Multi-label News Classification

Vanishing Gradients – COMP9444 – 2025 T1

Neitik Maheshwari (z5636903) Tahir Khawaja (z5363837) Nimish Aggarwal (z5260604)

Sanjay Govindan (z5531936) Shreyas Ananthula (z5360586)

I. Abstract

In today's flood of online news, there is a need to accurately and efficiently classify news articles. Effective classification supports recommendation systems and search engines, helping users better navigate and understand information. This project investigates deep learning architectures for classifying AG News articles into four categories: Business, Science/Technology, Sports, and World. The final proposed model, an LSTM with pre-trained GloVe embeddings, achieves an accuracy of 92.1%, outperforming practical benchmarks such as zero-shot large language models and other model architectures by up to 4%, while maintaining comparable inference times.

Keywords — Classification, AGNews, Deep Learning, Glove, LSTM

II. Introduction

The explosive growth of digital news has made it increasingly challenging to process and categorise vast amounts of information. News articles often span diverse topics, requiring fast, accurate classification for applications like personalised news feeds, recommendation systems, and information retrieval.

Effective categorisation is crucial for businesses, especially news organisations, which must continually identify and organise content to match audience interests and stay competitive. Manual classification, once viable, has become impractical at scale. It demands large editorial teams, is time-consuming, expensive, and ultimately unsustainable. Traditional rule-based and shallow machine learning methods also struggle to handle the nuanced, multi-topic nature of modern news articles.

Recent advances in artificial intelligence, particularly deep learning, have transformed text classification. Neural network architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) such as LSTMs, and Transformer-based models, have demonstrated significant improvements in tasks like sentiment analysis, topic modelling, and document classification. These techniques offer new opportunities to automate news categorisation with greater accuracy and efficiency.

This project investigates deep learning approaches for classifying articles from the AG News dataset, which contains thousands of articles labelled into four categories: *World*, *Sports*, *Business*, and *Science/Technology*. The objective is to develop and evaluate various neural network architectures, progressing from classical methods like Naive Bayes, to modern LSTM models with GloVe embeddings, and state-of-the-art transformer architectures such as BERT to develop a network which balances predictive accuracy with computational efficiency.

By comparing the strengths and limitations of each approach, the project aims to identify neural network models that best balance these trade-offs. Ultimately, the goal is to create intelligent systems capable of scaling automated news classification for real-world applications, shaping how users discover, access, and interact with information in an increasingly digital world.

III. Related Work (Literature Review)

Early natural language processing (NLP) methods were built on shallow neural networks, but their performance plateaued due to a limited ability to capture contextual information. This limitation drove the development of more sophisticated architectures, notably Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), which achieved significant improvements in modelling local and sequential patterns in text.

LSTMs, introduced by (Hochreiter and Schmidhuber 1997), addressed a fundamental limitation of standard recurrent networks: the vanishing gradient problem. By preserving information across long sequences, LSTMs enabled models to capture long-range dependencies, making them particularly effective for tasks where word order and temporal relationships are critical, such as sentiment analysis and machine translation (Sutskever et al., 2014). Nonetheless, LSTMs have a computational inefficiency due to being sequential in nature, limiting parallelisation in training (Vaswani et al., 2017).

Meanwhile, CNNs, originally designed for image processing, were adapted to NLP by treating text as a sequence of embeddings and applying convolutional filters to extract local N-gram features (Kim, 2014). CNNs learned hierarchical representations with fewer parameters compared to recurrent networks. However, their reliance on fixed-size kernels restricted their ability to model long-range dependencies, making them less effective for capturing global context in lengthy documents (Kalchbrenner et al., 2014).

The limitations of LSTMs and CNNs—particularly in computational scalability and handling long-range dependencies—motivated the development of attention mechanisms and transformer-based architectures. The Transformer model (Vaswani et al., 2017) revolutionised NLP by replacing recurrence with self-attention, enabling direct modelling of dependencies between all tokens, regardless of distance. This innovation proved highly effective for tasks requiring full-document comprehension, such as text classification and summarisation.

A major advancement in this space was BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), which introduced deep bidirectional pretraining and achieved state-of-the-art results across a range of benchmarks. However, despite their performance gains, transformer models posed challenges for real-world deployment due to large parameter sizes, extensive pretraining requirements, and high memory consumption (Strubell et al., 2019).

Consequently, recent research has focused on optimising transformers for efficiency through techniques such as knowledge distillation (Sanh et al., 2019) and sparse attention mechanisms (Child et al., 2019), aiming to make these powerful models more practical for scalable and resource-constrained applications.

