EchoTruth: A Fake News Detection System

Sarron Tadesse, Kalif Byrd, and Dev Raiyani

Submitted on January 21, 2024

Contents

1 Proposal

Background

We chose the EchoTruth project because of our firsthand observations of the impact of fake news on people and the significant damage it can cause. This realization was further amplified in our introduction to AI class, where we explored the concept of sorting emails into categories of 'spam' and 'ham'. This sparked an idea: why not extend this concept to a broader and more impactful domain? As we delved deeper into the realm of fake news detection, our curiosity and eagerness to engage with this challenge only grew. EchoTruth emerged from our desire to apply our computer science skills to a meaningful problem, combining our academic knowledge with the opportunity to learn and implement new technologies. The project aims to leverage straightforward machine learning techniques to develop a tool for identifying and flagging fake news articles, responding to a critical need in today's digital environment where misinformation is prevalent. The project's approach is to create a simple, user-friendly system, making it particularly beneficial for smaller entities and individuals who lack access to extensive fact-checking resources.

How we will approach

Our process begins with collecting a dataset of news articles labeled as true or fake, followed by basic preprocessing tasks like text cleaning and tokenization, preparing the data for machine learning analysis. The heart of EchoTruth is the creation of a classification algorithm, rooted in commonly used natural language processing techniques, designed to be uncomplicated yet effective in distinguishing between true and fake news. We plan to train this model on a significant portion of our dataset, then validate its accuracy and reliability with a separate set of data. Following the development, EchoTruth will undergo standard testing procedures, ensuring its basic functionality and effectiveness before its deployment as a cloud-based application for real-time news analysis.

Goals

Our aim is for EchoTruth to serve as an initial screening tool for news authenticity, supporting but not replacing in-depth fact-checking methods. The project's significance lies in its practical approach, offering an initial solution to the complex problem of fake news. It's an opportunity for us to put our computer science skills to practical use, exploring new technologies and contributing to a real-world issue that has tangible effects on society. In sum, EchoTruth is not just a project; it's our chance to make a meaningful impact in the digital world by addressing the pervasive issue of misinformation through a practical, AI-driven tool, reflecting a grounded and pragmatic approach in our journey as emerging computer scientists.

2 Software Requirements

EchoTruth will be a scalable machine learning-based software system for analyzing news content and classifying it as 'true' or 'fake'. The system aims to enhance media integrity by providing a reliable method to filter out fake news.

3 Development Plan

3.1 Project Planning and Research

- Research existing fake news detection models and techniques.
- Determine the technologies and tools to be used.

3.2 Data Collection and Preparation

- Gather a comprehensive dataset of news articles with labels for true or fake news.
- Preprocess the data including text cleaning, tokenization, and vectorization.
- Split the data into training, validation, and test sets.

3.3 Feature Engineering

- Identify and extract features significant for distinguishing true from fake news.
- Implement NLP techniques, sentiment analysis.

3.4 Model Development

- Choose a suitable machine learning algorithm.
- Train the model on the training dataset and validate using the validation dataset.

3.5 Evaluation and Optimization

- Evaluate the model's performance using accuracy, precision, recall, and F1 score.
- Optimize the model through various algorithmic, feature, and hyperparameter adjustments.

3.6 Testing and Validation

- Conduct comprehensive testing including unit, integration, and system tests.
- Validate the model in real-world scenarios.

3.7 Deployment

- Deploy the model onto a server or cloud platform for user access.
- Ensure a secure and scalable deployment environment.

3.8 Maintenance and Updates

- Regularly monitor system performance and accuracy.
- Update the model with new data and improved algorithms.

3.9 Documentation and Reporting

- Document the methodology, challenges, solutions, and results.
- Prepare a final report or presentation on EchoTruth.

4 Tools and Technologies

For the EchoTruth project, we will utilize a variety of technologies and tools, each chosen for its effectiveness and suitability to our project goals.

Programming Language

- Python - Already experienced and widely used for its versatility in data analysis and machine learning.

Data Analysis and Visualization

- Pandas - For data manipulation and analysis.
- Matplotlib and Seaborn - For plotting, data visualization, and advanced statistical visualizations.

Natural Language Processing

- NLTK - For text processing and basic NLP tasks.
- spaCy - For more advanced NLP operations, if required.

Machine Learning Framework

- Scikit-learn - User-friendly framework for model training and evaluation, suitable for our familiarity with AI.

Database/Storage

- SQLite - A lightweight SQL database for local data storage and testing.
- Kaggle - As a source for datasets.

Version Control and Collaboration

- GitHub - For code versioning and collaboration.

Web Framework (for API development)

- Flask - A lightweight framework, easier to use for creating simple web apps or APIs.

Cloud Platform for Deployment

- Heroku - Chosen for its ease of deployment and maintenance.

IDE/Code Editor

- Jupyter Notebook - For interactive development and experimentation.
- Visual Studio Code - A versatile text editor for general coding.

Testing and CI/CD

- PyTest - For testing Python code.
- GitHub Actions - For continuous integration and deployment, offering seamless integration with GitHub.

5 Configuration Management Document

The Configuration Management process for EchoTruth will involve using GitHub for version control. All changes and versions of the project code and documentation will be tracked through GitHub repositories. This will include feature branching and pull requests for managing different stages of development and ensuring that all changes are reviewed before being merged into the main project.

6 Testing

The testing strategy for EchoTruth will include unit tests, integration tests, and system tests. These tests will be designed to ensure the accuracy and reliability of the fake news detection model as well as the functionality of the user interface and overall system integration.

7 Timeline

- Week 1 and 2: Pitch Project
- Week 3: Begin Understanding Data and do EDA
- Week 4: Begin Model Deployment
- Week 5: Train Dataset
- Weeks 6 to 13: TBD