Using LLMs for Entity Extraction

& Deep Learning Experiments Report

Yunzhen WU

Date: May 18, 2025

1. Introduction and Problem Statement

This report focuses on advanced Natural Language Processing (NLP) techniques for entity and relation extraction, and genre classification using deep learning. The project is divided into two primary components: (1) entity and relation extraction from text using spaCy, BERT, and large language models (LLMs), and (2) text classification using Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models.

The motivation for this research stems from the growing need to structure unstructured text data for downstream tasks like question answering, recommendation systems, and information retrieval. This report explores both symbolic knowledge representation (via entity-relation extraction and KG visualization) and neural architectures (for classification), aiming to combine interpretable and performant NLP pipelines.

2. Data and Preprocessing

For Part 1, a set of ten manually selected movie *Oblivion* review documents were used to build knowledge graphs. Preprocessing included tokenization, lemmatization, named entity recognition, and syntactic parsing. spaCy was used to extract subject–verb–object triplets. For the enhanced pipeline, Hugging Face's bert-base-cased model and LLM were respectively combined with spaCy for improved NER tagging.

For Part 2, the full class corpus of movie reviews was used for genre classification. Preprocessing included removing stopwords, vectorizing using a text encoder, and organizing the data into train/validation/test splits. Padding and tokenization were applied before passing the data to deep learning models.

3. Research Design and Modeling Methods

3.1 Knowledge Graph Experiments

Three approaches were used to extract entities and relationships:

- spaCy-only: Rule-based extraction using dependency parsing and relation matching.
- BERT-based NER: Hugging Face pipeline with a pre-trained model (dbmdz/bert-large-cased-finetuned-conll03-english) for named entity recognition.
- LLM + spaCy hybrid: Sentence-level tuple extraction using GPT-3.5-turbo via OpenAI API, followed by DataFrame construction.

Each method resulted in a knowledge graph (kg_df or global_kg_df2), visualized with NetworkX. Additionally, subgraphs filtered by keywords (e.g., "Tom Cruise", "Oblivion", "director") were extracted to highlight contextual edges. Equivalent Classes (ECs) including "Tom Cruise - Jack Harper", "director - filmmaker", "lead actor - tom cruise" and "acts_in - stars_as/leads/is_actor" were manually normalized to reduce noise, improve conceptual integrity and consolidate KG nodes.

3.2 Deep Learning Classification Experiments

Text classification was conducted using TensorFlow/Keras with the following architectures:

- RNN Models with (i) 1 hidden layer with 32 Neurons, (ii) 1 hidden layer with 64 Neurons, (iii) 2 hidden layers with 32 Neurons each, (iv) 2 hidden layers with 64 Neurons each.
- LSTM Models with (i) 1 hidden layer with 32 Neurons, (ii) 1 hidden layer with 64 Neurons, (iii) 2 hidden layers with 32 Neurons each, (iv) 2 hidden layers with 64 Neurons each.

All models used text vectorization, a Bidirectional wrapper, embedding layers, and ReLU-activated dense layers with a final Softmax output. Early stopping was applied based on validation accuracy. Evaluation included accuracy/loss plots and confusion matrices.

4. Results

4.1 Knowledge Graphs

Entity and relation extraction was conducted using both spaCy-only and spaCy+LLM (via HuggingFace Transformers) pipelines. The spaCy-only model successfully extracted syntactic triples but often produced fragmented relations. Figure 1–4 present subgraphs extracted using spaCy for the keywords “Tom Cruise”, “Oblivion”, and “director”, as well as the full knowledge graph. These visualizations reveal sparse and often fragmented entity relationships. Entities are mostly literal noun phrases or named entities, and edge relationships are largely surface-level.

The BERT-based NER approach extracted a dense set of named entities, which were mapped to the KG using pattern matching. Figure 5 presents the knowledge graph extracted using a BERT-based NER model. This graph demonstrates improved entity labeling accuracy compared to spaCy-only extraction but lacks deep semantic structure or relation-level contextualization.

The LLM+spaCy pipeline demonstrated the most comprehensive semantic coverage, capturing nuanced and long-distance dependencies. Figures 6 to 9 display the results of the spaCy+LLM pipeline. Compared to spaCy-only extraction, the LLM-enhanced knowledge graphs exhibit denser connectivity, richer relationships, and improved co-reference resolution. For instance, in Figure 9 (Tom Cruise subgraph), context-specific associations such as “drone repairman”, “compelling hero”, and “in the picture” emerged, which were absent from the earlier spaCy graphs. Similarly, in Figure 8 (Oblivion subgraph), thematic connections involving the character Jack Harper and descriptors like “outstanding visuals” and “emotional detachment” were captured with significantly higher granularity.

4.2 Neural Network Classification

RNN and LSTM models were evaluated across multiple architectures. The training and validation accuracy/loss plots and confusion matrices (Figures 10–25) indicate performance trends. It was observed that LSTM models converged faster and reached higher accuracy levels compared to their RNN counterparts. In addition, the test accuracy results of eight models are summarized in the table:

Model	Test Accuracy
RNN (1x32)	13.04%
RNN (1x64)	34.78%
RNN (2x32)	13.04%
RNN (2x64)	8.70%
LSTM (1x32)	65.22%
LSTM (1x64)	60.87%
LSTM (2x32)	73.91%
LSTM (2x64)	91.30%

5. Analysis and Interpretations

5.1 Analysis of Knowledge Graph Extraction Methods

A comparative analysis of the spaCy-only, BERT-NER, and spaCy+LLM pipelines highlights differences beyond just entity count—particularly in semantic depth, relation coherence, and ontology structure.

The spaCy-only pipeline suffered from entity fragmentation. Entities such as “Joseph Kosinski” and “director Joseph Kosinski” were treated as distinct nodes, leading to duplication and weaker connectivity. The lack of coreference resolution further limited semantic consolidation across mentions like “he” or “the director.”

The BERT-NER approach improved named entity recognition, capturing key terms like “Tom Cruise” and “Jack Harper.” However, relation extraction remained shallow, as connections between entities were sparse and often lacked meaningful context—indicating limitations when using token-level models without fine-tuning.

In contrast, the spaCy+LLM pipeline showed stronger structure. The graph featured dense, semantically grouped clusters around concepts like “director,” “technology,” and “protagonist.” The LLM also enabled paraphrase abstraction, grouping phrases like “futuristic control caper” and “science fiction adventure” into common narrative arcs. This points to potential for automated equivalent class (EC) formation, which can consolidate varied expressions into unified semantic categories.

5.2 Analysis of Neural Network Classification

The comparison of deep learning models revealed a significant performance disparity between RNN and LSTM architectures. All RNN-based models performed poorly, with test accuracies below 35%. Increasing the number of neurons or layers did not improve RNN performance and in some cases led to degradation, likely due to the inability of standard RNNs to capture long-term dependencies and the vanishing gradient problem.

In contrast, LSTM models demonstrated strong classification capability. The best-performing model (LSTM 2×64) achieved an accuracy of 91.30%, with clear improvement observed as model complexity increased. The LSTM's gated architecture enabled effective retention and propagation of information over longer sequences, which proved essential in modeling movie review text.

Comparison with Assignment 2 results (using machine learning models such as Logistic Regression and SVM) indicates that LSTM models significantly outperformed traditional models. This supports the hypothesis that deep sequential models are more effective for capturing semantic and syntactic patterns in natural language texts.

6. Conclusions

This study demonstrated the comparative strengths of rule-based, NER-driven, and LLM-assisted pipelines for knowledge graph extraction, as well as the superiority of LSTM models for text classification tasks involving temporal dependencies. While spaCy-based extraction provided a foundational structure for entity linking, integration with a large language model substantially enhanced semantic richness and coherence.

On the classification task, LSTM models outperformed RNN and traditional machine learning baselines, with the best-performing model (LSTM 2×64) achieving a 91.30% test accuracy. This confirms the importance of long-range sequence modeling for genre prediction in movie reviews.

Overall, the integration of LLMs for knowledge representation and deep sequential models for classification proved highly effective, demonstrating the value of combining structured extraction with deep learning in NLP tasks. Future extensions may include fine-tuning transformer encoders for end-to-end KG construction and integrating knowledge graph embeddings into neural classifiers to bridge symbolic and neural reasoning.

