ITM 454 – Final Project Proposal
Project Title: Automated Fake News Detection System

Team Members: Virakyuth SRUN, Raksmey POL, Henglong LY, Sokati KEO

I. Problem Statement

Fake news spreads faster than real news on social media and creates serious problems for society, democracy, and public health. For example, research shows false news spreads six times faster than the truth on Twitter, and misleading health posts on Facebook gained billions of views during COVID-19. To help address this, we will build an **automated system that uses machine learning to detect fake news articles** in real time.

II. Proposed Approach

Our system will use **supervised machine learning** for text classification. The main steps are:

1. **Data Collection:** Use a labeled dataset of real and fake news from Kaggle.
2. **Preprocessing:** Clean and prepare the text by tokenizing, lowercasing, removing stopwords/punctuation, and applying stemming or lemmatization.
3. **Feature Engineering:** Convert text into numerical form using TF-IDF or word embeddings.
4. **Modeling:** Test classifiers such as Naive Bayes, Logistic Regression, and the Passive-Aggressive Classifier.
5. **Training & Evaluation:** Train models on part of the dataset, then evaluate them on test data.

III. Dataset

We will use the **Fake News Detection Dataset** (Kaggle, by Emine Bozkus, PhD). It includes over 25,000 news articles labeled as *real* or *fake*, along with fields like title, text, subject, and date. This dataset is balanced and suitable for NLP-based classification.

IV. Expected Outcomes & Evaluation

- **Target Accuracy:** At least 80% for Naive Bayes and Logistic Regression.
- **Metrics:**
  - Accuracy: Overall correct predictions.
  - Precision: Avoid false alarms (real news marked fake).
  - Recall: Catch as many fake articles as possible.
  - F1-Score: Balance between precision and recall.
  - Confusion Matrix: Detailed breakdown of results.

V. References

- Avaaz. (2020, May 14). *Global disinformation report: The toxic ten*.
- Vosoughi, S., Roy, D., & Aral, S. (2018, March 9). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
- Bisaillon, C. (2017). *Fake and real news dataset*. Kaggle. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset