UMassAmherst | Manning College of Information & Computer Sciences
Center for Data Science & Artificial Intelligence

Adobe   Media Cloud

Beyond Keywords: A Pipeline for Nuanced News Analysis

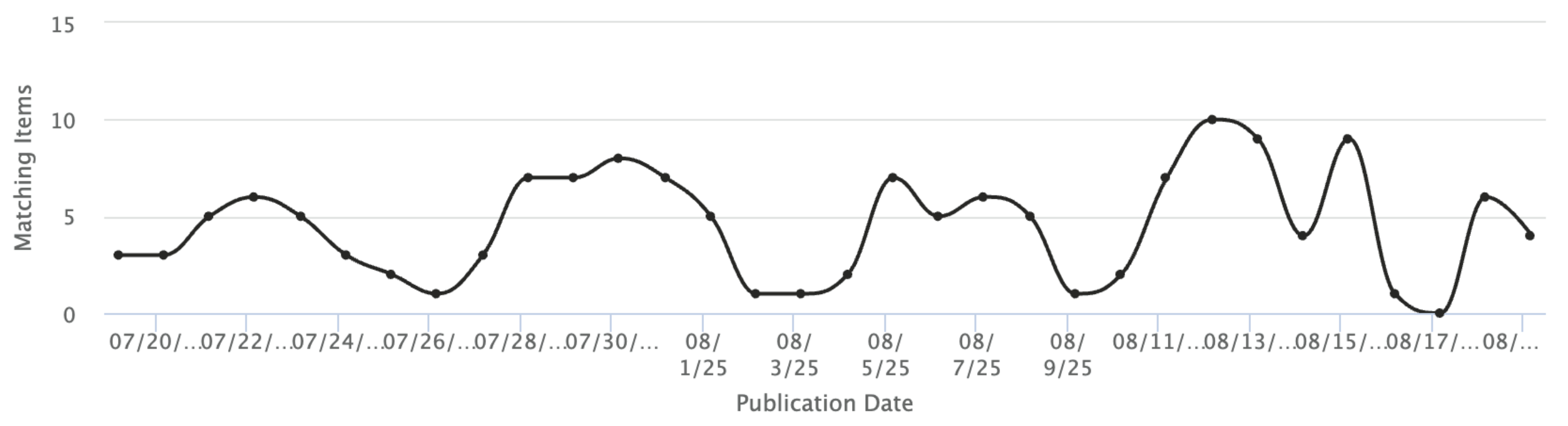
Arav Adikesh Ramakrishnan, Chandana Magapu, Stella Dey, Haamed Rahman, and Virginia Partridge

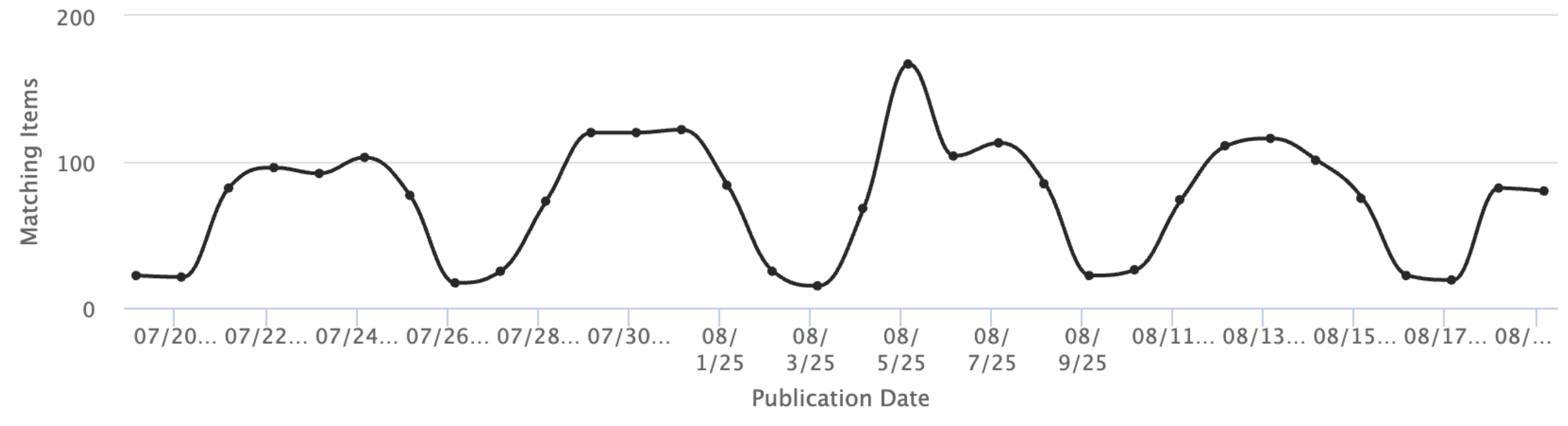
Summary

We have developed a streamlined, end-to-end pipeline that makes the process of building custom text classifiers both accessible and repeatable for news media analysis. By integrating data fetching, collaborative annotation, and model training into a single, manageable workflow, our project removes significant technical barriers, empowering researchers to spend less time on engineering and more time on critical anlalysis.