Appendix

Figure 1: spaCy-only KG graph

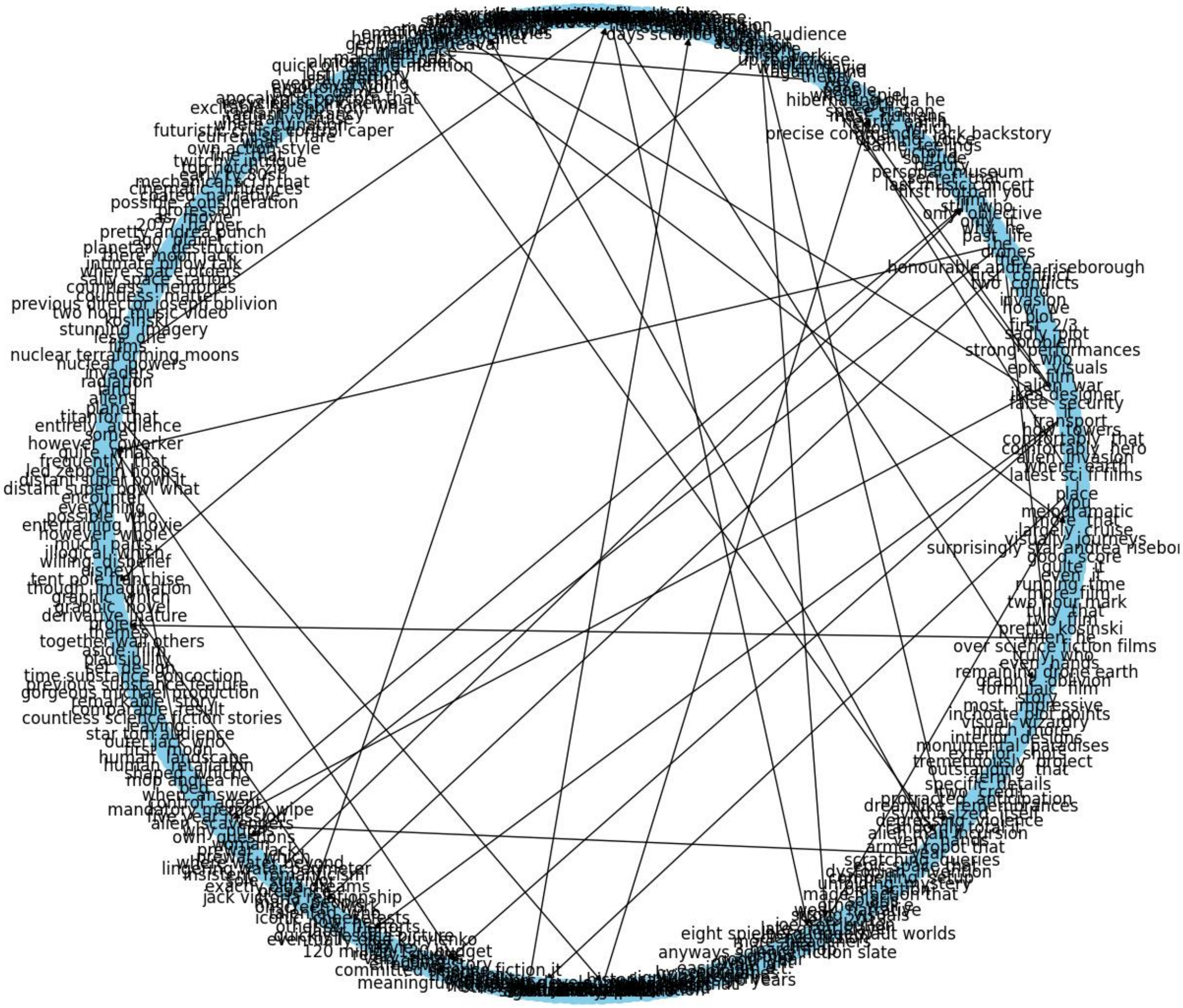


Figure 2: spaCy-only director subgraph

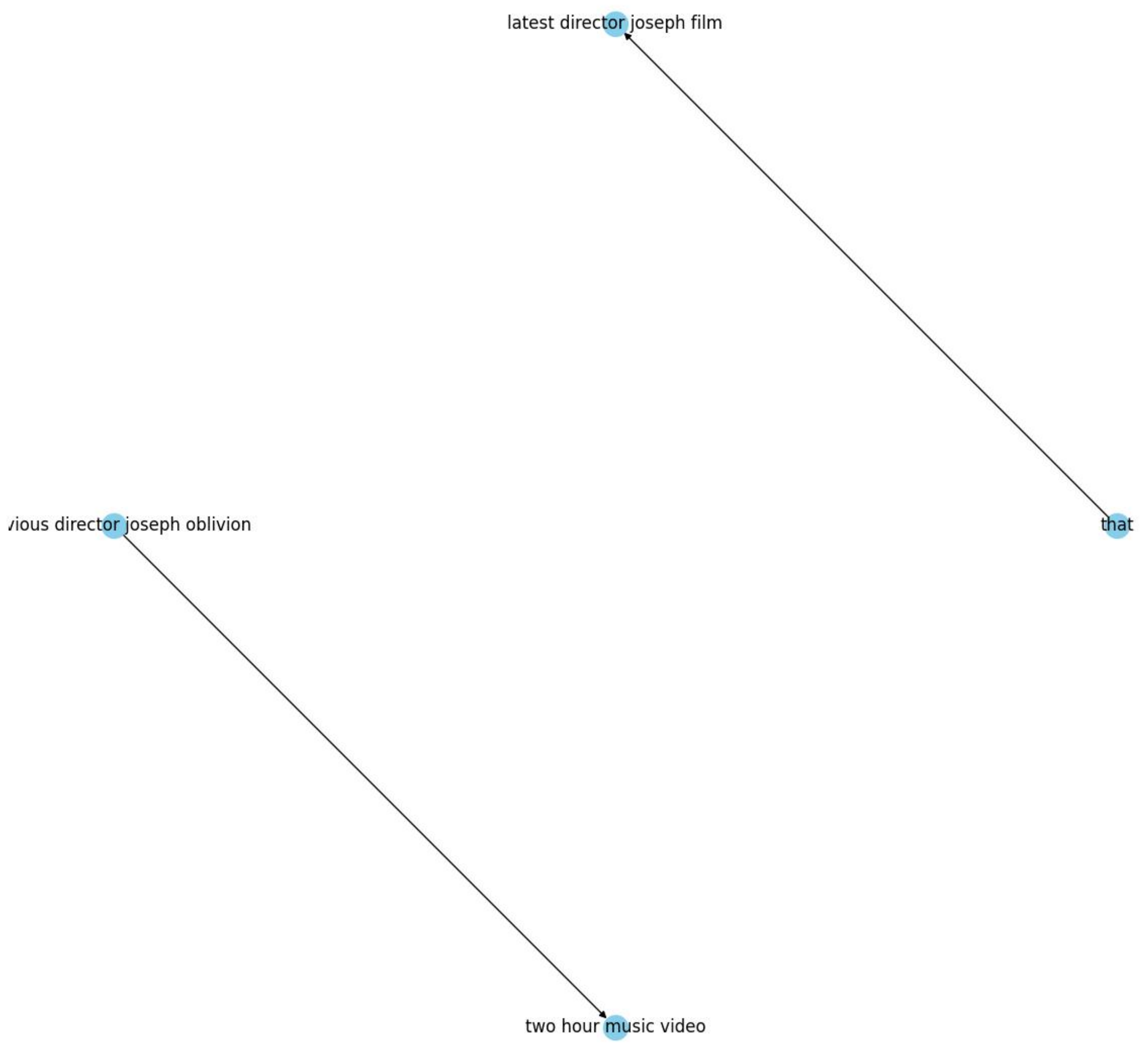


Figure 3: spaCy-only Oblivion subgraph

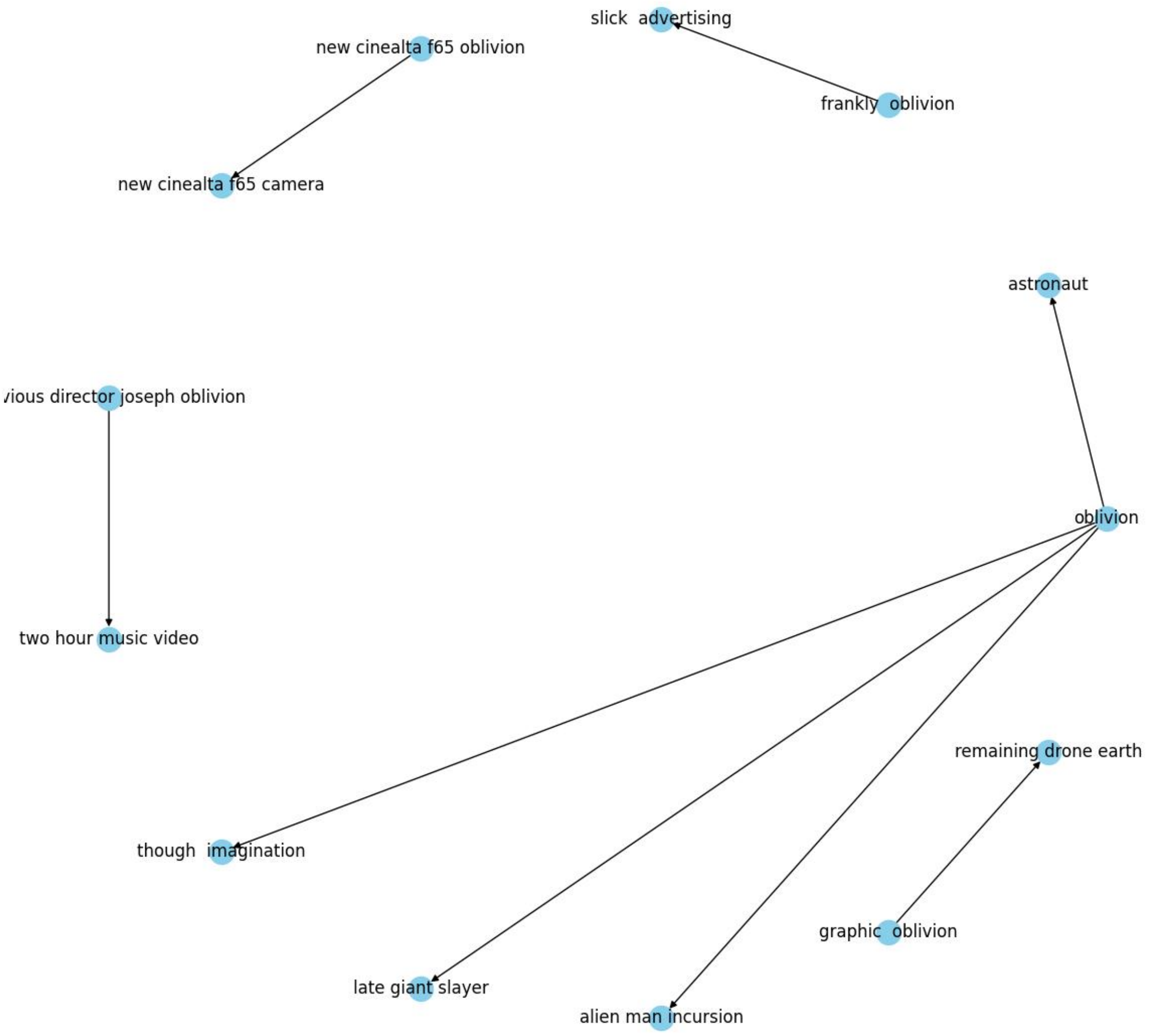


Figure 4: spaCy-only Tom Cruise subgraph

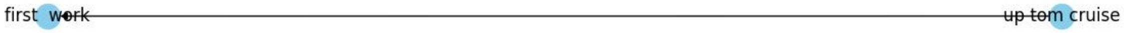


Figure 5: BERT NER extraction graph

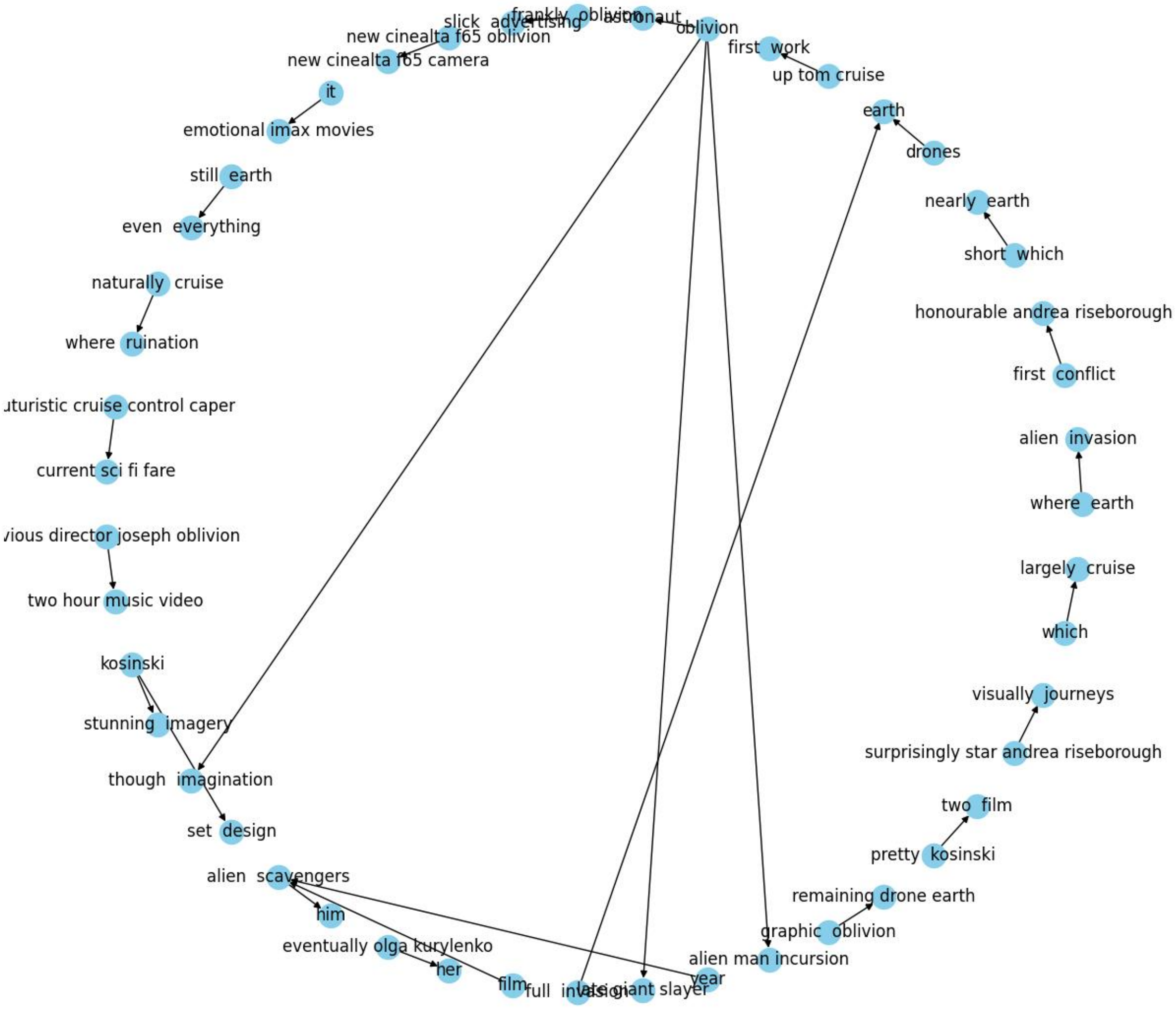


Figure 6: LLM+spaCy KG graph

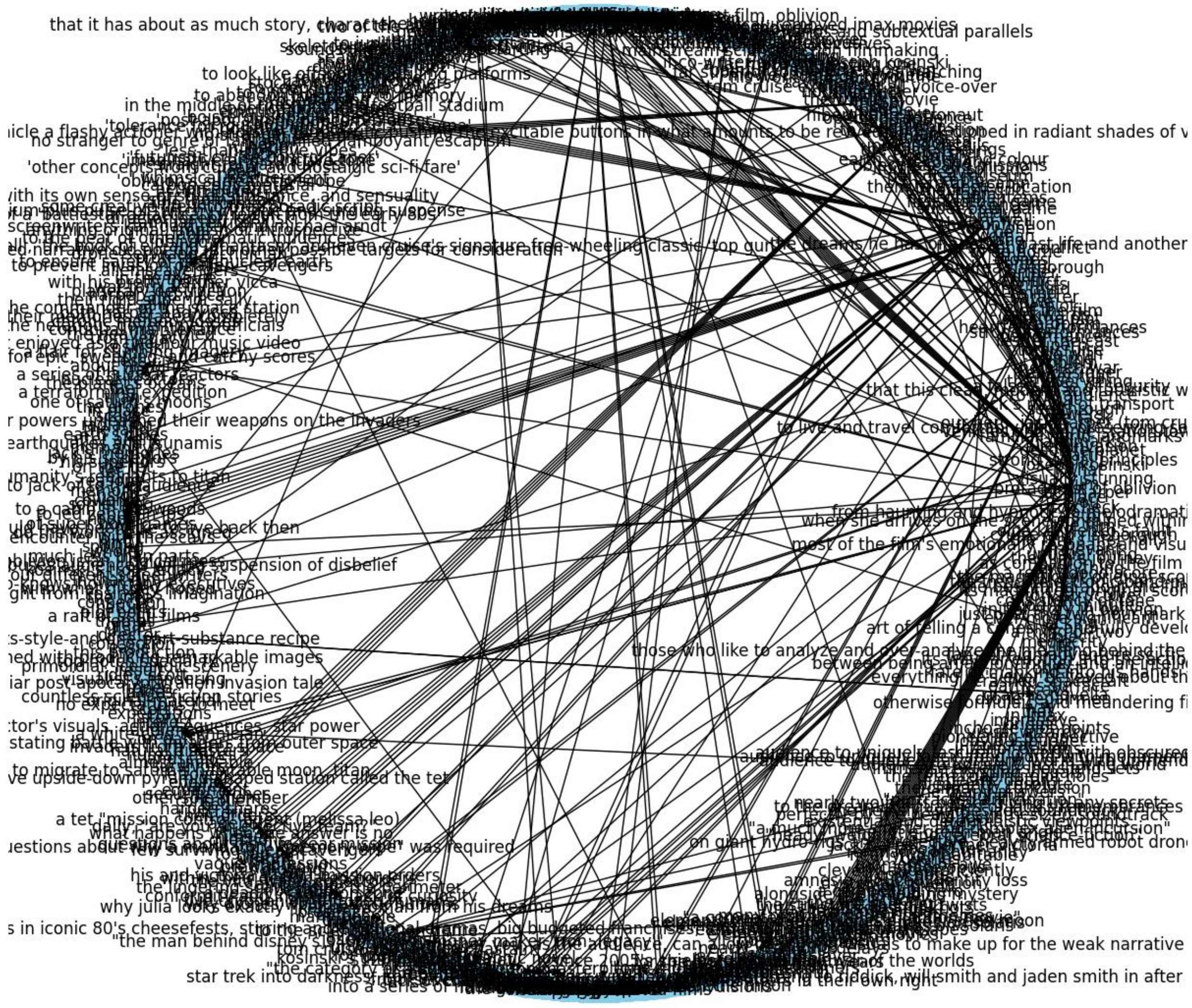


Figure 7: LLM+spaCy director subgraph

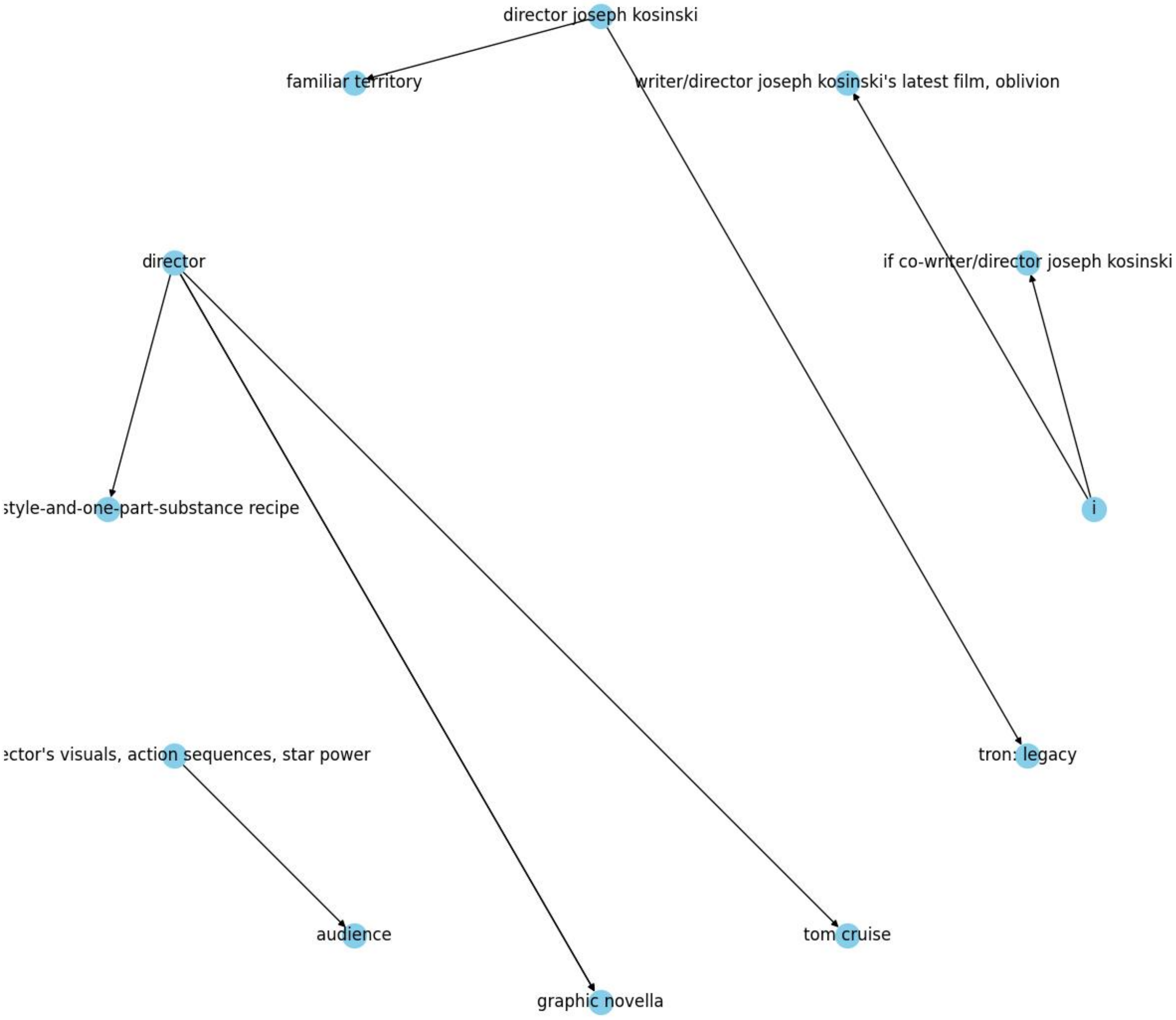

director joseph kosinski
familiar territory
writer/director joseph kosinski's latest film, oblivion
director
if co-writer/director joseph kosinski
style-and-one-part-substance recipe