IV. Methods

A total of 13 models were evaluated in this study: 4 benchmark models, 5 initial deep learning models, and 2 models developed under each of the two hypotheses. The rationale for each set of models was guided by insights gained from the performance of prior approaches and related works. Due to the number of models, only key results are presented in this report.

Benchmarking

To provide a practical benchmark the performance of two out of the box classification techniques were explored, Naive Bayes and zero- and one-shot classification with existing LLMs. The first one was a simple naive bayes with pre-processing that tracks the frequency of words in each category and does an effective regression on that. The pre-processing of this was used to take out stop words and common words, as can be seen in Figure 3 The various n-grams made it clear that there were sets of words such as "*in*", *"and"* and *"the"* which are not needed for the purposes of our processing.

Our secondary benchmark, zero- and one-shot classification used Google and DeepSeek models. No fine tuning has been conducted on these models, but we use their inherent knowledge to classify the articles. For Google we experimented with older Gemma 3 and newer Gemini 2.5 models, batching for both. Regarding DeepSeek we tried R1 without batching, but as our resource were limited, we had to resort to single line batching.

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%

Table 1: Results from practical benchmarks. Naïve Bayes produces the best against the zero shot and one short benchmarks from off the shelf LLMs.

Deep learning methods

We built deeper learning models in a structured and scalable way, progressing from simple baselines to more expressive architectures. Using the linear models as a launching off point as a simple trainable baseline, which aligned with Naïve Bayes (Table 1) but added the flexibility of learnable weights. This model however struggled, achieving just 34.8% accuracy due to limited representational capacity. By adding an embedding layer to the linear model, the model was able to learn richer text representations, boosting performance to 90.4% accuracy.

CNNs (Figure 1) are traditionally used for images but for short text contexts there is a method to use it effectively (Ajao, Bhowmik and Zargari, 2019). The method replaces 2D pixel grid with a sequence of word embeddings (dense vectors); a sentence becomes a matrix of shape (sequence length × embedding dimension). Following this augmentation the CNN continues, but with 1D convolutions. CNN used in this fashion captures local patterns (like phrases and key n-grams) regardless of their position. This provides a similar historical context architecture to that of Long Short-Term Memory modules whilst providing a fast, parallelizable, and suitable architecture to process large textual datasets like AG News.

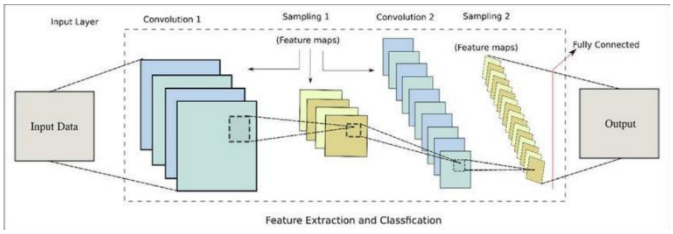

Figure 1: Sample CNN model architecture

Next, we introduced: Long Short-Term Memory networks (LSTMs) were specifically designed to handle sequential data. The LSTM architecture (Figure 2) provides a context module which stores important information across longer distances in the sequence, making them ideal for tasks like text classification where word order matters (e.g., "not good" vs "good"). Unlike CNNs (which focus on local patterns), LSTMs can model global sequence relationships. This might be important to analyse, as it is a text of news articles, and intuitively the order the text is written is important for understanding.

Similarly for Bi-LSTM (bi-directional LSTM) this model goes both forward and backward, allowing for a deeper understanding of the text. This should ideally be more accurate the LSTM as it has more information.

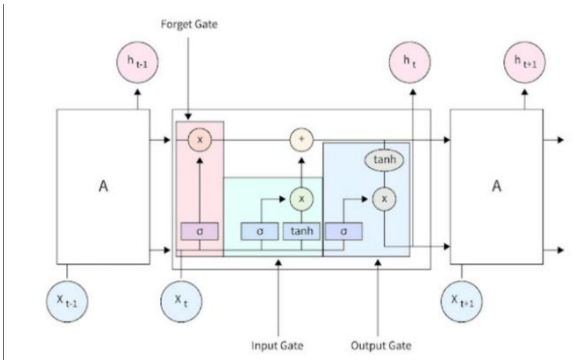

Figure 2: Example LSTM module architecture.

Overall, this approach reflects a deliberate and literature-backed progression, with models like LSTM and Bi-LSTM achieving over 90% accuracy.

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	4m	3,771,963	90.3%

Table 2: Results from deep learning architectures, highlighting Bi-LSTM as the leading model with the best accuracy.

