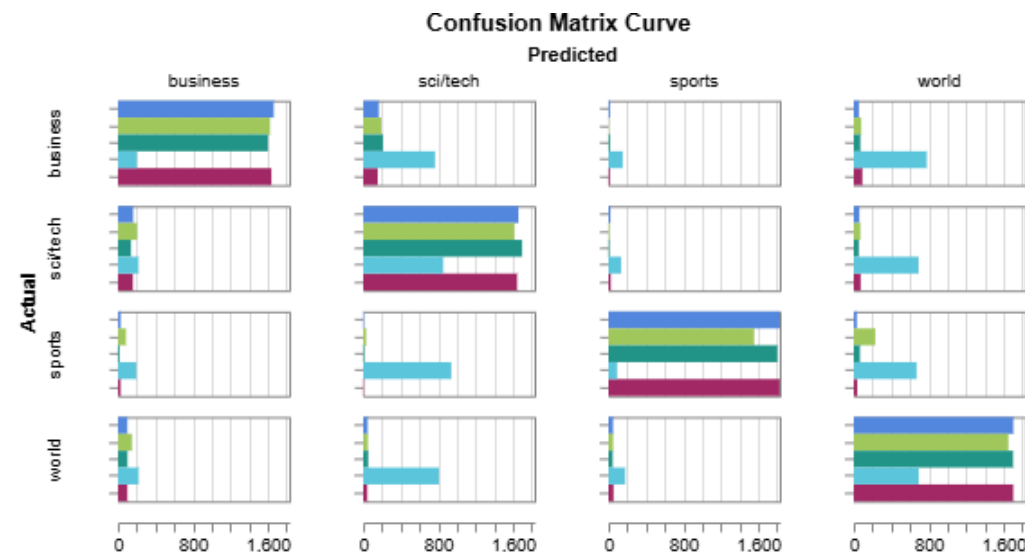
Deeper Learning Models

A summary of how we originally tackled the problem.

Model	Training Time	Parameters	Accuracy
Linear	2m	52,672	34.8%
Linear + Embedding	4m	3,010,504	90.4%
CNN + Embedding	3m	3,799,172	89.9%
LSTM + Embedding	3m	3,433,412	90.2%
Bi-LSTM + Embedding	4m	3,771,963	90.3%

Highlights:

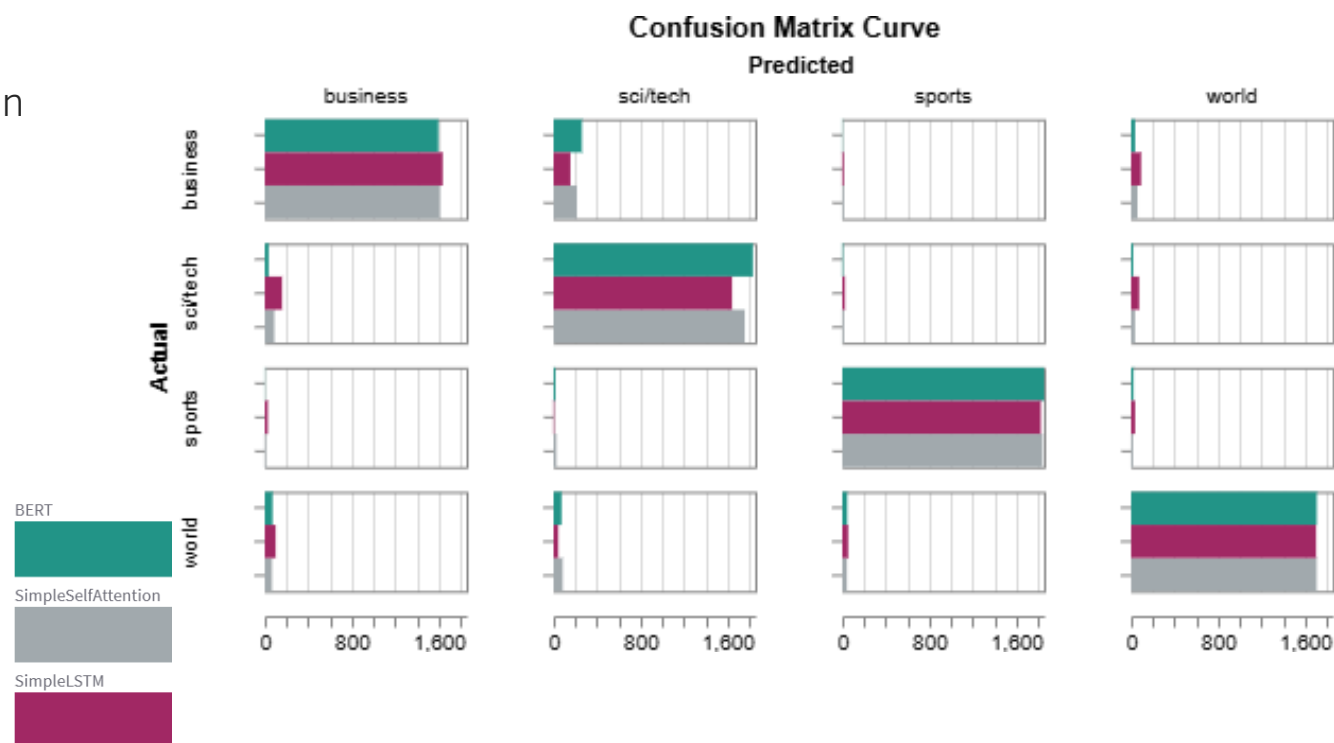
- Architectures influenced by Zhang et al. (2016), Gao et al. (2023), Wang et al. (2023), Liu (2024)
- Tokeniser only method, too few parameters
- CNN efficient to train
- Simple Linear and LSTM demonstrate higher accuracy – 91% and 90% respectively
- Confusion between tech and business



Method – Hypothesis 1

- 1st hypothesis: LSTM may not provide sufficient historical and contextual understanding for classifying news articles.
- We observed:
 - LSTM outperforms CNN, showing benefit from sequential modelling.
 - Bi-LSTM outperforms LSTM, highlighting value in bidirectional context.
 - However, performance plateaus around 90%, suggesting a limit to LSTM's capacity on this task.
- To overcome this, we tried using attention and transformer-based architectures [Vaswani et al., 2017]
 - BERT
- We note that adding attention doesn't provide any improvement.

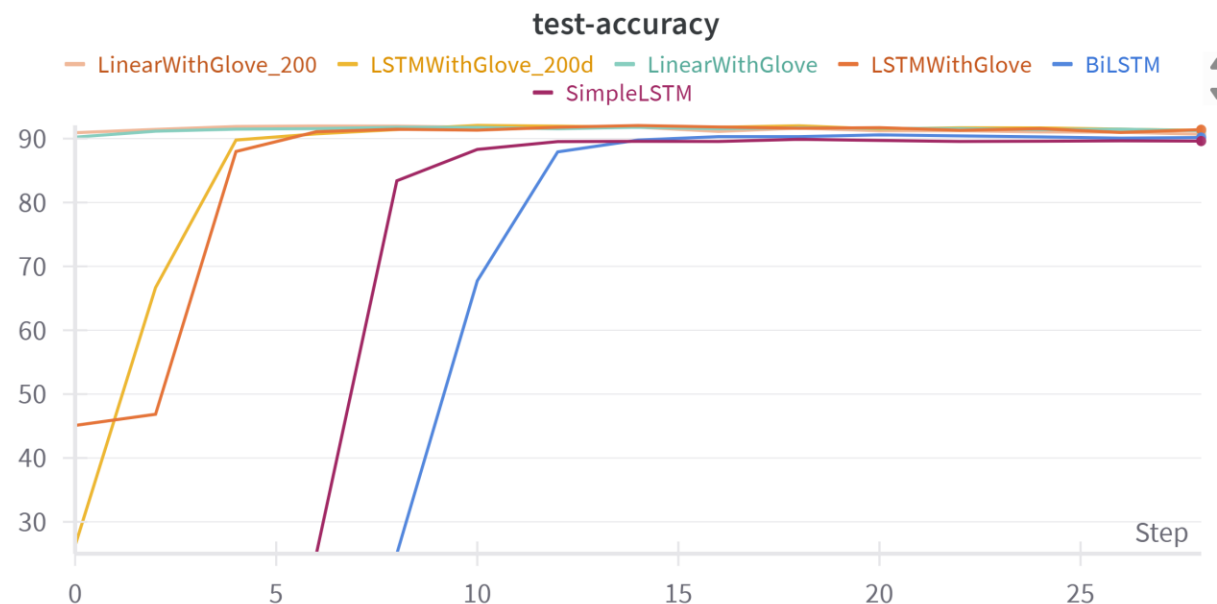
Model	Training Time	Parameters	Accuracy
Simple LSTM	3m	3,433,412	90%
Self Attention Simple	3m	7,779,716	89%
BERT	7m	109,485,316	90%



Method – Hypothesis 2

- 2nd hypothesis: Previous models may have extracted all they can from sequence patterns — performance can now be improved through richer word embeddings.
- We tested this by introducing pretrained GloVe embeddings into our models.
- Key observations:
 - GloVe *Pennington, J., Socher, R., & Manning, C. (2014)* boosts performance for both Linear and LSTM-based models.
 - Increasing embedding size (100d → 200d) offers minimal additional gain.
 - Notably, training converges faster, reducing required epochs compared to SimpleLSTM and BiLSTM.

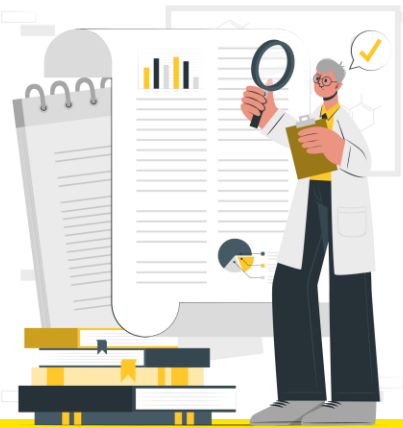
Model	Dimension	Training Time	Parameters	Accuracy
LSTM + GloVe	100d	2m	3,170,576	92.1%
	200d	2m	6,274,076	92.1%
Linear + GloVe	100d	3m	3,052,300	91.8%
	200d	3m	6,125,104	91.9%



Results

Category	Model	Training Time	Parameters	Accuracy
Benchmark	Gemini 2.5	N/A	~2.5T	88.6%
Deeper Learning	Bi-LSTM	4m	3,771,963	90.3%
Hypothesis 1: Transformer + Attention	BERT	7m	109,485,316	90.0%
Hypothesis 2: GloVe	LSTM + GloVe	3m	3,170,576	92.1%

- **Goal:** Maximise accuracy and efficiency
- **Best-performing model:** LSTM + GloVe (100d/ 200d)
 - High accuracy with fewer parameters and faster convergence.
- Bigger isn't always better
- Despite the power of transformer-based models like BERT, they did not outperform simpler LSTM + GloVe in this task.
- This suggests that task-specific tuning and representation quality can outweigh raw model complexity.



Results

Text	Actual	Gemini	Bi-LSTM	BERT	LSTM + GloVe
1994 Law Designed to Preserve Guard Jobs (AP) AP - A 1994 law strengthened job protections for National Guard and Reserve troops called to active duty. Here are major provisions of the Uniformed Services Employment and Reemployment Rights Act (USERRA).	World	Business	Business	Science	World
Promoting a Shared Vision As Michael Kaleko kept running into people who were getting older and having more vision problems, he realized he could do something about it.	Science	Sport	World	Science	Business
Justices to debate mail-order wine Being freelance wine critics may sound like a sweet gig, but Ray and Eleanor Heald have soured on it. Because their home state, Michigan, blocks direct shipments from out-of-state	Business	Science	World	World	Business
Times to Scrap Broadsheet Edition The Times is to scrap its broadsheet edition and go tabloid from Monday, it was confirmed today. The decision was made after a trial run of the compact edition proved a success, said editor Robert Thomson.	Sport	Sport	Sport	Science	Business

Discussion – Where Models Struggle?

- Why do models get confused?
 - Ambiguity in language
 - Vocabulary overlap
 - Short context windows
- Key insight: **Semantics matters**
 - Models perform better when they grasp meaning, not just sequence or pattern.
 - Deeper semantic understanding is what separates “just good” models from production-grade models.

Predicted / Truth	Article (GloVe + LSTM)
Science / Business	Important Rules for Phone Market Face F.C.C. Vote Next week, the FCC will likely change the rules on unbundled networks largely in ways favorable to the Regional bells.
Business / Science	New Rainbow Six Franchise for Spring 2005 SAN FRANCISCO, CA - November 30, 2004 - Ubisoft, one of the world #39;s largest video game publishers, today announced its plans to launch the next installment in the Tom Clancy #39;s Rainbow SixR franchise for the Sony PlayStationR2 computer entertainment system

Limitations & Future Research

Limitations:

- Limited context length: Only titles and descriptions were used — full articles might offer more signal.
- Domain generalisation: AG News is clean and balanced — real-world data may be noisier and domain-specific.
- Pretrained model constraints: Transformers used were off-the-shelf; minimal fine-tuning was done due to resource constraints.

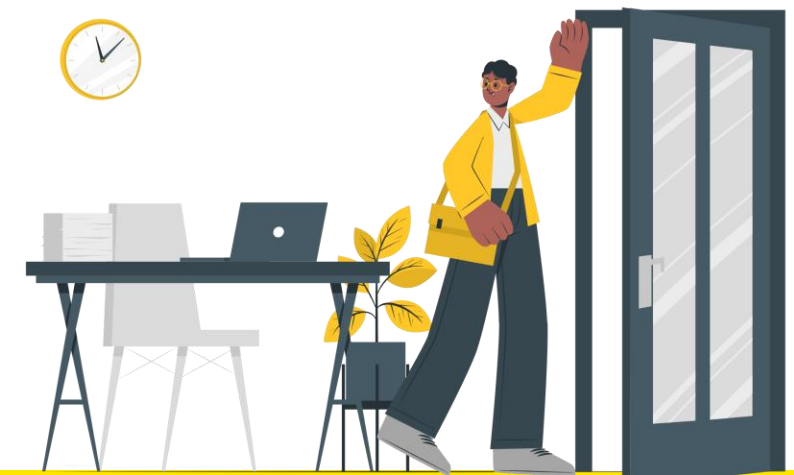
Future Research:

- Fine-tune transformer models like BERT and explore domain-specific variants (e.g., NewsBERT).
- Experiment with multimodal data — images, metadata, and social signals.
- Incorporate hierarchical models to better capture article-level structure.
- Evaluate on noisy, real-world news datasets to test robustness and adaptability.



Conclusion

- News classification is more than just text processing — it's a challenge of semantics, nuance, and scale.
- Balancing accuracy and efficiency is critical for real-world applications.
- LSTM + GloVe achieved the best results in this project, proving that smarter embeddings can outperform larger architectures.
- There's still plenty of room to grow — with fine-tuned transformers, hierarchical models, and real-world data stress testing.
- The goal isn't just better models — it's better understanding.



References

- [1] Jain, R., Kumar, P. and Jaiswal, A., 2021. News Article Classification Using Machine Learning Approaches. *Procedia Computer Science*, 192, pp.2312-2319.
- [2] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), pp.22-36.
- [3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint, arXiv:1907.11692*.
- [4] Ajao, O., Bhowmik, D. and Zargari, S., 2019. Fake News Identification on Twitter with Hybrid CNN and RNN Models. *Proceedings of the 9th International Conference on Social Media and Society*, pp.226-230.
- [5] Kowsari, K., Heidarysafa, M., Brown, D.E., Jafari Meimandi, K. and Barnes, L.E., 2019. Text Classification Algorithms: A Survey. *Information*, 10(4), p.150.

Questions?