i
ector's visuals, action sequences, star power
tron: legacy
audience
tom cruise
graphic novella

Figure 8: LLM+spaCy Oblivion subgraph

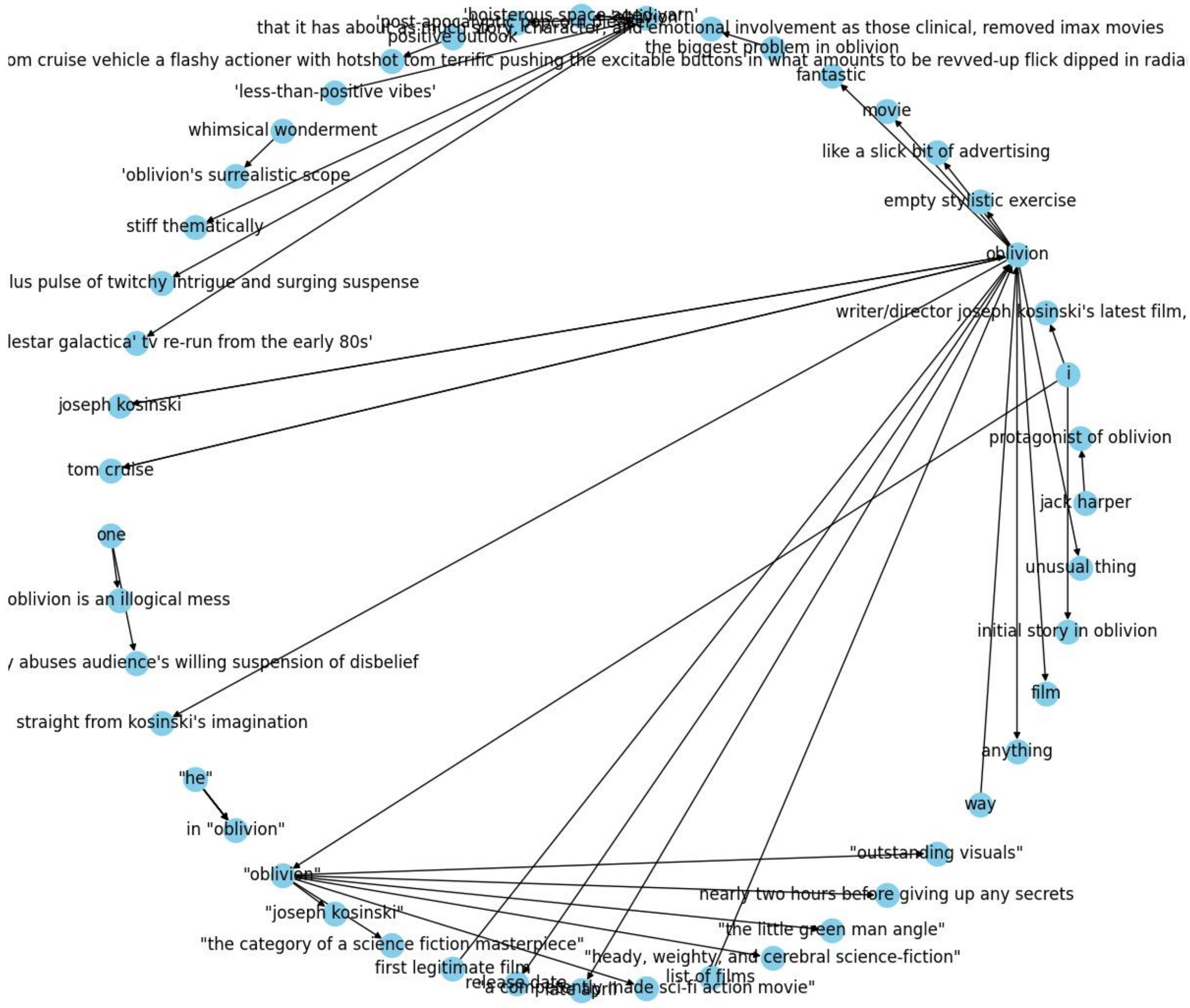

'boisterous space-age yarn'
'post-apocalyptic popcorn pleaser'
'oblivion'
that it has about as much story, character, and emotional involvement as those clinical, removed imax movies
positive outlook
the biggest problem in oblivion
om cruise vehicle a flashy actioner with hotshot tom terrific pushing the excitable buttons in what amounts to be revved-up flick dipped in radia
fantastic
'less-than-positive vibes'
movie
whimsical wonderment
like a slick bit of advertising
'oblivion's surrealistic scope
empty stylistic exercise
stiff thematically
oblivion
lus pulse of twitchy intrigue and surging suspense
writer/director joseph kosinski's latest film,
lestar galactica' tv re-run from the early 80s'
i
joseph kosinski
protagonist of oblivion
tom cruise
jack harper
one
unusual thing
oblivion is an illogical mess
initial story in oblivion
/ abuses audience's willing suspension of disbelief
film
straight from kosinski's imagination
anything
"he"
way
in "oblivion"
"outstanding visuals"
"oblivion"
nearly two hours before giving up any secrets
"joseph kosinski"
"the little green man angle"
"the category of a science fiction masterpiece"
"heady, weighty, and cerebral science-fiction"
first legitimate film
list of films
release date
"a competently made sci-fi action movie"
late april

