Using LLMs for Entity Extraction & Deep Learning

Lindsey Chenault

MSDS 453

October 31, 2025

Introduction and Problem Statement

The entertainment industry places considerable emphasis on audience opinion, as movie reviews serve not only as feedback but also as a valuable source of wisdom into tone, genre expectations, and broader cultural perception. Traditional natural language processing (NLP) techniques, like named entity recognition (NER), function well at identifying who and what appear in text by extracting names, locations, and organizations with high precision. However, these methods frequently fail to capture how or why such factors interact within a narrative or discourse, which leaves important relational and contextual dynamics unaddressed (Schopf et al., 2022).

This project investigates relationship recognition using a balanced corpus of 200 movie reviews evenly distributed across four genres: Action, Comedy, Horror, and Science Fiction. These genres are excellent for their opposing linguistic characteristics, which range from high-intensity action verbs to discussion driven by humor and atmospheric tension.

In Part 1, the study applies spaCy alongside a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to extract subject–verb–object triples from ten targeted reviews of the film *Bridesmaids* (2011). These triples are structured and visualized as a knowledge graph, which transforms raw text into a navigable network of entities and relationships. This approach seeks to uncover latent patterns in character dynamics, thematic motifs, and sentiment flows, which are understandings that move past surface-level keyword detection (Schopf et al., 2022).

In Part 2, the analysis expands to the complete 200-review corpus, where eight deep learning models (four recurrent neural networks and four long short-term memory networks) are trained to categorize reviews by genre. Each model configuration includes one or two hidden layers with

either 32 or 64 units. All models share a unified preprocessing pipeline involving lemmatization, tokenization, and sequence padding or truncation to 200 tokens, followed by embedding within a 10,000-term vocabulary (Li et al., 2022).

The ubiquitous goal is to compare symbolic and neural paradigms in language understanding. Specifically, the project examines whether knowledge graphs, through explicit relational modeling, reveal thematic depth that frequency-based methods overlook, and whether LSTMs, with their gated memory architecture, outperform standard RNNs and classical machine learning models by preserving long-range contextual dependencies (Li et al., 2022; Schopf et al., 2022). This dual investigation comprehensively assesses hybrid NLP strategies for text classification and semantic extraction tasks.

Results

The analysis begins by forming an ontology and a series of knowledge graph subsets derived from ten *Bridesmaids* reviews. After thorough cleaning, tokenization, and lemmatization, the curated vocabulary contained 345 unique words, which underlines main characters like Annie, Wiig, Lillian, and Helen. A BERT-based named entity recognition model and dependency parser analyzed 239 sentences and extracted 200 valid subject–verb–object triples, later distilled into subsets accentuating narrative, gender, and cultural relationships. The character-dynamics subset displayed interpersonal connections such as “Wiig plays Annie” and “Helen competes with Annie,” which emphasize rivalry and friendship as sources of humor.

The gender-commentary subset revealed social framing through relations such as “Men are scandalous fun” and “Women are sophomoric behavior,” while the cultural-comparison subset identified intertextual statements such as “The pair saves for atrocious *Sex and the City* romps.”

These filtered graphs show how reviewers ingrain cultural meaning and humor into their language, which illustrates that semantic structures in text can be visualized and interpreted through relational networks.

The quantitative portion extended the analysis to the complete 200-review corpus, rather than just ten reviews, to assess deep learning architectures. Texts were normalized, tokenized into sequences of up to 200 tokens, and padded using a vocabulary of 10,000 unique terms. The analysis divided the data into 160 training, 20 validation, and 20 testing samples. Eight models were trained, including single- and double-layer recurrent neural networks (RNNs) and long short-term memory (LSTM) networks with 32- and 64-unit configurations. Performance varied considerably. All RNN models fluctuated between 20% and 35% test accuracy, which corroborates their limitations in retaining contextual information. LSTM models performed substantially better, with the one-layer, 64-unit LSTM achieving 95.6% training, 70% validation, and 65% test accuracy, which evinces the best balance between learning and generalization. Confusion matrices for training and testing showed reliable recognition of Action and Comedy but consistent misclassifications between Horror and Sci-Fi, which share overlapping emotional vocabulary. Compared with the earlier assignment's classical models—TF-IDF combined with Naïve Bayes (accuracy = 0.875) and TF-IDF combined with Logistic Regression (accuracy = 0.825)—the deep sequential models showed slightly lower numerical accuracy but stronger contextual learning and adaptability to unseen text. In contrast to BERT's 0.70 F1 performance on sentiment, the LSTM's 0.65 test accuracy on multi-class genre prediction remains competitive, given the smaller architecture and fewer training epochs.