Hypothesis 1: Attention Mechanisms

After looking at the deep learning model and its results, as can be seen in Table 2 all of them approached approximately 90% and then stopped getting more accurate. Hence the 1st hypothesis formed, that LSTM may not provide sufficient historical and contextual understanding for classifying news articles. To explore original deep learning architectures, we investigated mechanisms to effectively extract meaning from the sequence of text inspired by (Vaswani et al. 2017). To explore this hypothesis, we use two models:

Linear simple with attention, allowing the model to selectively emphasise important tokens and potentially boost overall accuracy, without immediately resorting to heavier transformer models. We hypothesised that this model would strike a balance between efficiency and performance, offering a lightweight alternative to heavier models like full LSTMs or Transformers. To our surprise this model underperformed the Linear model as demonstrated in Table 3.

BERT, developed by (Devlin et al., 2018), introduces self-attention mechanisms that allow the model to process and relate all parts of a sentence simultaneously, capturing rich, bidirectional context. Instead of learning from scratch, we fine-tune a pre-trained "bert-base-uncased" model with a classification head tailored for AG News classification. By using BERT, we leverage its pre-learned deep understanding of language (from massive corpora like Wikipedia) to potentially break through the accuracy ceiling we observed with earlier models, and more effectively capture the directional data and depth that may be missing in Bi-LSTM , linear and CNN

Model	Parameters	Accuracy
Linear	3,010,504	**90.4%**
Self-attention simple	7,779,716	89.1%
Bert	109,485,316	90.1%

Table 3: Results from hypothesis 1 exploring attention integration. Demonstrates how attention does not contribute to an improvement in model accuracy.

Hypothesis 2: Embeddings

After evaluating BERT and noting that it did not significantly improve classification accuracy as much as was hoped, the team considered a second hypothesis. There is a strong argument that the core limitation lies not in sequence modelling, but in the semantic richness of the input embeddings. If models like LSTM and Bi-LSTM achieve similar results it suggests that sequential information is already being effectively extracted. Therefore, the issue may not be "how" we are modelling the sequence, but rather "what" information the embeddings contain.

To explore this idea, we test whether improving the semantic quality of embeddings could boost model performance. Specifically, we introduce GloVe embeddings of 100d and 200d(dimensions) into our best-performing LSTM and linear models. GloVe embeddings are trained on large corpora to capture broader word relationships and meanings, and they offer a simple yet effective way to evaluate whether enhanced semantic information leads to better classification outcomes.

Model	Parameters	Accuracy
LSTM	3,433,412	89.9%
LSTM with Glove	3,170,576	**92.1%**
LSTM with Glove 200	6,274,076	92.1%
Linear	3,010,504	91.6%
Linear with Glove	3,052,300	91.8%
Linear with Glove 200	6,125,104	91.9%

Table 4: Hypothesis 2 results demonstrating LSTM with Glove as the best performing model with one of the lowest parameters counts.

As can be seen from Table 4 LSTM with Glove Gives the highest accuracy across all the models with 92.1%, across both 100d and 200d having similar accuracies, the 100d model was chosen as it has fewer parameters.

V. EXPERIMENTS

The source of the dataset we used was AG news (**https://huggingface.co/datasets/wangrongsheng/ag_new**

s) it contains 120,000 training samples and 7,600 test samples. Each sample contains the news article title, description and the associated category/topic. There are four news category classes. Sci/Tech, Business, Sports and World. The Categories are all split evenly and neatly, allowing for us not having to wight the sample one way or the other.

To explore the dataset, we used a variety of methods, including n-grams, word clouds, and length of texts, per category. By analysing the n-grams and word clouds, we observed a high presence of stop words in the dataset. Additionally, within the n-grams, we identified a substantial number of coupled words, such as “Reuters Reuters”, “AP AP”, and place names such as “New York” and “San Francisco”. These findings were used for the pre-processing of the Naive bayes.

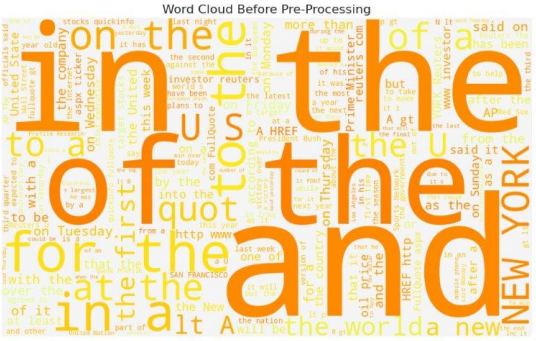

Figure 3: Word analysis pre cleaning is dominated by stop words

The evaluation strategy focused on both accuracy and computational efficiency. Accuracy was used as the primary performance metric, while training time and model size were considered to assess efficiency. Across all models, a maximum of 15 training epochs were used as a standard benchmark during training.