Figure 9: LLM+spaCy Tom Cruise subgraph

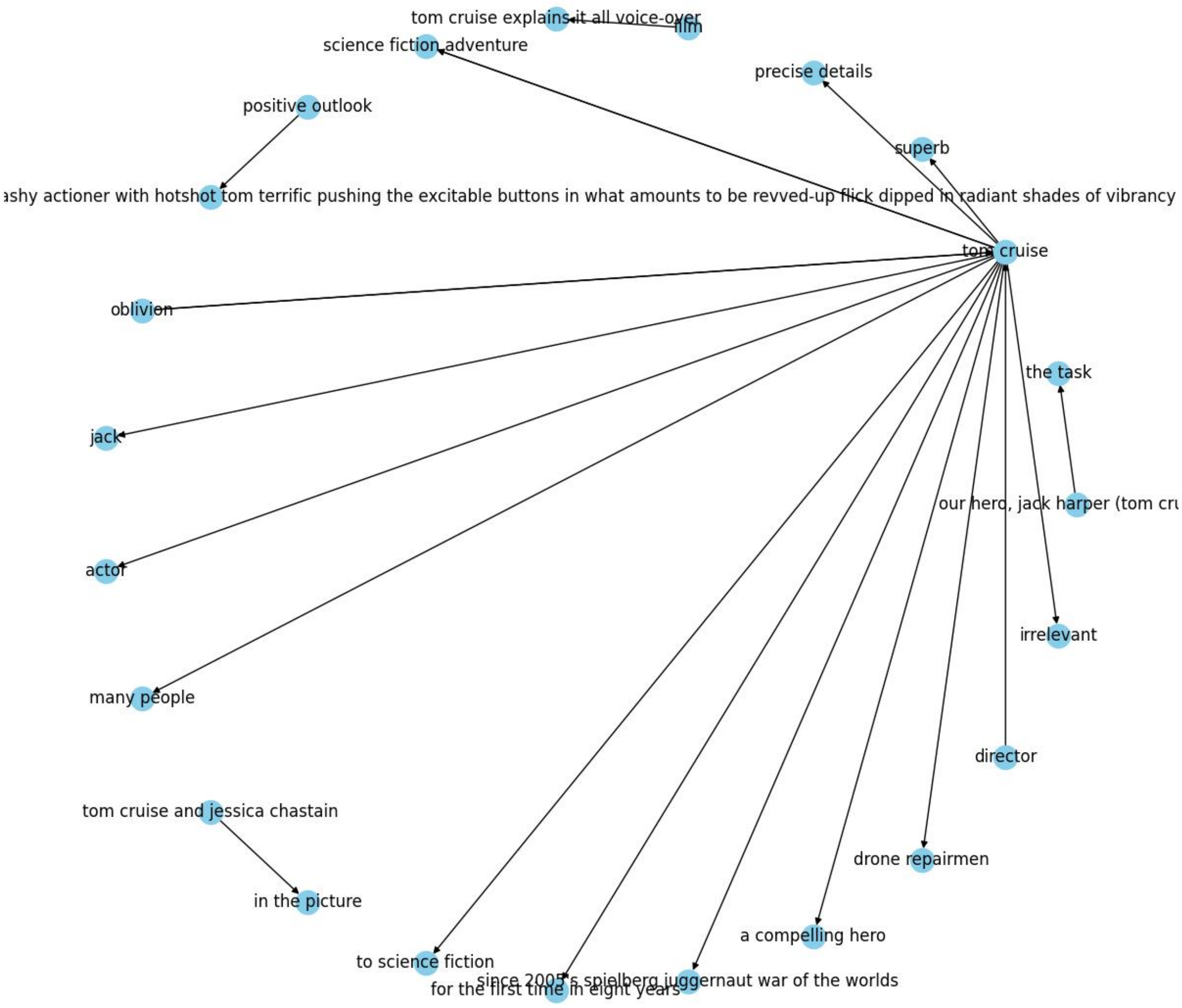
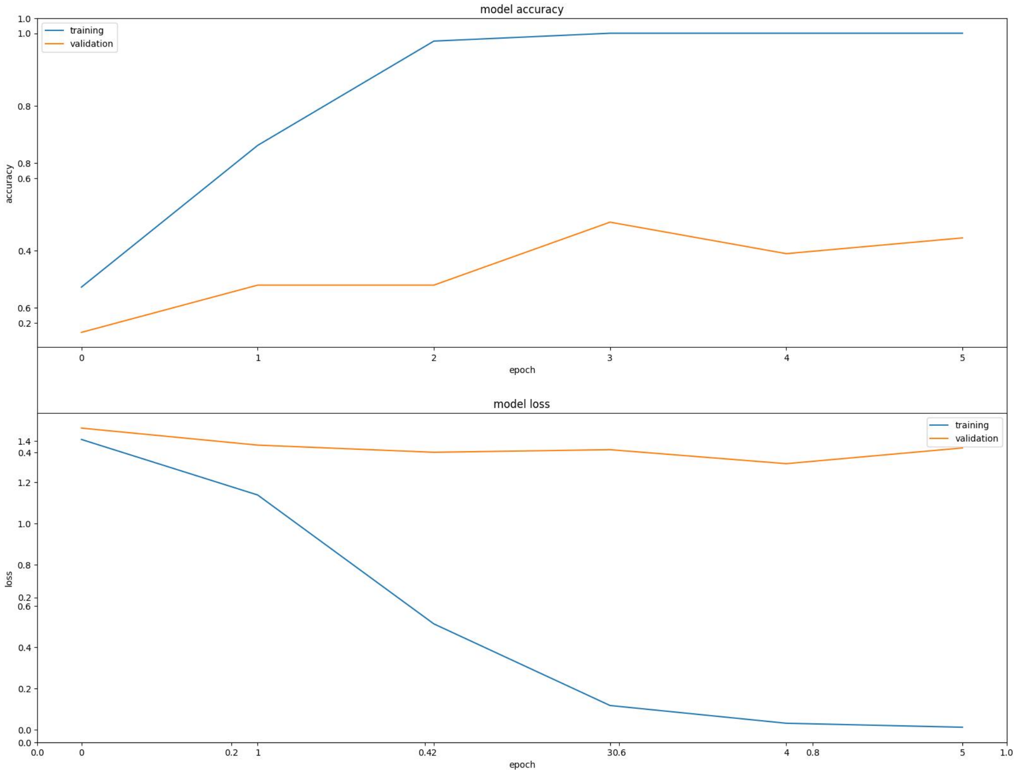

tom cruise explains it all voice-over
film
science fiction adventure
precise details
positive outlook
superb
ashy actioner with hotshot tom terrific pushing the excitable buttons in what amounts to be revved-up flick dipped in radiant shades of vibrancy
tom cruise
oblivion
the task
jack
our hero, jack harper (tom cru
actor
irrelevant
many people
director
tom cruise and jessica chastain
drone repairmen
in the picture
a compelling hero
to science fiction
since 2005's spielberg juggernaut war of the worlds
for the first time in eight years

Figure 10&11: RNN model accuracy/loss & confusion matrices (1 layer, 32 neurons)



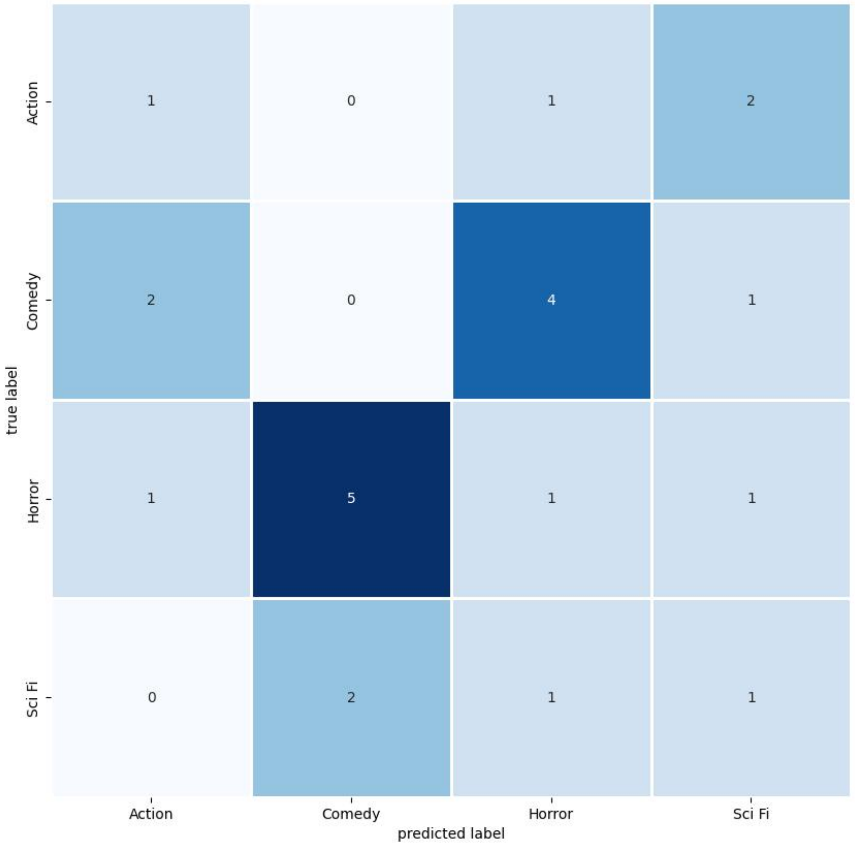
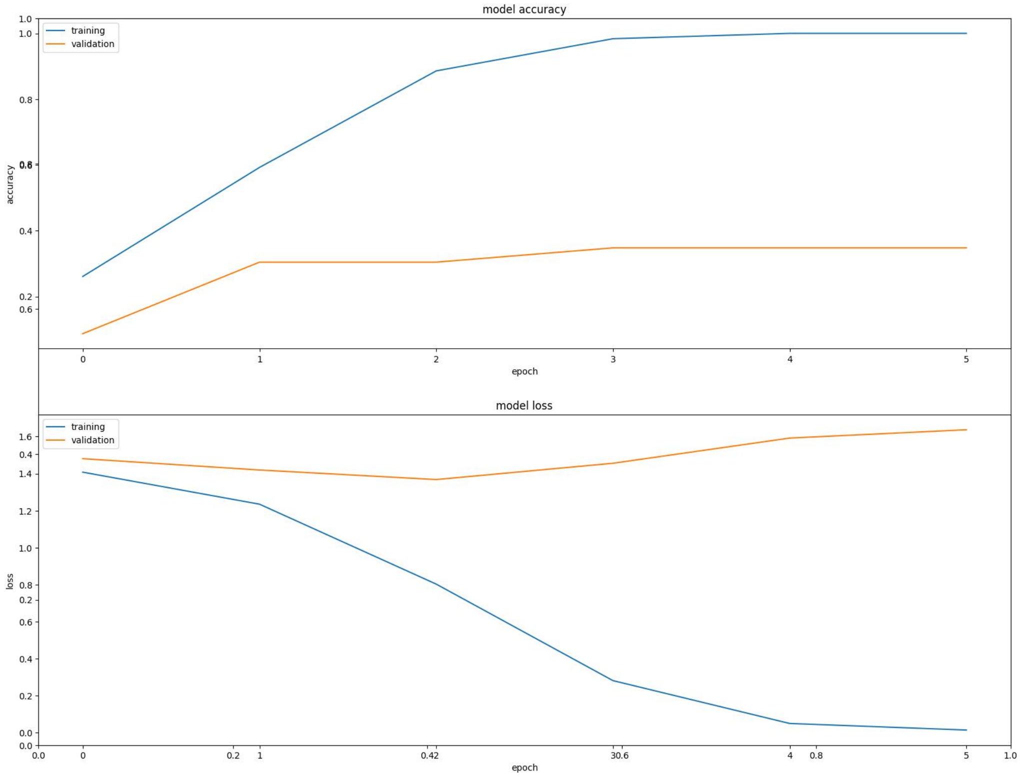