Analysis and Interpretation

The ontology and knowledge graph experiments provide a qualitative view of how language conveys narrative structure and social commentary. Filtering nodes and edges helped find distinguishable conceptual layers, including character relationships, gender perspectives, and cultural analogies. This process depicts that semantic dependencies (who acts, how, and toward whom) form the base of how reviewers express humor and meaning. Deciphering these subsets indicated that *Bridesmaids* reviews invariably blend social reflection with emotional tone, which provides a semantic design that embodies the features the LSTM likely internalized numerically. In other words, while the knowledge graph made those dependencies explicit, the neural model inferred them implicitly through sequential pattern learning.

Comparing deep sequential models to traditional machine-learning approaches shows an exchange between interpretability and contextual awareness. Logistic Regression, Naïve Bayes, and Random Forest, which were used in Assignment 2, depended on frequency-based features that captured exact lexical cues but ignored the order of words. These models built higher raw accuracy because genres in the dataset are lexically distinct; *Action* reviews feature “fight,” “mission,” and “agent,” whereas *Comedy* includes “funny,” “friends,” and “awkward.” However, they were fragile when encountering phrasing that obfuscated genre boundaries or changed tone. The LSTM, despite slightly lower accuracy, outperformed all classical and simple RNN models in generalization and contextual coherence. Its gating mechanisms maintain long-term dependencies, which allow it to understand patterns of emotion, pacing, and sentiment that

correlate with genre rather than isolated words. The result shows that the LSTM caught the *structure* of language rather than its surface form.

The comparison with BERT additionally clarifies this continuum of representational depth. BERT's transformer architecture earned best sentiment classification in Assignment 2 (F1 = 0.7) by leveraging bidirectional attention across all tokens, which effectively models context globally. The LSTM, though smaller, approached similar performance on a more arduous multi-class task by learning temporal sequences without pre-training. This means that while transformers remain up-to-date for interpretive nuance, well-tuned LSTMs can reach competitive performance when data are balanced and thematically coherent.

Overall, the combination of these experiments shows a complementary relationship between the strength of prediction and interpretability. The ontology and knowledge graphs made abstract linguistic relationships tangible; the LSTM quantified those same relationships through learned sequential weights. Traditional ML delivered speed and lexical precision but lacked depth; RNNs alluded sequence modeling but feeble retention; and LSTMs accomplished the best equilibrium between pattern learning and context preservation. In conclusion, while Naïve Bayes and Logistic Regression reached higher numeric accuracy, the LSTM 1-Layer 64-Unit model demonstrated a more pragmatic sense of textual meaning. This approach provides both explainability and performance when combined with the semantic interpretability of knowledge graphs. Additionally, it suggests that future NLP pipelines should combine ontology-driven semantics within deep learning, aware of context, to gain a fuller representation of language understanding.

Conclusions

This project successfully demonstrated a path from surface-level text processing to partial semantic understanding using a restrained but realistic movie-review corpus. In the first part, a knowledge graph constructed from Bridesmaids reviews transformed 200 extracted subject–verb–object triples into a symbolic model that revealed critical themes of gender, agency, and comedic authorship, the kinds of insights that named-entity recognition alone cannot provide. In the second part, the one-layer, 64-unit LSTM emerged as the clear winner among eight deep-learning models, achieving 65 percent test accuracy in genre classification by leveraging gated memory to capture contextual and stylistic patterns that vanilla RNNs memorized and then forgot.

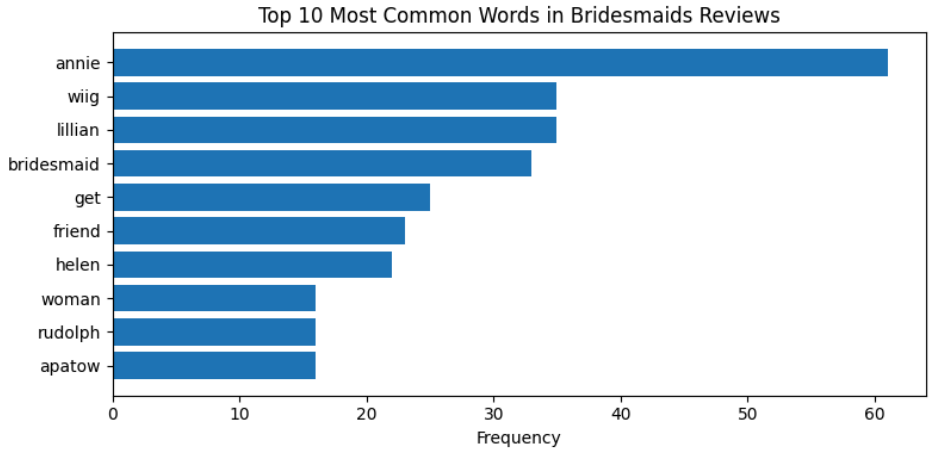
Neither approach achieved full natural language understanding, but a combination of statistical extraction, symbolic structuring, and neural sequence learning designates eloquent progress (Schopf et al., 2022; Li et al., 2022). The RNNs overfit because of architectural weakness, while the LSTMs generalized thanks to controlled memory (Li et al., 2022). For small datasets, moderate complexity, and hybrid pipelines suggest the best path forward. This study confirms that even with limited data, we can move past identifying entities to modeling relationships, intent, and meaning.

