Semester Project Proposal

Jinpeng Zhang, Ahmed Hasan, Mátyás Kispéter,
Rubens Onzi, Farzana Afrin, Sonaly Akther

October 16, 2025

1. Title

Information Origin Tracker

2. Abstract

This project addresses the critical issue of information transparency by developing a web application designed to trace the origins of digital content. The system analyzes text from user-submitted web pages to identify and verify the provenance of individual claims. In an era marked by the rapid spread of misinformation, disinformation, and AI-generated articles, this tool provides users with a vital mechanism to critically assess the credibility of their sources, fostering a more informed and discerning public.

3. Context and Problem

3.1. General Context

The broader domain of a news origin tracker project falls under several fields which are primarily:

- Media transparency
- Misinformation detection
- Computational journalism

The importance of this project addresses some of the most relevant issues of modern information ecosystem: it engages misinformation, restores trust in news sources, enhances media literacy through transparency and supports ethical journalism.

3.2. Problem Statement

The project addresses the main problem of lack of transparency on the internet. It is difficult for users to determine if a piece of information is reliable especially if the website lacks transparency. The people who tipically consume information on internet are the main beneficiaries of this solution, they will be empowered with a tool that allows them to check for themselves and verify the sources and origins of a piece of information.

4. Possible Solutions (Software)

4.1. Solution(s) in Terms of Software

The system is going to be structured as follows:

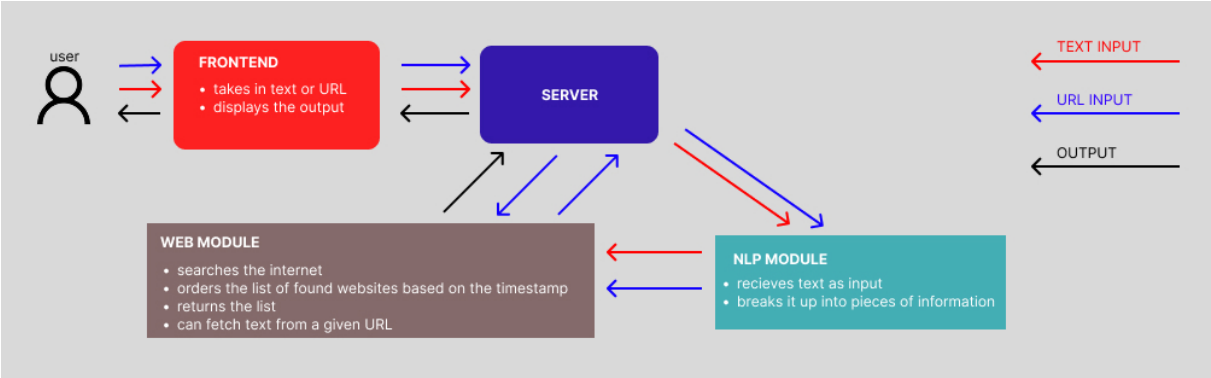

Figure 1: System's Structure

As shown in Figure 1 the system primarily consists of 4 main components which are:

- **Frontend:** the user interface of the web application, through which the user provides input.
- **Server:** functions as the middleware layer.
- **Web Module:** responsible for all operations related to data retrieval from the web.
- **NLP Module:** the core component for text analysis.

4.2. How it Solves the Problem

The program finds the original sources of information and presents them to the user, then the user can decide whether to trust that source or not. This way we give people a tool that helps them combat misinformation and not just blindly trust the first source they come across. The origin of a source is defined as the earliest time that an information has been published on the internet.

5. Relation to Internet / Distributed Systems

5.1. Concepts/Components in Distributed Systems

This project is based on different distributed system concepts:

- **Client-Server Model :** A traditional model where clients send requests, and servers provide the required services or data.
- **Service-Oriented Architecture (SOA):** The four main components are designed as distinct, loosely coupled services. The Server acts as middleware, making requests to the Web Module for data and to the NLP Module for analysis. This design allows for independent development, deployment, and scaling of each component.

- **Parallelism and Concurrency:** In order to improve performance the system can perform multiple operation simultaneously by initiating multiple parallel web scraping tasks to search for different sentences
- **Message Passing:** Communication between the different modules (e.g., from the Server to the Web Module and NLP Module) can be implemented through message passing, such as using a message queue. This would enable asynchronous processing and decouple the services further.
- **Fault Tolerance :** Robust system that works even if one of its components fails.

5.2. Scalability

To handle large data, many users, or high loads, the project will be designed with scalability and resilience in mind. The application will be containerized using Docker, which packages the code and its dependencies into a single, portable unit. These Docker containers will then be deployed and managed across a cluster of different computers using an orchestration platform like Kubernetes. Kubernetes will automatically distribute the workload ensuring the system remains highly available and performant under heavy traffic. This approach ensures that the system is not limited by the capacity of a single machine and can efficiently handle increasing demands.

5.3. Sustainability

The project's main goal is transparency on the internet since it provides every user the tool necessary to retrieve information, thanks to this it also tackles some of the points of the 17 goals set by the UN which are:

- Quality education.
- Peace, justice and strong institutions.

This approach is also future proof given the fact that the amount of unreliable information sources are rising.

6. Possible Challenges and Scope

- **Possible Challenges:** The NLP process will be the most difficult part of the project because it engages in text analysis.
- **Minimal Product (MVP):** A website where the user can input a piece of text and the program will return an ordered list of URLs that mention it.
- **Maximum Ambition:** Website and browser extension that can search for the origins of information and also display all websites that reference the information on a visually appealing timeline.

7. Required Learning

In order to build the information origin tracker, we'll need to learn new topics that we haven't had a chance to explore in depth before. First, we must master Natural Language Processing (NLP) to handle text analysis, which will involve learning and utilizing relevant Python libraries like NLTK or spaCy. Second, we will need to understand and apply web scraping techniques to efficiently and ethically gather information from various online sources. Finally, to prepare for a potentially large user base, we will need to understand the scalability considerations for a browser extension, which often differs from a typical web application.

8. MoSCoW Prioritization

- **Must have:** The core functionality will include web scraping to gather content and Natural Language Processing (NLP) to analyze and extract information.
- **Should have:** A user-friendly and aesthetically pleasing frontend.
- **Could have:** We will consider adding more advanced features like fake news detection and AI-generated text detection as optional improvements if time permits.
- **Won't have (this semester):** However a comprehensive fake news detection and AI text detection will not be included in the initial release.

9. Small Bibliography

- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Srivastava, A. N., & Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*. CRC Press.
- Mueller, M. L. (2010). *Networks and States: The Global Politics of Internet Governance*. MIT Press.