Motivation

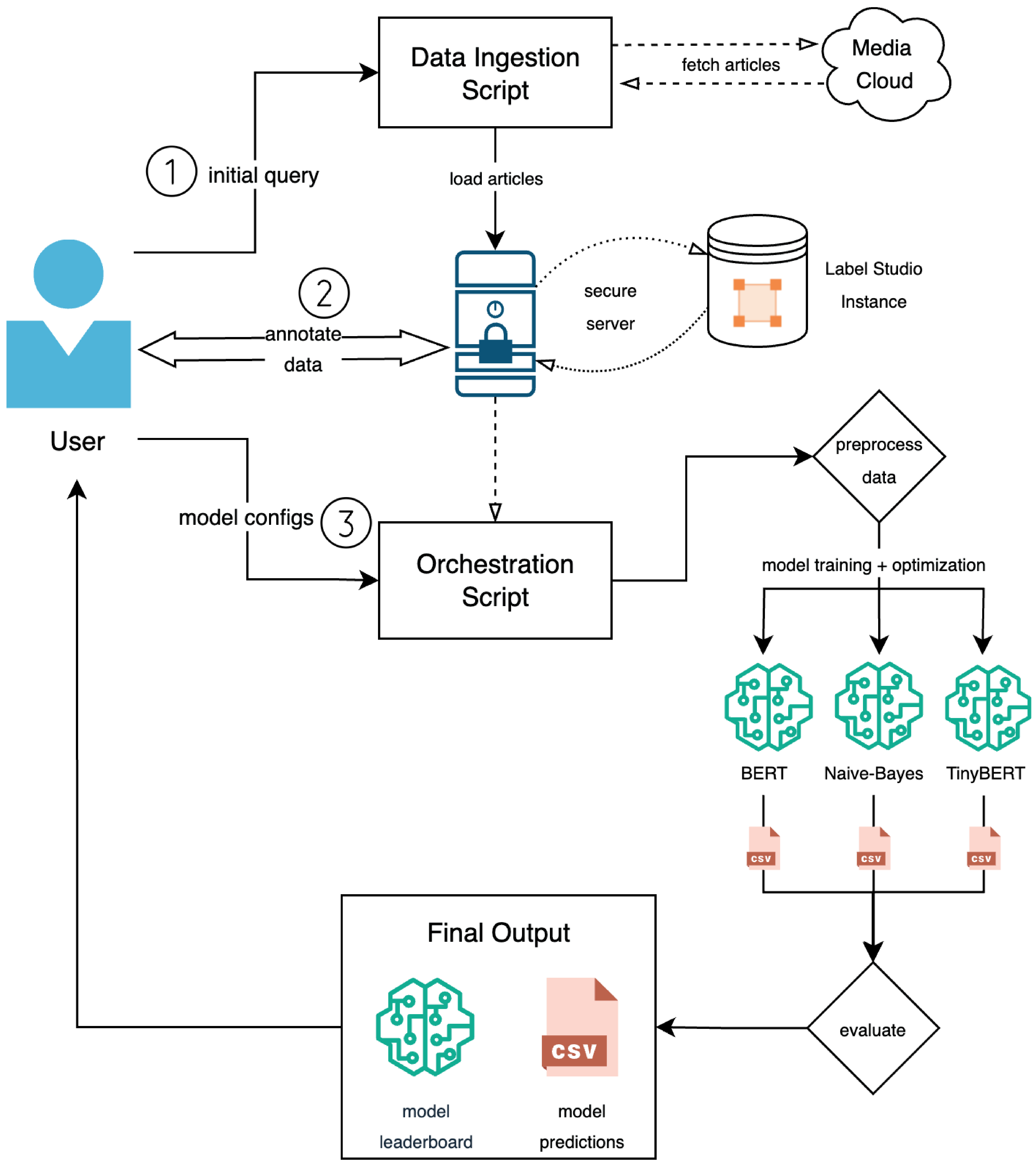
Researchers using the massive Media Cloud news archive often need to find articles about complex, nuanced topics like "solutions-based journalism" or "feminicide." Simple keyword searches are often insufficient for these tasks, as they can miss relevant articles or include many false positives. This creates a significant barrier to performing large-scale analysis of media narratives on critical topics.





Goal: Researchers need a way to train their own custom models to classify articles based on meaning and context.

Architecture



Approach

Our pipeline consists of three stages:

1. **Data Ingestion**: Users create a boolean query to fetch relevant news articles from the Media Cloud API, which is then loaded into a LabelStudio project for annotation.
2. **Annotation:** Articles are collaboratively labeled by researchers to create a custom dataset.
3. **Training & Evaluation**: The labeled data is used to train multiple classifiers, allowing researchers to evaluate a range of options i.e., **BERT-based** or **Naive-Bayes**

Future Directions

The current pipeline for custom model creation has a solid foundation but can be enhanced with:

- AutoML Tools: Add support for tools like AutoGluon or Ludwig to achieve further model abstraction.
- Model Explainability: Add tools like LIME or SHAP for transparency in model decisions.
- Enhanced User Interface: Create a more intuitive dashboard for easier management and monitoring.
- Active Learning: Implement strategies that lower annotator burden and allow full complete cyclical model development.

Codebase Here!