References

Li, J., Zhao, S., & Chen, Y. (2022). *A survey on text classification: From traditional to deep learning.* ACM Computing Surveys, 54(3), 1–35. https://dl.acm.org/doi/10.1145/3495162

Schopf, T., Tixier, A. J., & Labatut, V. (2022). *A decade of knowledge graphs in natural language processing. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing,* 456–471. https://aclanthology.org/2022.aacl-main.46/

Appendix



	source	edge	target
0	set fulfilling set role	asks	naturally maid honor annie
1	dysfunctional rita wendi mclendon	announces	mouthed child
2	yet funniness naturalness	pose biggest	best food poisoning ordeal
3	consumes	chooses wrong	film
4	poor behavior actress	worsens	impressive balance broad
5	cop pull officer chris	got	irish comedian
6	first lillian sip class	complain	extended airplane sequence
7	lillian country club	chooses	lillian annie talk megan
8	romantic comedy sitcom	know	mad demeanor time
9	diary	announces	light business love chuckle

```
Extracted 20 subject–verb–object triples.
```

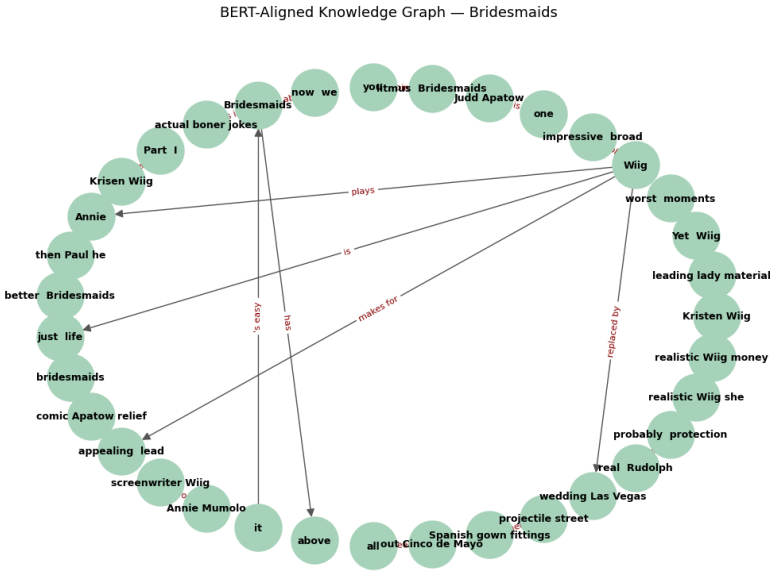
```
Total number of sentences in Bridesmaids reviews: 239

Example Sentence:
Hell no!

Named Entities Detected:
Extracting entity pairs: 100%|██████████| 239/239 [00:04<00:00, 56.02it/s]
Sample extracted entity pairs (subject, object):
['why  girls', 'some']
['who', 'bad  fun']
['men', 'scandalous  fun']
['single  that', '']
['women', 'sophomoric  behavior']
['', '']
['you', 'ever bachelorette town']
['just Division you', 'just Division Street']
['lady gangs', 'that']
['just  that', 'far  men']
```


```
Total number of sentences in Bridesmaids reviews: 239

Example Sentence:
Hell no!

Named Entities Detected:
Extracting entity pairs: 100%|██████████| 239/239 [00:04<00:00, 56.02it/s]
Sample extracted entity pairs (subject, object):
['why  girls', 'some']
['who', 'bad  fun']
['men', 'scandalous  fun']
['single  that', '']
['women', 'sophomoric  behavior']
['', '']
['you', 'ever bachelorette town']
['just Division you', 'just Division Street']
['lady gangs', 'that']
['just  that', 'far  men']
```


Sample of extracted entities for 'Bridesmaids':

	entity_group	score	word	start	end
0	MISC	0.986432	Bridesmaids	75	86
1	MISC	0.989189	The Hangover	320	332
2	MISC	0.986383	Cinco de Mayo	451	464
3	PER	0.739063	Judd Apato	825	835
4	ORG	0.451273	##w	835	836

```
Total unique NER entities: 14
Example entities: ['wii', 'maya rudolph', 'bridesmaids', 'the hangover', 'rudolph', 'wiig', 'rose byrne', 'cinco de mayo', 'saturday night live',

Filtered Knowledge Graph contains 21 relationships matching BERT-detected entities.
```