For the final proposed model — the LSTM with pre-trained GloVe embeddings — key hyperparameters included a learning rate of 0.001, a batch size of 64, and the use of early stopping based on validation accuracy to prevent overfitting. Other hyperparameters were optimised using grid search to further improve model performance while maintaining computational efficiency.

VI. RESULTS

Examining the results Table 5 from the original Deep Learning benchmarks and Hypothesis 1 and 2 We find that the LSTM model enhanced with GloVe embeddings provides the best overall performance, while maintaining comparable efficiency to other approaches.

Category	Model	Training Time	Accuracy & Parameters
Benchmark	Gemini 2.5	N/A	88.6% (~2.5T)
Deep Learning	Bi-LSTM	4m	90.3% (3.8M)
Hypothesis 1: Attention	BERT	7m	90.0% (109M)
Hypothesis 2: GloVe	LSTM + GloVe	**3m**	**92.1% (3.1M)**

Table 5: Final results comparing the accuracy, parameters and training times of different techniques across our experiments. Hypothesis 2, LSTM with GloVe produces the best result.

Our results echo those of Pennington et al. (2014) demonstrating that GloVe embeddings capture both co-occurrence statistics and semantic structure, enabling more effective learning across NLP tasks.

The proposed solution applies an LSTM model with pre-trained GloVe embeddings for text classification. Unlike transformer-based models such as BERT Hamidzadeh (2021) and XLNet Yang et al. (2019), which are the current state-of-the-art models in achieving the highest accuracy on the AG News dataset, this approach prioritises computational efficiency. While XLNet achieves higher accuracy (above 95% on AG News) through complex attention mechanisms and large-scale pretraining Yang et al. (2019), it requires over 110 million parameters and extensive GPU resources for fine-tuning. In comparison, the proposed LSTM model contains approximately ~3 million parameters, trains in under half an hour on a single GPU and maintains strong performance with accuracy above 92%.

Although the LSTM model does not reach XLNet's absolute accuracy, it offers a significant reduction in model size, training time, and resource requirements, making it better suited for deployment in real-time or resource-constrained environments.

VII. CONCLUSION

The LSTM with GloVe model achieved an accuracy of 92.1%, showcasing that effective architecture can rival richer models with tractable computational costs. The notable findings indicated GloVe embeddings improved performance on linear and LSTM models with minor improvements from large embedding dimensions (100d-200d). The methodology also exhibited a faster convergence than Simple LSTM and Bi-LSTM variants with reproducible standards for news classification tasks. Limitations present in the model are its reliance on titles and descriptions rather than full article text that could potentially miss contextual signals. The clean, balanced AG News dataset may not fully represent noisy real-world data, and we used off the-shelf transformers with modest fine-tuning due to resource constraints. These factors could impact the model performance in production environments. Future research would entail fine-tuning domain-specific transformers (i.e., NewsBERT) and incorporating multimodal data like images and metadata. Examining hierarchical models may more effectively capture article-level structure, and evaluation on noisy, real-world datasets would evaluate robustness. These improvements would render the model more applicable in practice to a variety of news classification applications

References

[1] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), pp.1735–1780. doi:https://doi.org/10.1162/neco.1997.9.8.1735.

[2] Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, pp.3104–3112

[3] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. doi:https://doi.org/10.48550/arxiv.1408.5882.

[4] Kalchbrenner, N., Grefenstette, E. and Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.1404.2188.

[5] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org. Available at: https://arxiv.org/abs/1810.04805.

[6] Strubell, E, Ganesh, A & McCallum, A 2019, 'Energy and Policy Considerations for Deep Learning in NLP', *arXiv (Cornell University)*, Cornell University.

[7] Sanh, V, Debut, L, Chaumond, J & Wolf, T 2020, *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, arXiv.org.

[8] Child, R., Gray, S., Radford, A. and Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. doi:https://doi.org/10.48550/arxiv.1904.10509.

[9] [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.

[10] Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28 (NIPS 2015).

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, pp.5998–600

[12] Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532–1543.

[13] AG News BERT Classification, Mansoor Hamidzadeh (Hugging Face), 2021.

[14] [XLNet: Generalised Autoregressive Pretraining for Language Understanding, Yang et al., NeurIPS, 2019.

[15]