Internship Report

Internship at Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), India

Intern: **Mohamed Ahmed Mansour**

Supervisor: **Prof. Ramakrishna**

Duration: 6 Weeks (Ongoing)

Table of Contents

1. Introduction

I completed a 6-week internship at Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), India, under the guidance of Prof. Ramakrishna. The internship centered on building a **Dual-Branch Fake News Detection Framework** that combines a **BERT**-based semantic branch with a **TransE**-based knowledge-graph branch. The motivation is to detect misinformation by using both **context** (what the text says) and **factual consistency** (what a knowledge graph knows).

Along the way, I studied core NLP concepts—polysemy, preprocessing, tokenization, embeddings—and developed supporting projects (traditional ML baselines and a news category classifier). This report summarizes weekly progress, key skills, datasets, challenges, and planned future work.

2. Weekly Work Summary

Week 1

• Read two research articles recommended by Prof. Ramakrishna to understand state-of-the-art approaches.
• Studied **polysemy in NLP** and how different writing styles can lead to multiple interpretations in news.
• Learned text preprocessing fundamentals: cleaning, tokenization (sentence/word/character), stopword handling.
• Repository: Polysemy NLP Project

Week 2

• Deep dive into text processing: **stemming vs. lemmatization**, **POS tagging**, and **n-grams**.
• Explored representations: One-Hot, TF-IDF, and dense embeddings (Word2Vec/GloVe/BERT).
• Built the **first prototype** of the dual-branch framework; trained BERT and TransE separately but no fusion—pipeline unstable.

Week 3

• Restructured the pipeline and implemented **version 0.2** with an initial fusion module (did not converge).
• Implemented a **Fake News Detection** baseline using Logistic Regression, SVM, and Random Forest—good baseline performance.
• Repository: Fake News Detection (ML Baselines)

Week 4

• Developed **version 0.3**: integrated **REBEL** for triplet extraction and fuzzy matching to a knowledge graph; stabilized training.
• Combined BERT (text semantics) with TransE (triplet logic) using a more reliable fusion layer; first functional end-to-end prototype.
• Repository: Dual-Branch Framework (v0.3)

Week 5

• Built a **News Category Classification** model to classify articles by topic; improved documentation and reproducibility across repos.
• Added a simple **GUI** (Gradio), a **license**, and refined the project structure for the dual-branch framework.
• Wrote the **first draft** of the research paper.
• Repository: News Category Classification

Week 6 (Ongoing)

• Prepared this internship report and started refining the research paper for review and submission.

3. Skills and Tools Learned

Technical: Text preprocessing, tokenization, stemming, lemmatization, POS tagging; ML algorithms (LogReg, SVM, Random Forest); deep learning with **BERT** and **TransE**; triplet extraction with **REBEL**; fuzzy matching; GUI with Gradio; Hugging Face Transformers, PyTorch, OpenKE.

Soft: Research reading & summarization, debugging complex ML pipelines, documentation, time management, iterative experimentation.

4. Datasets Used

• **FakeNewsNet** (PolitiFact + GossipCop): news content with labels and social context.
• **PolitiFact**: political claims/articles with real/fake labels.
• **GossipCop**: celebrity/entertainment news with credibility labels.
• **LIAR**: short political statements labeled on a truthfulness scale; useful for text-branch training/validation.

5. Challenges and Solutions

• **Unstable pipeline in early versions**: Fusion missing or mis-specified → *Solution*: re-architected training loop; added robust fusion and better preprocessing.
• **Integrating semantics with factual knowledge**: Hard to align triplets with KG → *Solution*: introduced REBEL extraction + fuzzy matching before TransE scoring.
• **Resource constraints**: Large model training time → *Solution*: tuned batch sizes/epochs; used smaller baselines to guide expectations.

6. Conclusion and Future Work

The internship took me from theory (polysemy, preprocessing) to practice (end-to-end framework and baselines). Next steps include improving fusion, experimenting with better triplet scoring and KG coverage, and extending to multimodal inputs (text + images). I will also polish the research paper and evaluate on multiple datasets for stronger benchmarks.

7. References

- Polysemy NLP Project
- Fake News Detection (ML Baselines)
- Dual-Branch Framework (v0.3)
- News Category Classification
- FakeNewsNet, PolitiFact, GossipCop, LIAR datasets