# Project Milestone 1
## (Initial Submission)

Deep Learning and Software Engineering

Agustín Tamagnone

Birk Bregendahl

Karol Swiderski

# Milestone 1 – Initial

## Index

# 1. Context and Requirements

This project focuses on developing a Fake News Detection System designed to assist users (general public, editors, analysts, students, etc.) in identifying potentially unreliable or misleading online articles. The solution employs a Deep Learning (DL) text-classification model trained on the ISOT Fake News Dataset, created by the University of Victoria in Canada.
The dataset includes articles labeled as *True* or *Fake* and contains four key columns: title, text, subject, and date. Using these features, the model analyzes both the title and the body of an article to predict whether the content is likely to be fake or real. The system's main objective is to demonstrate how natural-language models can recognize linguistic and stylistic cues that commonly appear in either credible or deceptive news writing.

## 1.1 System Overview

The proposed system will be implemented as a lightweight web application that allows users to paste or upload an article for evaluation. In the initial stage, this interface can take the form of a simple notebook or prototype webpage. In later milestones, it will evolve into a fully deployed web platform connected to a backend API.

At the core of the application lies the API service, which handles prediction requests. The endpoint (for instance, "/predict") accepts article data in JSON format and returns the model's prediction together with a probability score. This structure ensures that the system can easily integrate with other tools or visual dashboards in the future.

Before reaching the model, the text passes through a preprocessing module responsible for cleaning and vectorizing the data. TF-IDF is used to convert the text into numerical form so that it can be processed by the neural network.

The Deep Learning model itself is a baseline Dense Neural Network trained on these text features. After training, the model is saved as a serialized file (for example, ".h5" or ".pkl") and loaded by the backend whenever predictions are made.

A decision and logging component interprets the model's probability output, assigns the *Fake* or *True* label, and records anonymized logs for later evaluation. The system also includes an evaluation module that calculates standard performance metrics such as accuracy, precision, recall, and F1-score. These metrics can be viewed through a dashboard or within a notebook to support continuous monitoring and comparison with more advanced models in future milestones.

## 1.2 Functional Requirements

The system is designed to allow users to input or upload news articles directly through a simple web interface. Each submission includes at least a title and main text, while optional metadata such as subject and publication date can also be provided. Once an article is received, the system cleans and standardizes the text, extracting the relevant elements required for further processing. The normalized text is then passed to the Deep Learning model, which generates a probability score between 0 and 1, indicating the likelihood that the article is fake.

Based on this score, the system assigns one of three interpretive labels: *likely fake* when the score falls below 0.35, *uncertain* when it lies between 0.35 and 0.65, and *likely true* when it is equal to or above 0.65. These results are displayed on the user interface together with the confidence value. In addition, the interface may highlight the most influential words or phrases that contributed to the prediction, giving users a clearer understanding of how the model reached its conclusion.

To support broader integration, the project also includes a REST API endpoint that enables programmatic access to the prediction service. This interface allows other applications to send article data in JSON format and receive predictions automatically. The API validates the structure of incoming data and rejects unsupported or malformed submissions. All predictions and user interactions are logged in anonymized form, ensuring traceability and enabling future improvements. Furthermore, version information about both the dataset and the model is stored to maintain transparency and reproducibility across experiments.

## 1.3 Non-Functional Requirements

The system must operate efficiently on standard hardware. The deep learning component should process an article within approximately three seconds on a CPU, while the web interface is expected to display results within two seconds of receiving a prediction. Although designed primarily for research and demonstration purposes, the application aims to maintain around ninety-five percent uptime during use. In the event of a technical error—such as a model failure—the system will respond gracefully by showing an informative message rather than crashing.

All communication between the web client and the backend service is encrypted through HTTPS, and API access is restricted to authenticated users when deployed. To protect privacy, article texts and predictions are stored only when explicitly required for evaluation, and any stored data are anonymized to comply with basic privacy principles.

The system architecture has been designed with modularity in mind. Separate components handle preprocessing, model inference, and user interaction, which makes the solution easier to maintain and extend. This design also supports future

scalability, allowing more advanced models—such as transformer-based architectures like DistilBERT—to be integrated later without the need for major structural changes.

## 1.4 Success Criteria (Milestone 1)

At this stage, the focus is on defining and preparing the baseline model architecture rather than evaluating its performance. No formal testing has been conducted yet; however, model training and evaluation will be performed in the following milestones. The goal for future stages is for the Dense Neural Network using TF-IDF features to achieve an F1-score of at least 0.85 on a held-out test set, with single-article predictions executed within approximately three seconds on standard CPU hardware. The training process will be designed for full reproducibility, and a brief analysis of misclassified examples will be performed once evaluation results are available.

