ELKP: Knowledge-Powered Rumor Detection

Enhancing social media fact-checking with external knowledge graphs. A minimalist implementation exploring how "world knowledge" bridges the gap between language understanding and factual verification.

THE CHALLENGE

Why Language Models Get Fooled

The Core Problem

Standard models focus on grammar and style, not facts. A fake tweet can sound perfectly official.

The Limitation

BERT analyzes linguistic patterns—tone, formality, structure—but lacks access to factual databases.

The Risk

"The Eiffel Tower has been sold to a private investor" sounds credible, yet is completely fabricated.

The Gap

Without external verification, AI cannot distinguish well-written fiction from truth, leading to false positives.

CORE CONCEPT

The ELKP Framework

What is ELKP?

Entity-Link Knowledge-Powered framework—a system that extracts entities from claims and cross-references them against trusted sources.

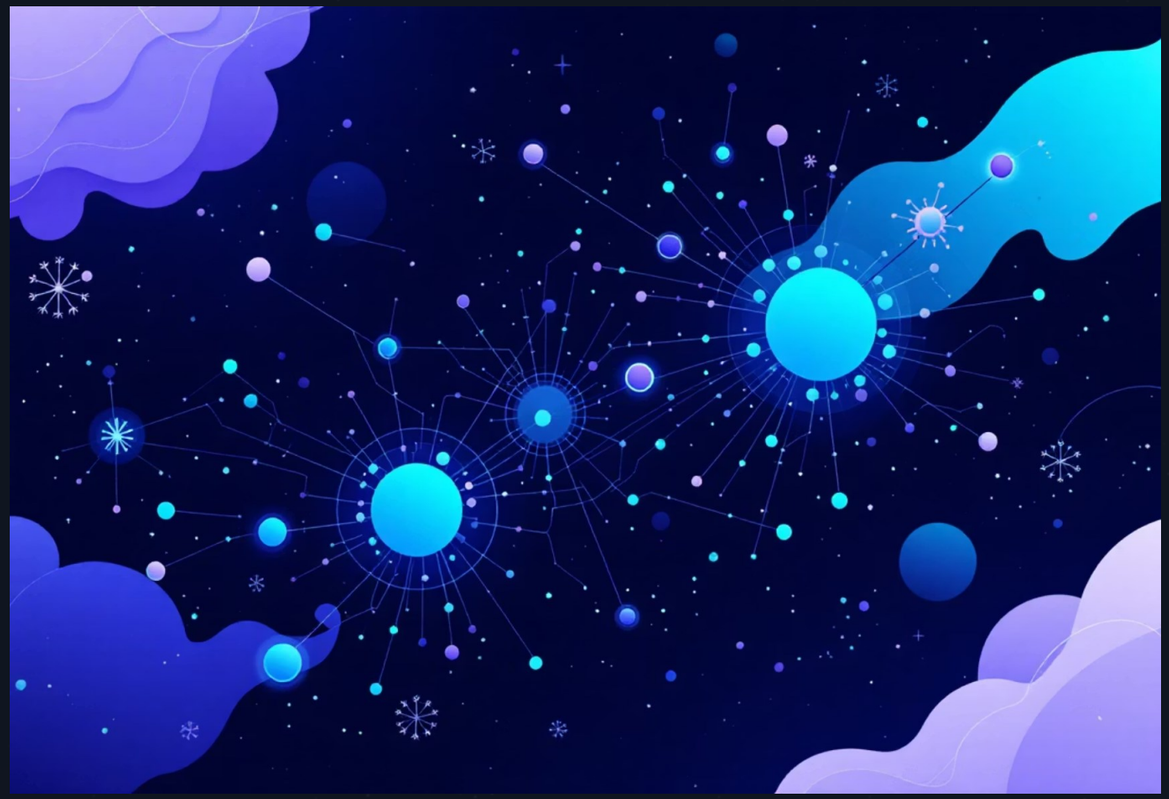
The Strategy

Instead of immediate classification, we first "fact-check" claims against Wikipedia's knowledge base before making decisions.

The Paradigm Shift

Moving from "Does this sound like a rumor?" to "Does this contradict documented facts?"—a fundamental rethinking of verification.

Bridging the Knowledge Gap



The research paper identifies that rumors frequently contain verifiable entities—people, places, events—that can be cross-referenced against established knowledge sources.

01

Entity Identification

Extract core entities from the claim using NER

02

Knowledge Retrieval

Query trusted sources for contextual information

03

Context Fusion

Combine retrieved knowledge with original content

04

Classification

Analyze augmented text for contradictions

SYSTEM ARCHITECTURE

The ELKP Pipeline

The system transforms a raw tweet like "Michael Jackson seen on an island" by detecting entities, retrieving contradictory facts ("Died in 2009"), and feeding the augmented context to BERT for final classification.

Modular & Clean Architecture

prepare_data.py

The ETL Pipeline

Parses complex nested JSON from PHEME dataset into clean CSV format for analysis.

preprocess.py

The Knowledge Engine

KnowledgeEngine class handles Spacy NER and Wikipedia API to inject contextual knowledge.

main.py

The Trainer

Manages data splitting to prevent leakage, fine-tunes BERT, calculates performance metrics.

demo_notebook.ipynb

Interactive Testing

Real-time interface for testing the system with custom inputs and exploring predictions.

RESULTS

Performance Metrics

85.14%	0.81	0.88
Accuracy	Rumor F1	Non-Rumor F1
Overall classification accuracy	F1-score for rumor detection	F1-score for verified content

Model Configuration

Architecture: BERT-Base-Uncased fine-tuned for 3 epochs on augmented PHEME dataset.

Key Observation: Significant improvement over baseline raw BERT by leveraging injected knowledge context. The model demonstrates stronger performance on non-rumor classification, suggesting effective knowledge integration.

ERROR ANALYSIS

When the Model Fails

> "The PM's office releases a statement about #sydneysiege."

Prediction: Real (Incorrect) | Actual: Rumor

Critical Lesson: System accuracy depends heavily on entity recognition quality. Better NER is essential for reliable knowledge augmentation.

NER Failure

Failed to recognize "#sydneysiege" as a key event entity—hashtag format confused the detector.

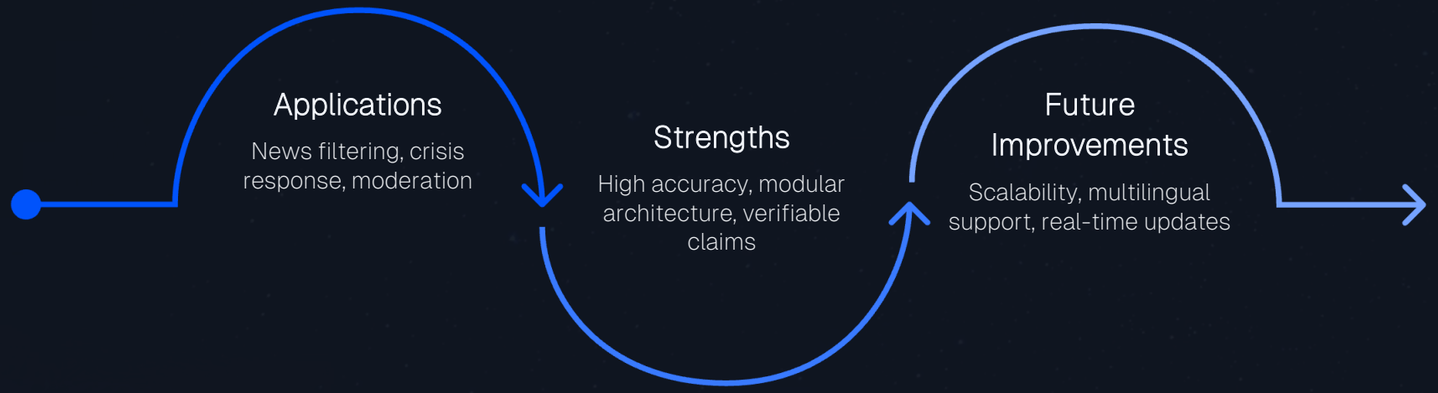
Bad Retrieval

Retrieved Wikipedia info about the word "About" instead of the Sydney siege event.

Garbage In, Garbage Out

Without proper context, model defaulted to analyzing formal tone, missing the rumor entirely.

Where Do We Go From Here?



Real-World Applications

- Automated news filtering systems
- Crisis management platforms
- Social media content moderation

System Strengths

High accuracy on verifiable claims with robust, modular architecture enabling easy experimentation and deployment.

Future Improvements

- Upgrade to Transformer-based NER (RoBERTa)
- Integrate real-time search APIs (Google) instead of static Wikipedia
- Multi-lingual knowledge retrieval for global coverage

Thank You

Questions & Discussion

Exploring the intersection of natural language processing, knowledge graphs, and misinformation detection.