Deliverable 2: Detailed Technique Report

Algorithm Overview

The credibility score function follows a hybrid approach, combining rule-based heuristics and machine learning techniques in its evaluation process with the first step being text extraction. This step extracts content from articles from a provided URL using BeautifulSoup, a Python library used to pull data out of HTML and XML files. The next step is evaluating a domain’s trustworthiness in which a domain authority score is assigned based on the source’s reputation. In this case, sources with high authority such as academic journals, receive higher trust scores. The following step is evaluating content relevance which uses Hugging Face’s sentence-transformers/all-mpnet-base-v2 model to measure and compute semantic similarity between user queries and extracted text. This measures how closely the extracted text aligns with the user query using cosine similarity. Next, fact-checking is implemented by querying the Google Fact Check API to determine if the content has been verified by credible fact-checking organizations. Finally, a weighted aggregation of domain trust, semantic similarity, and fact-check confidence produces the final credibility score. The way this is done is by using a weighted formula where 40 percent is the domain trust score, 30 percent is the semantic similarity score, and 30 percent is the fact-check score. This is then followed up by normalizing the final score between 0 and 1, where 0 is least credible and 1 is the most credible.

Literature Review

Several machine learning models have been proposed to assess credibility and misinformation detection. One of these proposals was a supervised learning approach that integrates social context, metadata, and linguistic features to classify sources based on their credibility. The study done by the authors of “Predicting Factuality of Reporting and Bias of News Media Sources” demonstrates that models that combine multiple feature sets outperform single-feature models. The authors of “A Transformer-Based Approach for Detecting Misinformation in Online Content” utilized transformer-based architectures, including RoBERTa and BERT, to evaluate the credibility of online content. Their study demonstrated that attention-based NLP models can determine factual accuracy effectively when trained on fact-checked datasets. Finally, the authors of “FEVER: A Large-scale Dataset for Fact Extraction and Verification” introduced the *FEVER dataset*, which provides a benchmark for fact-checking systems by using verifiable claims linked to Wikipedia articles. This dataset played a significant role in training deep learning models for claim verification.

Justification of Chosen Methodologies

There were multiple methodologies chosen for this project, those being machine learning and rule-based methods. Machine learning was chosen because transformer-based models provide robust semantic similarity measures, enhancing content relevance evaluation. Additionally, fact-checking APIs leverage vast knowledge bases to cross-verify claims dynamically. Rule-based methods were chosen because domain trust scores provide a quick filter for identifying unreliable sources with hybrid scoring ensures interpretability and mitigates machine learning biases.

Sources

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). **Predicting Factuality of Reporting and Bias of News Media Sources**. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Retrieved from <https://aclanthology.org/D18-1389.pdf>

Gupta, A., Mittal, A., & Sharma, A. (2021). **A Transformer-Based Approach for Detecting Misinformation in Online Content**. *Journal of Computational Social Science, 4(2)*, 245-268. doi:10.1007/s42001-021-00152-9

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). **FEVER: A Large-scale Dataset for Fact Extraction and Verification**. *Proceedings of NAACL-HLT 2018*. Retrieved from https://aclanthology.org/N18-1074.pdf