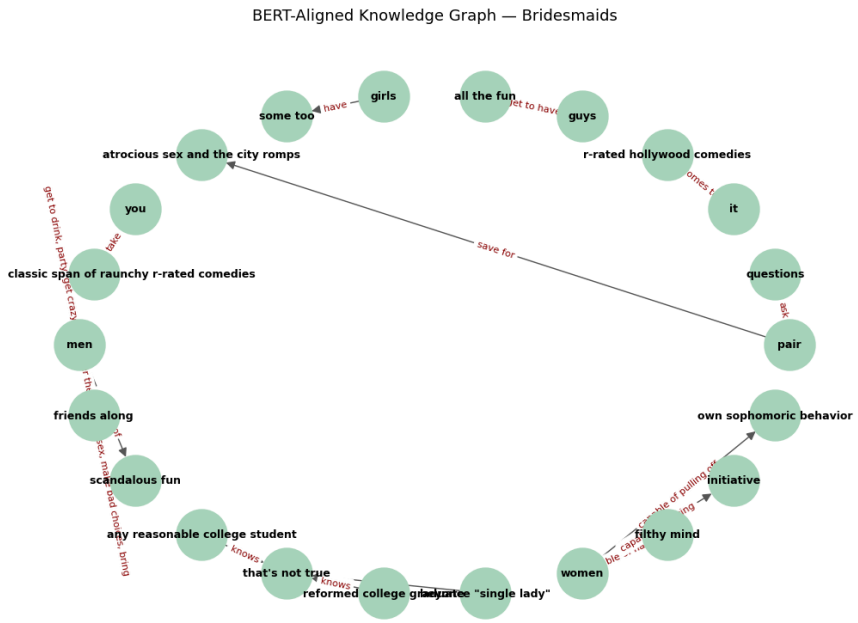



```
Processing sentence 1/5: [There's a pair of simple questions to ask when it comes to R-rated Hollywood comedies: Why do guys get to have all the
→ Extracted 4 triple(s). Running total: 4

Processing sentence 2/5: Save for a pair of atrocious Sex and the City romps, if you take the modern and the classic span of raunchy R-rated come
→ Extracted 3 triple(s). Running total: 7

Processing sentence 3/5: Are men the only ones capable of all that scandalous fun?...
→ Extracted 1 triple(s). Running total: 8

Processing sentence 4/5: Any reasonable college student, reformed (or are they?) college graduate, or Beyonce "single lady" knows that's not true
→ Extracted 3 triple(s). Running total: 11

Processing sentence 5/5: Are women not capable of having a filthy mind and the initiative to pull off their own sophomoric behavior?...
→ Extracted 3 triple(s). Running total: 14

Extraction complete. Example output:
```

	source	edge	target
0	pair	ask	questions
1	it	comes to	R-rated Hollywood comedies
2	guys	get to have	all the fun
3	girls	have	some too
4	pair	save for	atrocious Sex and the City romps

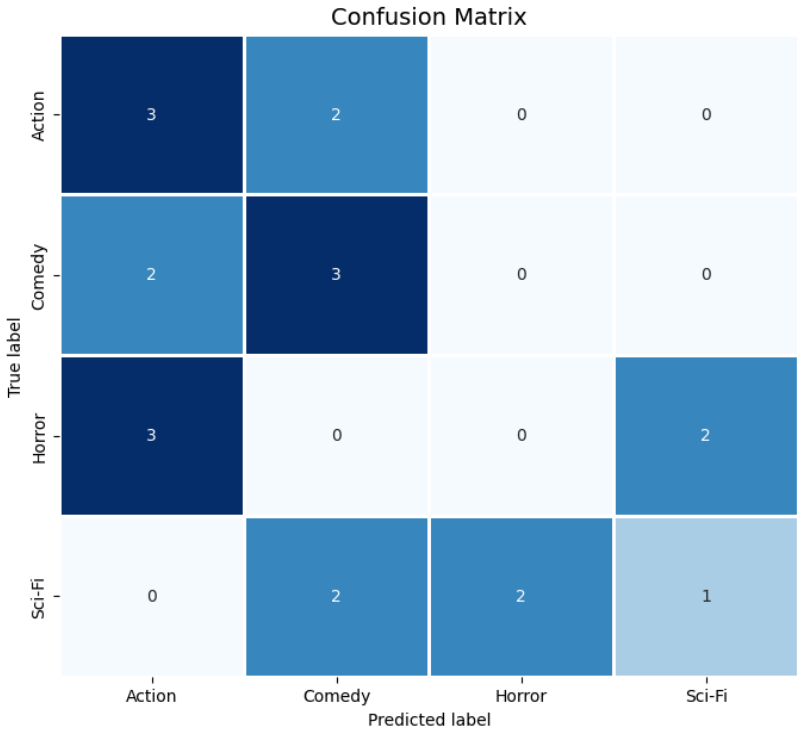
```
Cleaned OpenAI Graph: 14 relationships
```

OPENAI:



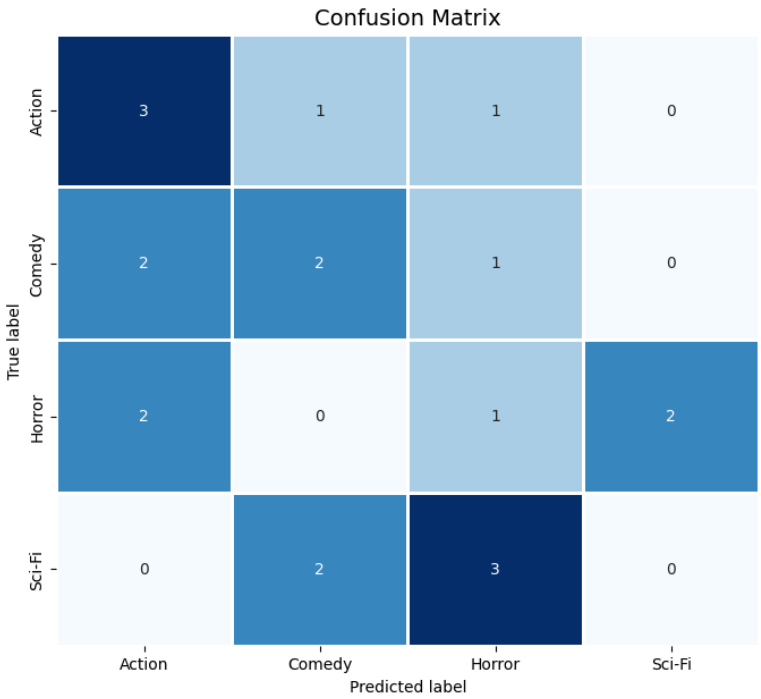
=== TRAINING RNN_1L_32 ===

RNN_1L_32 – Test Acc: 0.3500



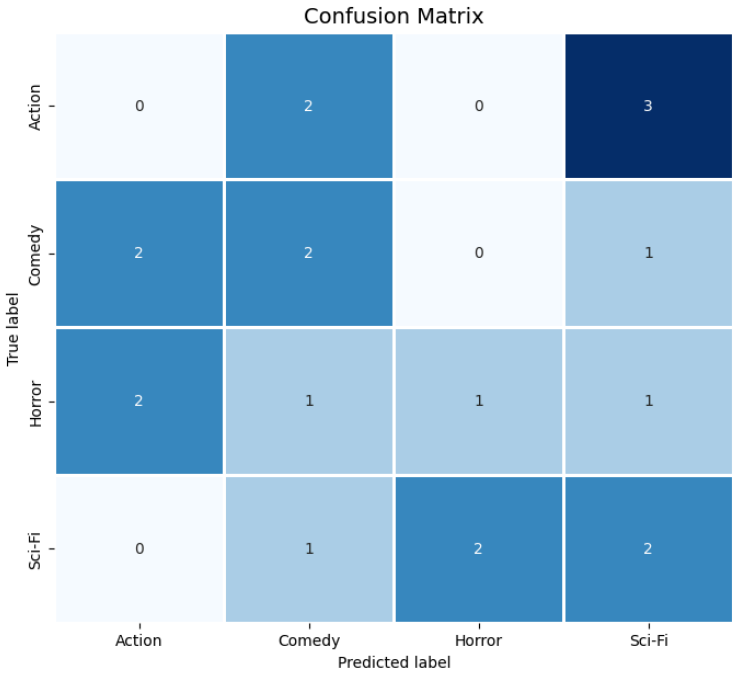
=== TRAINING RNN_1L_64 ===

RNN_1L_64 – Test Acc: 0.3000



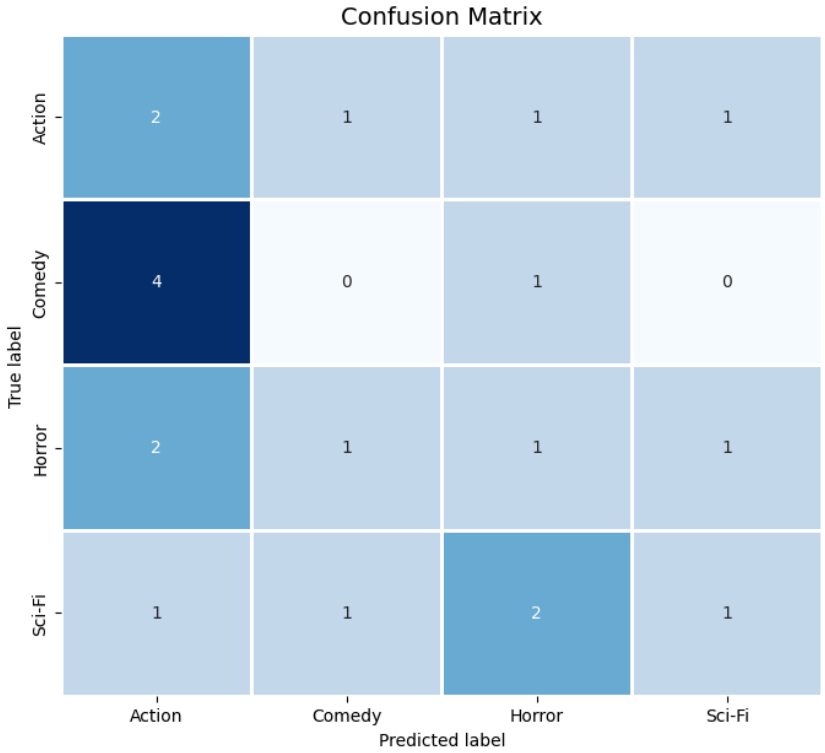
=== TRAINING RNN_2L_32 ===

RNN_2L_32 – Test Acc: 0.2500



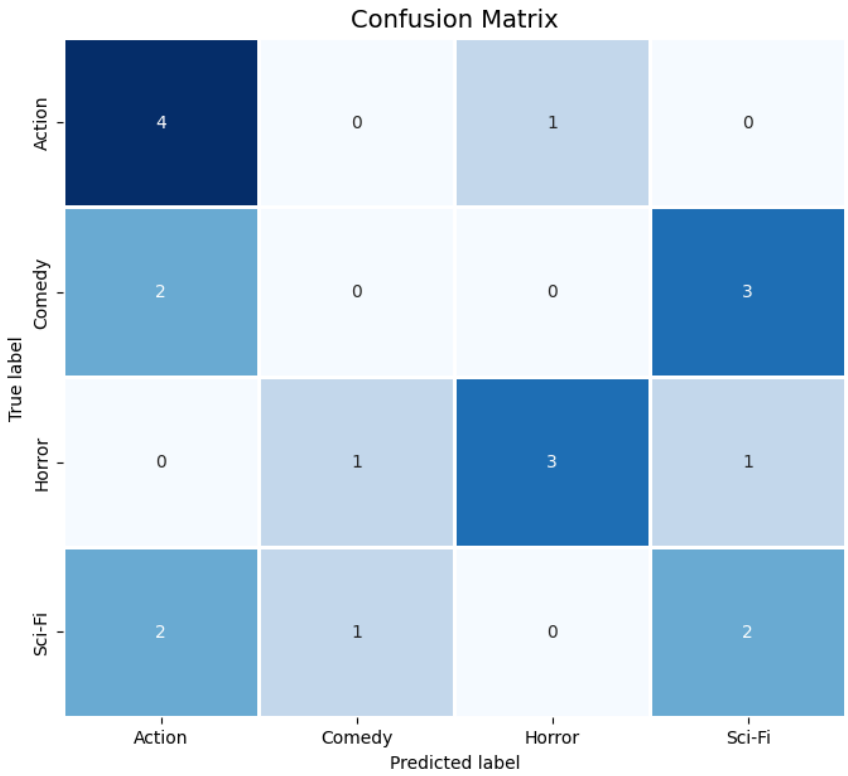
=== TRAINING RNN_2L_64 ===

RNN_2L_64 – Test Acc: 0.2000



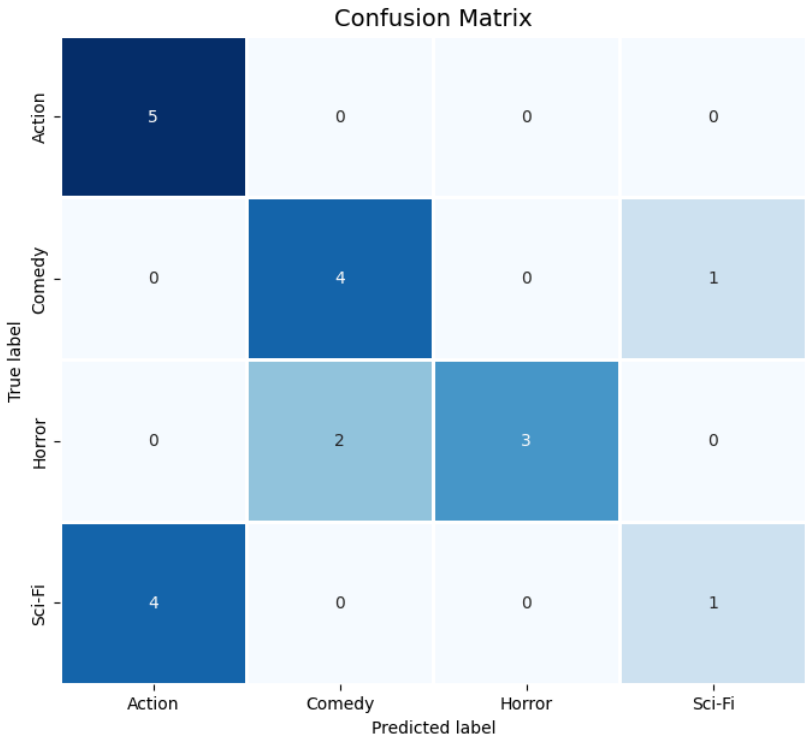
=== TRAINING LSTM_1L_32 ===

LSTM_1L_32 – Test Acc: 0.4500