Figure 12&13: RNN model accuracy/loss & confusion matrices (1 layer, 64 neurons)



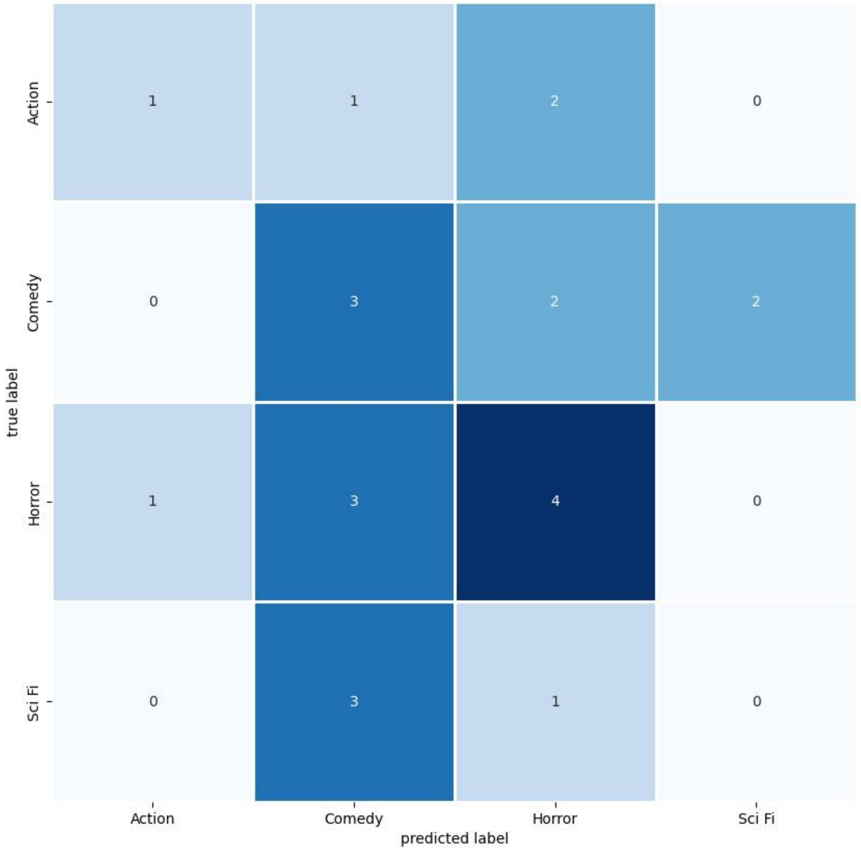
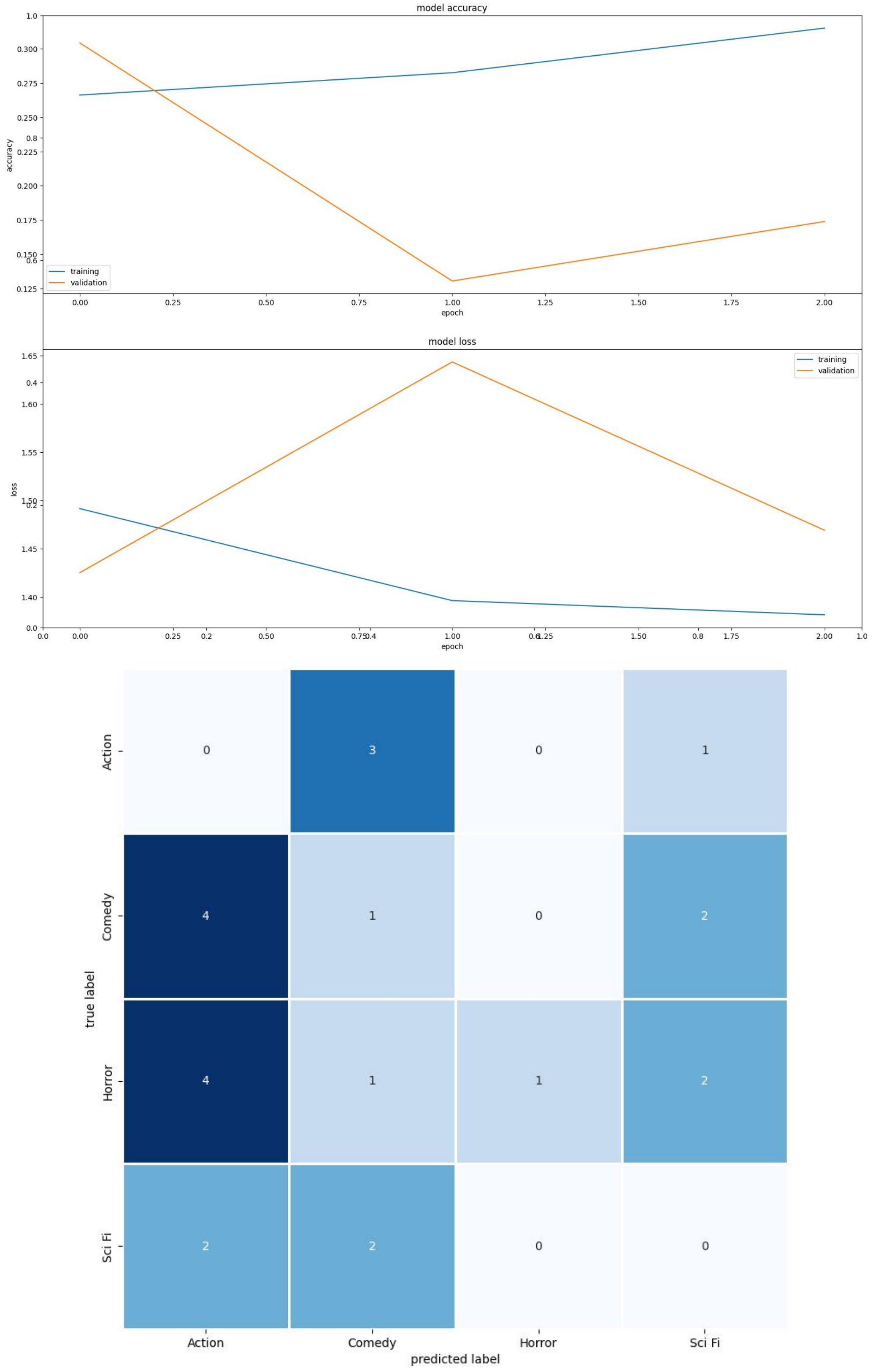
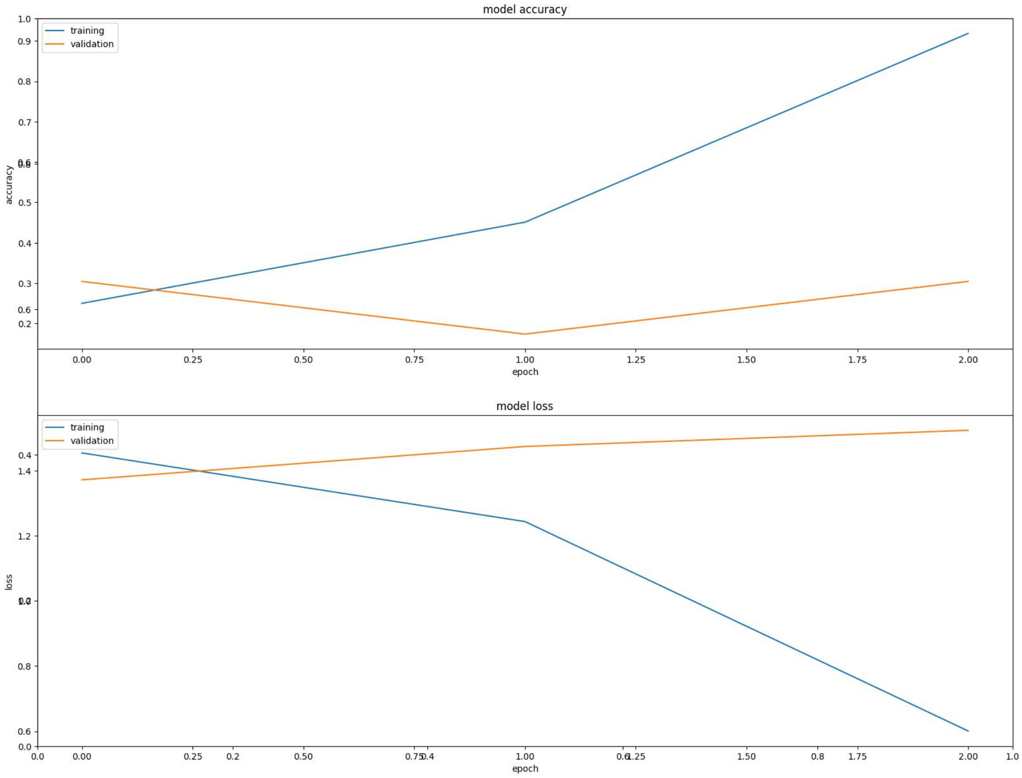


