## 



Project Proposal: Explainable Fake News Detection System

28.03.2025

─  
*Aymane Sbai  
Hanie Boukmakh  
Imane Guessous*  
MsBDA

Al Akhawayn University

CSC 5354 - NLP for Big Data

# Motivation

## Problem Significance

The spread of fake news has become a critical global issue, specifically in social media, with misinformation spreading faster and farther than factual content. Studies show that false news stories are 70% more likely to be retweeted than true stories (Vosoughi et al., MIT Sloan 2018). This has led to a large volume of societal harms, including:

* Public health crises (we all remember the covid era)
* Political polarization and election interference (as seen in the last Trump campaign)
* The death of trust in legitimate journalism

Current solutions rely heavily on manual fact-checking by organizations like PolitiFact and Snopes and the fact-checker provided by X, which cannot scale to match the volume of content shared daily on social media platforms.

## Our Solution

We propose an automated, explainable NLP system that:

* **Flags Potentially Misleading Claims:**
  + Uses LIAR’s labeled data (PolitiFact/SNOPES-verified statements) as a reference.
  + Predicts whether new claims resemble known true/false patterns.
* **Explains Its Reasoning:**
  + Highlights phrases that contributed to the prediction (e.g., "third-trimester abortions on demand").
* **Operates Transparently:**
  + Clearly states it cannot verify truth—only identifies similarities to historical misinformation.

This will ease the load on human fact-checkers, and independent non-biased journalists to provide accurate checks and verifications without having to go through all the news headlines.

For example, social media platforms (e.g., X, Facebook) use similar systems to flag content for human review.

Therefore, this will reduce manual fact-checking workload by prioritizing **suspicious claims.**

# Linkage to Course Material

## NLP Techniques Applied

Our project builds directly on course concepts while extending them with advanced applications, the core concept we will tackle:

* **Text Classification:** The NLP task we will be tackling on this project, as seen in the course material, using transformers models.
* **Feature Engineering:** Incorporating metadata such as speaker party as auxiliary feature.
* **Transformer Models:** Fine-tuning DistilBERT (a distilled version of BERT) for binary classification available on the huggingface website on LIAR dataset (<https://huggingface.co/docs/transformers/en/model_doc/distilbert>)

## Beyond Classroom Coverage

our project extends the course materials by:

* Implementing LIME (Local Interpretable Model-agnostic Explanations) to explain predictions
* Combining textual analysis with metadata features for richer predictions
* Bias Auditing: We’ll test if model performance differs by speaker party.

# Schedule and Phases

## Project Phases Overview:

| Week | Phase | Tasks | Team Member |
| --- | --- | --- | --- |
| Week 1-2 | Data Preparation and Preprocessing | Clean LIAR data, map labels | Hanie |
| Week 3-4 | Model Training | Fine-tune DistilBERT + party feature | Aymane |
| Week 5 | Explainability | Implement LIME on test cases (a subset of tests) | Imane |
| Week 6-7 | UI Development & Testing with Documentation | Build Streamlit app with disclaimers and Bias audit | All |

(These are subject to change, and any change will be visible in the report by the end of the project)

## Members Tasks

| Member | Tasks | Metrics |
| --- | --- | --- |
| Aymane | Model Training, Hyperparameter Tuning | Try to achieve a high validation accuracy |
| Hanie | Data Cleaning, UI Backend | Clean Dataset and functional API endpoints |
| Imane | LIME explanations and simple frontend | Explainable examples. |

# Sources & References

## Primary Dataset

**1. LIAR Dataset**

* Wang, W. Y. (2017). "LIAR: A Benchmark Dataset for Fake News Detection"
* Paper Link: <https://arxiv.org/abs/1705.00648>
* Content: 12,800 manually labeled political statements (20102016) from PolitiFact.
* Labels: Six fine-grained truth ratings (true, mostly-true, half-true, barely-true, false, pants-fire).
* Download: <https://www.cs.ucsb.edu/~william/data/liar_dataset.zip>

## Models & Libraries

**2. DistilBERT**

* Sanh, V., et al. (2019). "DistilBERT, a Distilled Version of BERT"
* Paper Link: <https://arxiv.org/abs/1910.01108>
* Implementation: <https://huggingface.co/distilbert-base-uncased>

**3. LIME (Explainability)**

* Ribeiro, M. T., et al. (2016). "Why Should I Trust You?" Explaining Predictions of Any Classifier"
* Paper Link: <https://arxiv.org/abs/1602.04938>
* Code: <https://github.com/marcotcr/lime>

## Supporting Research

**4. Misinformation Spread**

* Vosoughi, S., et al. (2018). "The Spread of True and False News Online"
* Article Link: <https://science.sciencemag.org/content/359/6380/1146>

**5. Bias in Fact-Checking**

Nguyen, A., & Vu, H. (2019). "Testing Popular News Discourse on the Rise of Fake News"

The Github Repo for this project: <https://github.com/Aymzzz/fake_news_detector>