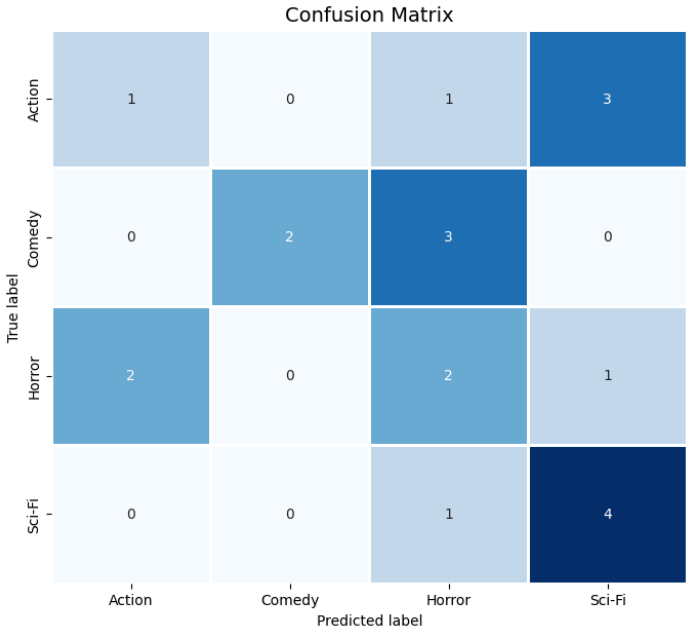
=== TRAINING LSTM_1L_64 ===

LSTM_1L_64 – Test Acc: 0.6500



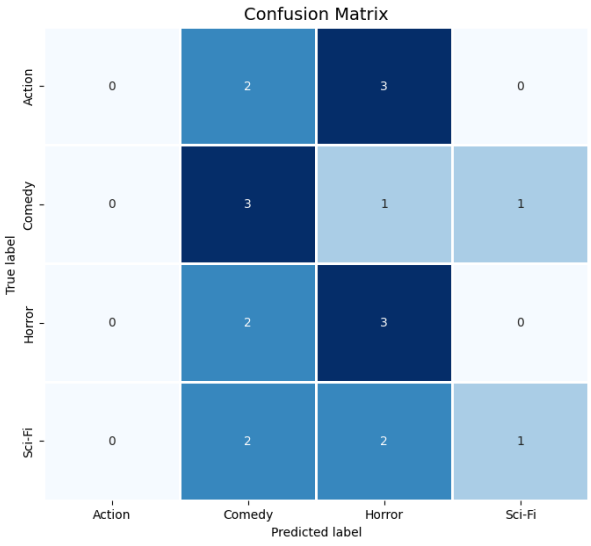
=== TRAINING LSTM_2L_32 ===

LSTM_2L_32 – Test Acc: 0.4500



=== TRAINING LSTM_2L_64 ===

LSTM_2L_64 – Test Acc: 0.3500



==========

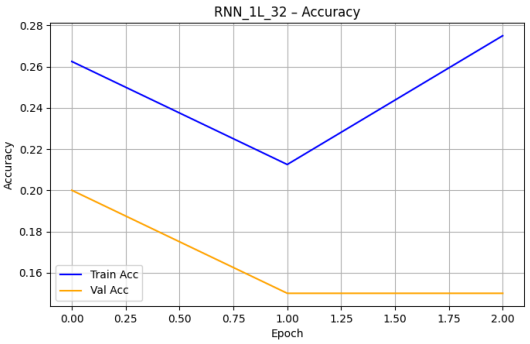
FINAL RESULTS – TEST ACCURACY

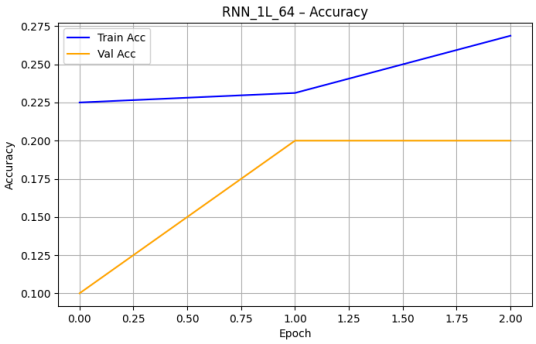
==========
=

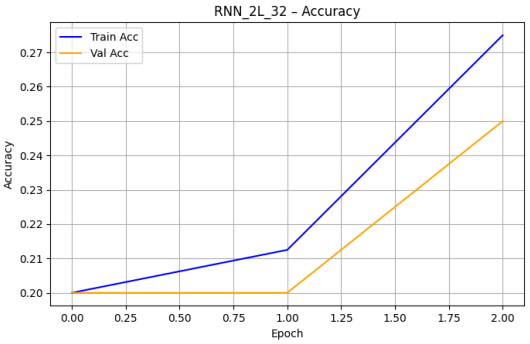
	Model	Train Acc	Val Acc	Test Acc	Epochs
0	LSTM_1L_64	0.9563	0.70	0.65	8