Figure 14&15: RNN model accuracy/loss & confusion matrices (2 layers, 32 neurons)



	Action	Comedy	Horror	Sci Fi
Action	0	3	0	1
Comedy	4	1	0	2
Horror	4	1	1	2
Sci Fi	2	2	0	0

Figure 16&17: RNN model accuracy/loss & confusion matrices (2 layer, 64 neurons)



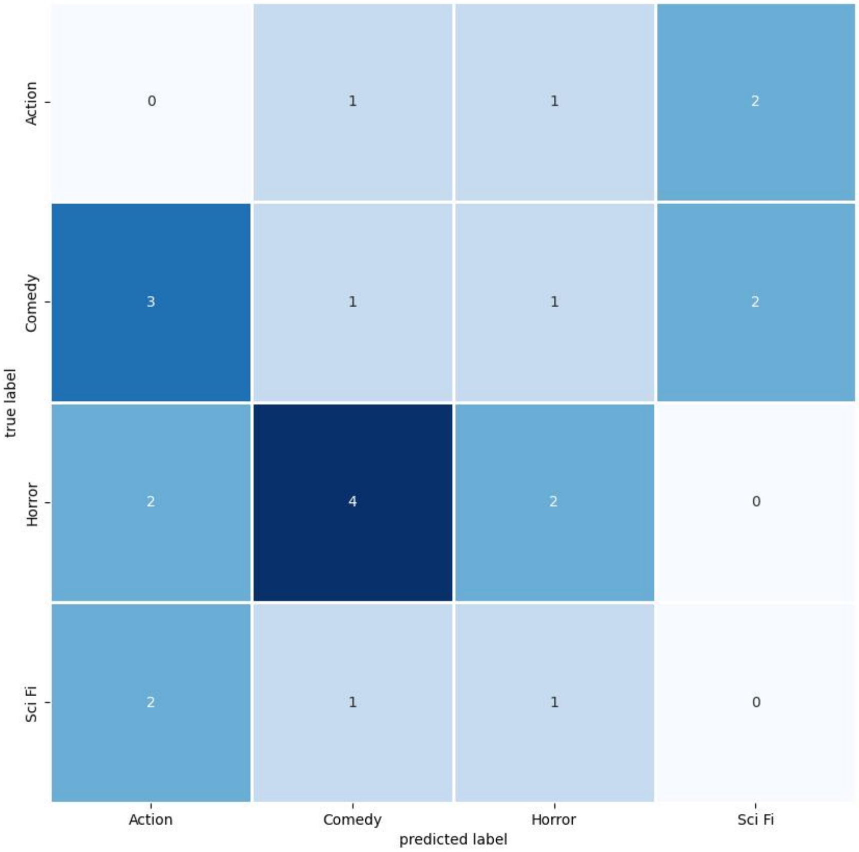


Figure 18&19: LSTM model accuracy/loss & confusion matrices (1 layer, 32 neurons)

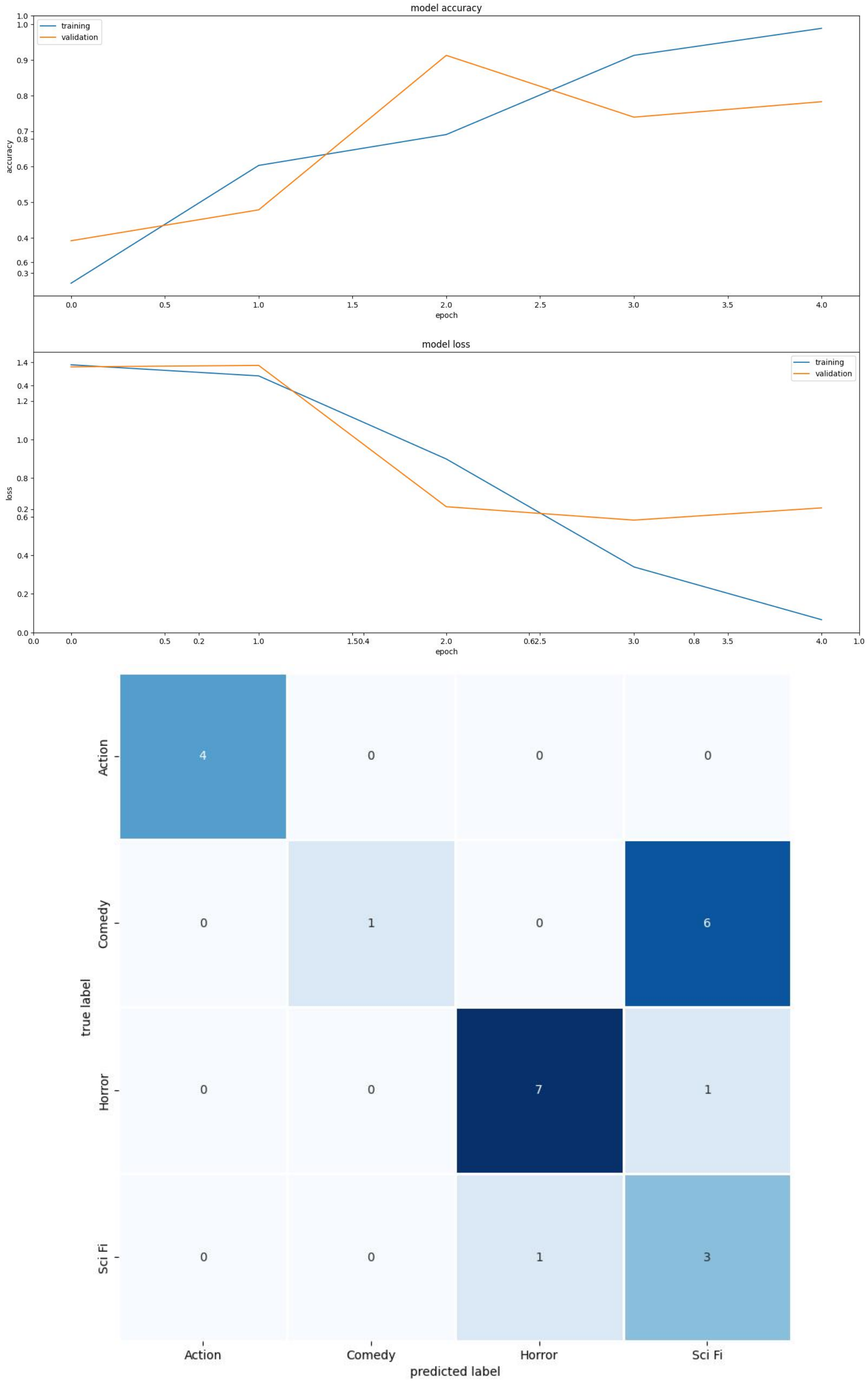


Figure 20&21: LSTM model accuracy/loss & confusion matrices (1 layer, 64 neurons)