## 1.5 Key Risks and Mitigations

The project faces several common challenges, including bias, data imbalance, and overfitting. Because the ISOT dataset mainly represents political news and stylistic differences between sources, the model may incorrectly associate emotional or sensational language with fake content rather than factual inaccuracy. This limitation will be documented, and results will be evaluated by subject category to detect potential bias.

Imbalances in class distribution or hidden patterns in titles could also influence model performance. These risks will be reduced through stratified data splits and the use of regularization techniques. To minimize overfitting on sparse TF-IDF features, dropout layers and early stopping will be applied during training.

## 1.6 Ethics and Privacy

The system is intended solely for educational purposes, illustrating how deep learning can recognize patterns of language and style in news writing. It does not verify factual accuracy. Each prediction will include a confidence score and, where possible, a brief explanation of the most influential words contributing to the classification. All data is processed locally, and no personal or identifying information is stored. The project follows responsible AI principles by prioritizing transparency, privacy, and human oversight in interpreting results.

# 2. High-Level Description of the Model

The goal of our deep learning component is to classify news articles as real or fake based on their textual content. The model receives the concatenated title and body of an article as input and outputs a probability indicating how likely it is to be fake. The problem is therefore framed as a binary text classification task.

To represent the textual data numerically, the input text will be transformed into TF-IDF vectors (term-frequency–inverse-document-frequency). The TF-IDF representation captures the relative importance of words across documents while remaining computationally efficient. The TF-IDF features will then be passed to a feed-forward neural network composed of several fully connected (Dense) layers with ReLU activations and dropout for regularization. The final output layer will use a sigmoid activation to return a single probability between 0 and 1. The model will be trained using binary cross-entropy loss and optimized with the Adam optimizer.

Several model families were considered for this task:

- Recurrent Neural Networks (RNNs / LSTMs): effective for sequential data but computationally heavier and slower to train.

- Convolutional Neural Networks (CNNs): can capture local word-pattern features but are less interpretable for global context.

- Transformer-based models (BERT, DistilBERT): state-of-the-art for text understanding but require significantly more data and compute resources.

Given the project scope and the current milestone's focus on defining a baseline model rather than optimizing performance, the team selected the feed-forward Dense Neural Network as a balanced starting point. It is simple, interpretable, and well-suited to TF-IDF representations of text. This choice allows us to establish a clear baseline for later comparison. In future milestones, we plan to explore Transformer-based architectures (such as DistilBERT) to assess whether contextual embeddings can improve classification accuracy and robustness once the system foundation is complete.

# 3. Model Card

## *Model Details*

- Developed by students at Universidad Politécnica de Madrid, 2025, v1.
- It uses a Dense Neural Network trained on TF-IDF.
- Binary classification, the model predicts if a news article is *False* or *True*.

## *Intended Use*

- The model is designed to help users identify potentially fake articles and support human review.
- Intended to aid users (general public, editors, analysts, students, etc.) identify fake news and increase awareness about misinformation.
- It should not be used to make final decisions about truth or credibility, as the model gives a probability, but it does not actually verify if the news is true or false.

## *Factors*

- Results may vary depending on:
    - Topic of the article (for example, politics, world news, sports)
    - Writing style
    - Language (our dataset only contains English)

## *Metrics*

- Main metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-score

- The goal is to reach an F1-score of at least 0.85 once the model is trained.
- Predictions will be based on a threshold of 0.5 (above = real, below = fake)

## *Training Data*

- **Dataset:** Fake and Real News Dataset (Kaggle, by Clement Bisaillon)
    - The dataset mentioned above is a copy/cleaned version of the ISOT dataset created by Clement Bisaillon in University of Victoria, Canada.

- The dataset contains about 44,000 English (language) news articles, separated into *real (21,417)* or *fake (23,481)*.
- Each record includes: title, text, subject, and date.
- First clean and prepare the text, then turn it into numbers using TF-IDF so we can train a basic neural network. Later, we plan to use the raw text to fine-tune a more advanced model called DistillBERT, which understands language better.

## *Evaluation Data*

- A portion of the dataset *'Fake and Real News'*, will be used for testing after training.
- Evaluation will measure how well the model predicts the correct label and meets the target F1 score.
- Data will be stratified to keep real/fake balance and ensure fair comparison.

## *Ethical Considerations*

- Opinion and biases from the dataset (tone and topic imbalance) could affect predictions.
- No personal or sensitive information is used. All texts come from public sources.
- Results should always be reviewed by humans before making a conclusion.

### *Caveats and Recommendations*

- The model may work best with English-language articles similar to those in the dataset.
- It may mislabel opinion or emotional pieces as fake because of the writing tone.

# 4. Data Card

## 1. *Dataset Description*

Homepage: Kaggle - Fake News Dataset By Clement Bisaillon

Repository: Kaggle

Paper: Originally based on the ISOT Fake News Dataset (University of Victoria, 2017).

Point of Contact: Clement Bisaillon (Kaggle Author), ISOT Lab - University of Victoria, Canada.

### *1.1.* Dataset Summary

- This dataset contains news articles labeled as *Fake* or *Real*.
- It is used to train and test text-classification models that detect misinformation in online news.
- The articles come from public sources and include a title, text, subject, and date.
- It helps researchers and students understand how linguistic patterns differ between fake and real news.

### 1.2 Languages

- Language: English only
- Writing style: formal and journalistic

# 2. Dataset Structure

## 2.1. Data Fields

Table 1 - Dataset card Data Fields

| Field | Type | Description | Structure | Usage |
|---|---|---|---|---|
| ID | Integer | Unique identifier for each article | Single integer per article | Metadata |
| Date | String (date format) | Date of the article. | Single timestamp | Metadata |
| Subject | String | Topic of the article | Single categorical value | Input |
| Title | String | Headline of the article | Word level, contiguous text | Input |
| Text | String | Main body of the article | Word level, contiguous text | Input |
| Label | Binary (String or Integer) | 1=Real, 0=Fake | Single label per instance | Output |

## 2.2 Data Splits

Table 2 - Data Splits

| | Train | Validation | Test |
|---|---|---|---|
| *Input Sentences* | 35,916 (80%) | 4,490 (10%) | 4,490 (10%) |
| *Average Sentence Length* | 405.45 | 403.54 | 410.84 |
| *Purpose* | Model training | Fine-Tune | Evaluate performance |

# 3. Dataset Creation

## 3.1. Curation Rationale

- The dataset was created to help general public interest in detecting fake news automatically using text-classification. Dataset can also be used by editors, students, analysts, etc.
- It is useful for studying misinformation, language tone, and article writing styles.

## 3.2. Source Data

### 3.2.1. Initial Data Collection and Normalization

- The data was collected from online news websites between 2015 and 2017.
- Real news came from reliable sources (e.g., Reuters, BBC).
- Fake news came from unreliable or satirical websites.
  The data was cleaned and saved as two CSV files: *Fake.csv* and *True.csv*.
- No detailed preprocessing steps are documented.

### 3.2.2. Who are the Source Language Producers?

- The text was written by journalists and online article authors (humans).
- The dataset creators did not change the text, only collected it from existing sources.
- No compensation was given because the data was already public.

## 3.3. Annotations

### 3.3.1. Annotation Process

- Articles were labeled as *Fake* or *True* according to the source websites.
- No human annotators were hired for this dataset.

#### 3.3.2. Who are the Annotators?

- The labels were produced automatically when collecting the data, based on the website category (trusted or untrusted).

### 3.4. Personal and Sensitive Information

- The dataset does **not** contain personal, private, or sensitive data.
- All articles were publicly available and do not include names, addresses, or identifiable information.

## 4. Considerations for Using the Data

### 4.1. Discussion of Biases

- The dataset is limited to English and mostly political content, which could bias results.
- Writing style and emotional tone may influence predictions (for example, strong opinions might appear as “fake”).
- Models trained on this dataset might not generalize to other languages or informal text.

### 4.2. Other Known Limitations

- The data is from 2015 to 2017, so it might not reflect today’s news style, especially that it may not capture modern misinformation challenges such as AI-generated content.
- No multi-language or multimedia (image/video) data is included.
- Some topics are overrepresented (especially politics)

## 5. Additional Information

### 5.1. Dataset Curators

- Kaggle Author: Clement Bisaillon
- Original Source: ISOT Lab, University of Victoria (Canada)

### 5.2. Licensing Information

- License: For research and educational use only.
- CC BY-NC-SA 4.0

## 5.3. Citation Information

- Bisaillon, C. (2020). Fake and Real News Dataset. Kaggle.
- Originally collected by ISOT Lab, University of Victoria (2017).

## 5.4. Contributions

- Thanks to Clement Bisaillon (Kaggle) and ISOT Lab for providing dataset to the research community.