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

Internship Report

Intern: Mohamed Ahmed Mansour
Supervisor: Prof. Ramakrishna

Duration: 6 Weeks (Ongoing)

Abstract

This report documents my 6-week internship at Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), India, under the supervision of Prof. Ramakrishna. The primary focus was the development of a **Dual-Branch Fake News Detection Framework** that combines semantic reasoning (BERT) with factual verification (TransE + Knowledge Graphs).

The project aimed to address the growing challenge of misinformation by integrating both *linguistic understanding* and *factual consistency*. This approach differs from traditional fake news detection systems which often rely only on textual features. The internship also included studying polysemy in NLP, implementing machine learning baselines, and experimenting with a news category classification model. Weekly progress involved moving from theoretical study to building practical ML/DL pipelines and producing a draft research paper.

The experience enhanced my expertise in NLP and deep learning, improved my research and documentation skills, and prepared me for future academic and industry challenges.

1. Introduction

Fake news has become one of the most pressing problems of the digital era. Social media platforms and online news sites enable rapid information dissemination, but they also facilitate the spread of misinformation. Automated methods for fake news detection are therefore critical to maintaining information integrity.

This internship provided the opportunity to work on an advanced NLP project under academic guidance at MIT MAHE. The research problem was to detect fake news by combining **semantic analysis** and **knowledge-based reasoning**. The framework integrates BERT embeddings for contextual understanding with TransE embeddings derived from triplets aligned to a knowledge graph. This dual-branch design seeks to balance *how news is written* with *what facts it conveys*.

Previous research has mostly explored either deep contextual embeddings (e.g., BERT, RoBERTa) or knowledge graph embeddings (e.g., TransE, DistMult) in isolation. The novelty of this project lies in fusing both paradigms.

The report is structured to present weekly progress, skills learned, datasets used, challenges encountered, solutions applied, and future directions.

2. Weekly Work Summary

Week 1

- Read two research articles recommended by Prof. Ramakrishna to understand state-of-the-art approaches.
- Studied the concept of **polysemy in NLP** and its impact on interpretation of news.
- Learned preprocessing: cleaning, tokenization, stopword removal.
- Repository: magentaPolysemy NLP Project.

Week 2

- Explored stemming vs. lemmatization, POS tagging, and n-grams.
- Studied word representations: One-Hot, TF-IDF, Word2Vec, BERT.
- Built the first version of the framework (BERT + TransE) but fusion failed.

Week 3

- Improved the pipeline and implemented **version 0.2** with an early fusion module (still unstable).
- Built ML baselines for fake news detection: Logistic Regression, SVM, Random Forest.
- Repository: magentaFake News Detection (ML Baselines).

Week 4

- Created **version 0.3** with REBEL triplet extraction + fuzzy matching.
- Achieved the first stable prototype combining BERT and TransE.
- Repository: magentaDual-Branch Framework.

Week 5

- Developed a **News Category Classification** model.
- Enhanced version 0.3: added GUI (Gradio), license, and improved documentation.
- Drafted the first version of the research paper.
- Repository: magentaNews Category Classification.

Week 6 (Ongoing)

- Writing this report and refining the research paper.

3. Skills and Tools Learned

Technical Skills

- Preprocessing, tokenization, stemming, lemmatization.
- ML: Logistic Regression, SVM, Random Forest.
- DL: BERT, TransE, and fusion models.
- Knowledge Graph alignment and triplet extraction using REBEL.
- Libraries: PyTorch, Hugging Face Transformers, OpenKE, Gradio.

Research and Soft Skills

- Research reading and summarization.
- Debugging and improving ML pipelines.
- Academic paper writing.
- Documentation and reproducibility in GitHub projects.
- Collaboration and time management.

4. Datasets Used

The project leveraged four major datasets. Table **??** summarizes their scope.

Dataset	Domain	Size	Labels
FakeNewsNet	Mixed (news + social)	∼23K articles	Real / Fake
PolitiFact	Political news	12K claims	Real / Fake
GossipCop	Entertainment news	22K articles	Real / Fake
LIAR	Political statements	12.8K short claims	Truthfulness scale

Table 4.1: Datasets used in the internship project.

Each dataset provided unique challenges. FakeNewsNet offered combined article and social-context information. PolitiFact and GossipCop enabled balanced fake/real news classification. LIAR provided short text samples with fine-grained labels, useful for BERT fine-tuning.

5. Challenges and Solutions

- **Unstable pipeline in early versions:** Fusion modules failed to converge. *Solution:* Re-architected training loop, added better preprocessing, used smaller baseline models to stabilize.
- **Integrating semantics with factual knowledge:** Hard to align extracted triplets with KG facts. *Solution:* Introduced REBEL triplet extraction + fuzzy matching before TransE scoring.
- **Resource constraints:** Large model training required significant compute. *Solution:* Tuned batch sizes, epochs, and optimized pipeline. Used ML baselines to set performance benchmarks.

6. Conclusion and Future Work

This internship combined theoretical learning with practical implementation. The highlight was the dual-branch framework that integrates semantics and factual reasoning for fake news detection.

Future Work:

- Improve fusion methods with advanced neural architectures.
- Extend framework to multimodal inputs (text + images).
- Expand experiments to more datasets.
- Finalize and publish the research paper.

Acknowledgements

I sincerely thank **Prof. Ramakrishna** for his guidance, valuable feedback, and encouragement throughout this internship. I am grateful to **MIT MAHE** for providing the resources and opportunity. I also thank my peers and the supportive research environment that enabled me to overcome challenges and make steady progress.

References

- magentaPolysemy NLP Project
- magentaFake News Detection
- magentaDual-Branch Framework
- magentaNews Category Classification
- Datasets: FakeNewsNet, PolitiFact, GossipCop, LIAR