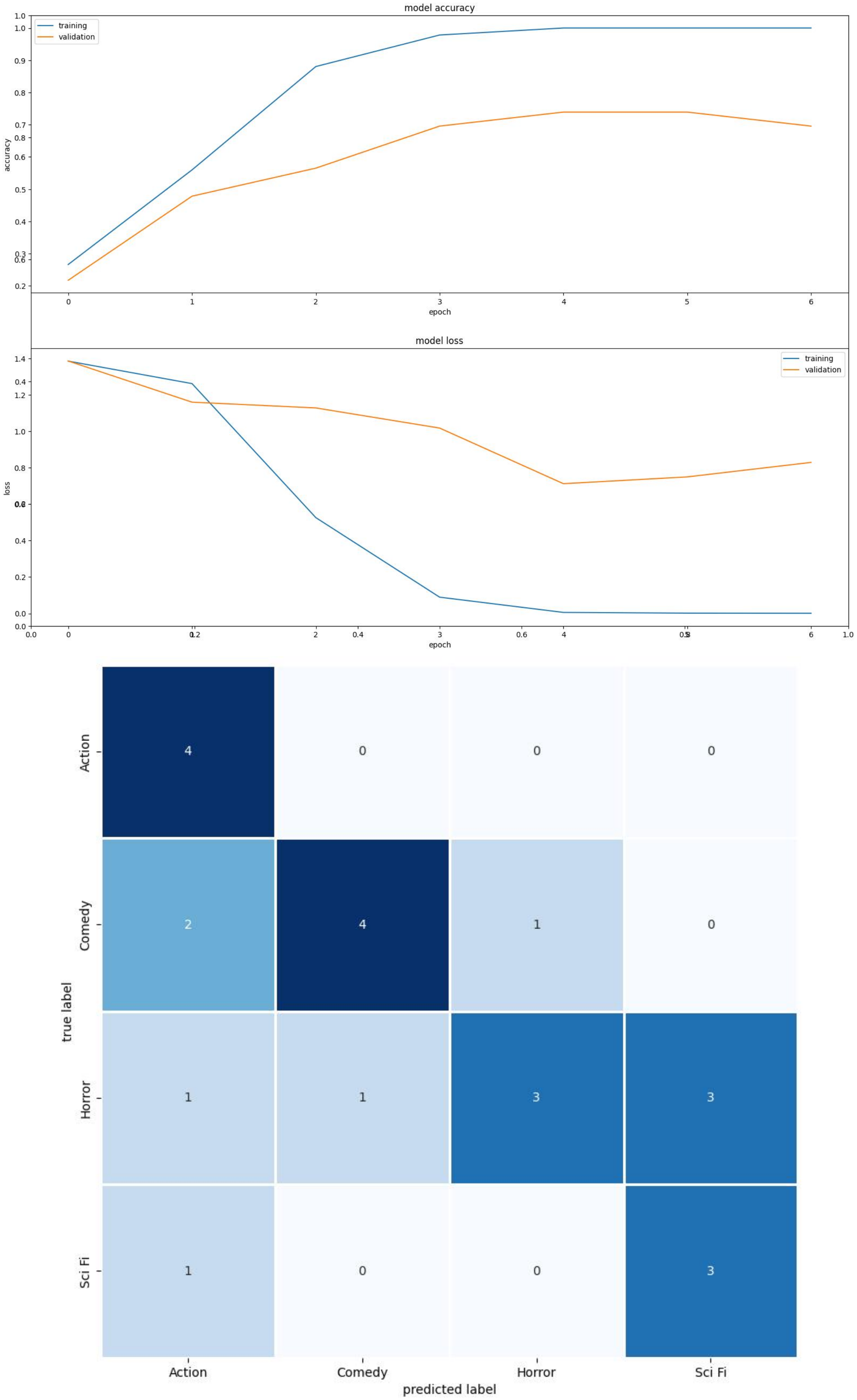


Figure 22&23: LSTM model accuracy/loss & confusion matrices (2 layers, 32 neurons)

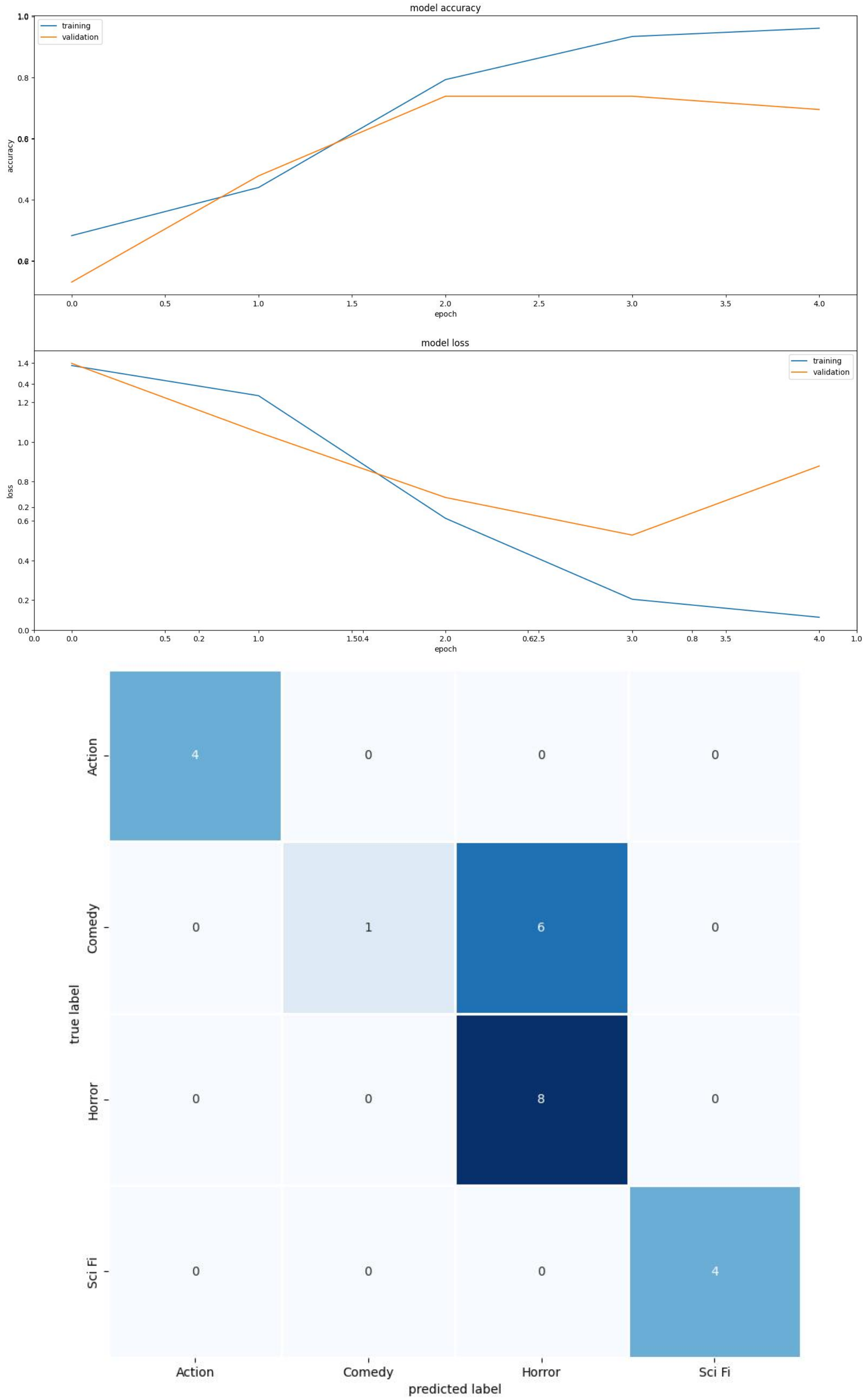
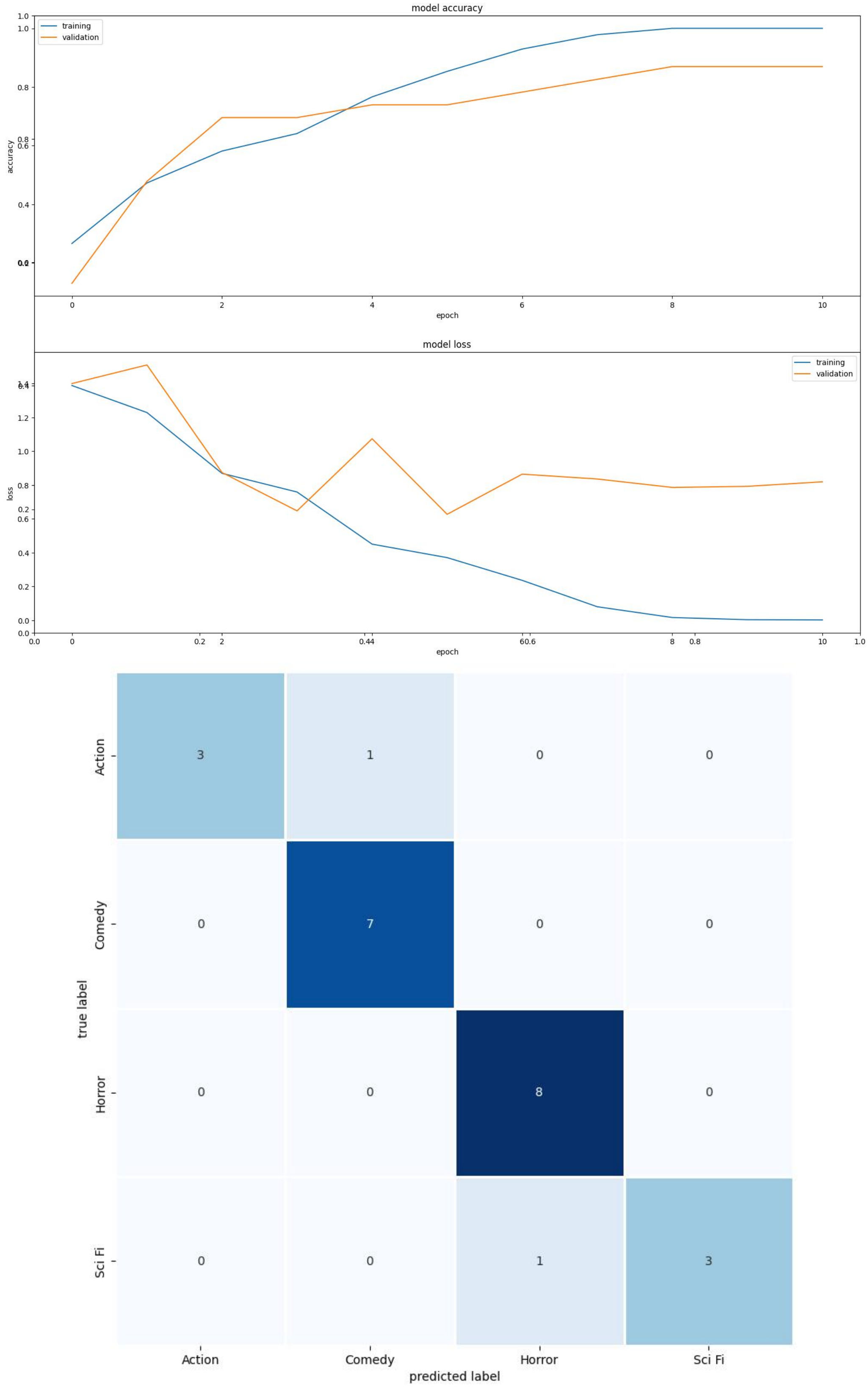


Figure 24&25: LSTM model accuracy/loss & confusion matrices (2 layer, 64 neurons)


model accuracy
training
validation
accuracy
epoch
model loss
training
validation
loss
epoch
true label
Action
Comedy
Horror
Sci Fi
predicted label
Action
Comedy
Horror
Sci Fi