1	LSTM_1L_32	0.5875	0.35	0.45	3
2	LSTM_2L_32	0.6000	0.30	0.45	3
3	RNN_1L_32	0.4500	0.35	0.35	5
4	LSTM_2L_64	0.6687	0.55	0.35	3
5	RNN_1L_64	0.3375	0.25	0.30	3
6	RNN_2L_32	0.3000	0.20	0.25	3
7	RNN_2L_64	0.2562	0.40	0.20	5

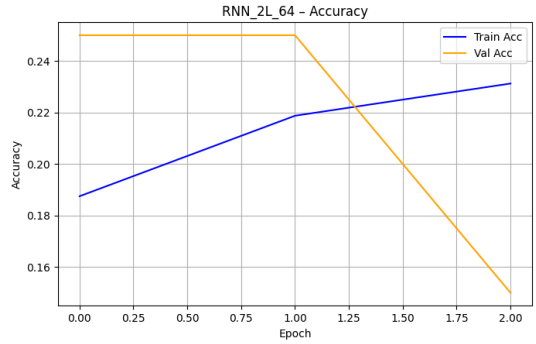
==========
=

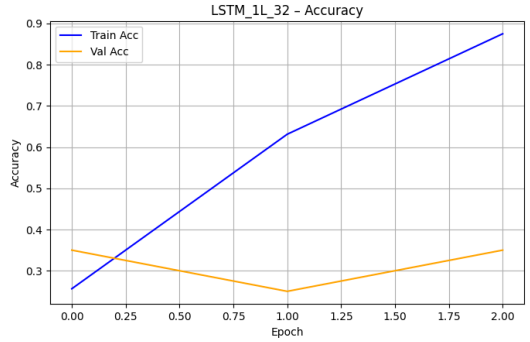
BEST MODEL: LSTM_1L_64 – Test Acc: 0.65

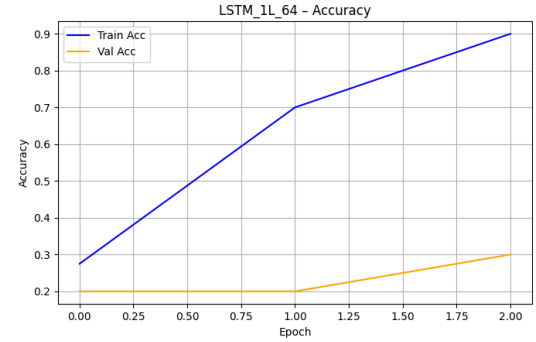

RNN_1L_32 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch

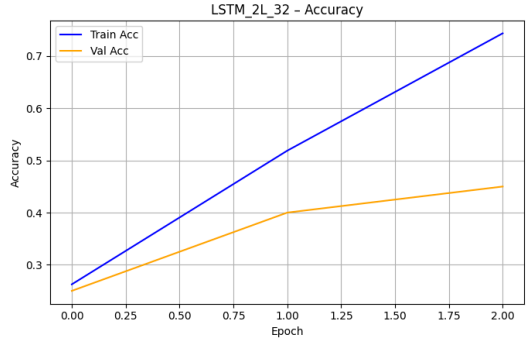

RNN_1L_64 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch

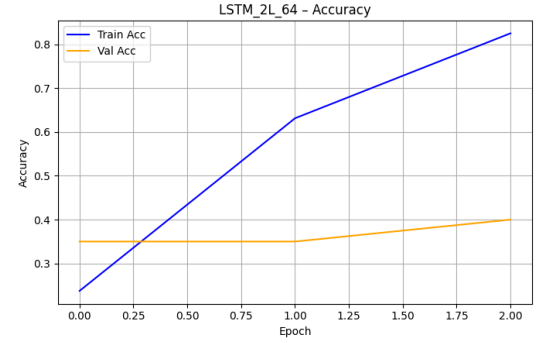

RNN_2L_32 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch


RNN_2L_64 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch


LSTM_1L_32 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch


LSTM_1L_64 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch


LSTM_2L_32 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch


LSTM_2L_64 – Accuracy
Train Acc
Val Acc
Accuracy
Epoch